Andre Chalom

Exploração de espaços de parâmetros de modelos biológicos sob diferentes paradigmas estatísticos

São Paulo
2014

Andre Chalom

Exploração de espaços de parâmetros de modelos biológicos sob diferentes paradigmas estatísticos

Dissertação de mestrado apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática - Universidade de São Paulo

Orientador: Paulo Inácio de Knegt López de Prado

São Paulo
2014

Chalom, A.
Exploração de espaços de parâmetros de modelos biológicos sob diferentes paradigmas estatísticos
128 páginas
Dissertação (Mestrado) - Programa Interunidades de Pós-Graduação em Bioinformática.

1. Análise de sensibilidade
2. Análise de incerteza
3. Modelagem numérica
4. Verossimilhança

I. Universidade de São Paulo. Programa Interunidades de Pós-Graduação em Bioinformática.

Comissão Julgadora:

Prof. Dr.
Nome

Prof. Dr.
Nome

Prof. Dr.
Paulo Inácio de Knegt López de Prado

"In an ideal world, all data would come from well-designed experiments and would be sufficient to simultaneously estimate all parameters using rigorous statistical procedures. The world is not ideal. One must often combine estimates from different experiments, or supplement high-quality data (...) with uncertain data, or even assumptions (...). Think carefully about whether your conclusions may be artifacts of your assumptions or calculations, and document those assumptions and calculations so that your reader can ask the same question, and then carry on."

(H. Caswell, Matrix Population Models)

"There comes a time in the life of a scientist when he must convince himself either that his subject is so robust from a statistical point of view that the finer points of statistical inference are irrelevant, or that the precise mode of inference he adopts is satisfactory. Most will be able to settle for the former, and they are perhaps fortunate in being able to conserve their intellectual energy for their main interests; but some will be forced, by the paucity of their data or the complexity of their inferencies, to examine the finer points of their own arguments, and in so doing they are likely to become lost in the quicksands of the debate on statistical inference."

(A.W.F Edwards, Likelihood)

"To say that the probability of rain is 60% is neither to say that 'it will rain', nor that 'it will not rain': however, in this way, everyone is in a better position to act than they would be had metereologists sharply answered either 'yes' or 'no'."

(B. de Finetti, Philosophical Lectures on Probability)

"Uncertainty is a personal matter; it is not *the* uncertainty but *your* uncertainty."

(D. Lindley, Understanding Uncertainty)

Resumo

A formulação e o uso de modelos matemáticos complexos têm recebido grande atenção no estudo da ecologia nos últimos anos. Questões relacionadas à exploração de espaços de parâmetros destes modelos - executada de forma eficiente, sistemática e à prova de erros - são de grande importância para melhor compreender, avaliar a confiabilidade e interpretar o resultado destes modelos. Neste trabalho, apresentamos uma investigação de métodos existentes para responder as questões relevantes da área, com ênfase na técnica conhecida como Hipercubo Latino e com foco na análise quantitativa dos resultados, e realizamos a comparação entre resultados analíticos de incerteza e sensibilidade e resultados obtidos do Hipercubo. Ainda, examinamos a proposta de uma metodologia paralela baseada no paradigma estatístico da verossimilhança.

O capítulo 1 introduz uma investigação a respeito dos conceitos históricos sobre a natureza da probabilidade, situando o conceito da verossimilhança como componente central da inferência estatística.

O capítulo 2 (em inglês) traz uma revisão bibliográfica sobre o estado da arte em análises de incerteza e sensibilidade, apresentando dois exemplos de aplicação das técnicas descritas a problemas de crescimento populacional estruturado.

O capítulo 3 examina a proposta de uma metodologia baseada na verossimilhança dos dados como uma abordagem integrativa entre a estimação de parâmetros e a análise de incerteza, apresentando resultados preliminares.

Durante o progresso do presente trabalho, um pacote de funções na linguagem **R** foi desenvolvido para facilitar o emprego na prática das ferramentas teóricas expostas acima. Os apêndices deste texto trazem um tutorial e exemplos de uso deste pacote, pensado para ser ao mesmo tempo conveniente e de fácil extensão, e disponível livremente na internet, no endereço http://cran.r-project.org/web/packages/pse.

Palavras-chave: análise de sensibilidade, análise de incerteza, modelagem numérica, verossimilhança

Abstract

There is a growing trend in the use of mathematical modeling tools in the study of many areas of the biological sciences.

The use of computer models in science is increasing, specially in fields where laboratory experiments are too complex or too costly, like ecology.

Questions of efficient, systematic and error-proof exploration of parameter spaces are are of great importance to better understand, estimate confidences and make use of the output from these models.

We present a survey of the proposed methods to answer these questions, with emphasis on the Latin Hypercube Sampling and focusing on quantitative analysis of the results. We also compare analytical results for sensitivity and uncertainty, where relevant, to LHS results. Finally, we examine the proposal of a methodology based on the likelihood statistical paradigm.

Chapter 1 introduces a brief investigation about the historical views about the nature of probability, in order to situate the concept of likelihood as a central component in statistical inference.

Chapter 2 (in English) shows a revision about the state-of-art uncertainty and sensitivity analyses, with a practical example of applying the described techniques to two models of structured population growth.

Chapter 3 examines the proposal of a likelihood based approach as an integrative procedure between parameter value estimation and uncertainty analyses, with preliminary results.

During the progress of this work, a package of **R** functions was developed to facilitate the real world use of the above theoretical tools. The appendices of this text bring a tutorial and examples of using this package, freely available on the Internet at http://cran.r-project.org/web/packages/pse.

Keywords: uncertainty analysis, sensitivity analysis, numerical modeling, likelihood

Contents

Capítulo 1

Sobre probabilidade e inferência

1.1 A história e a natureza da probabilidade

John Venn, no prefácio do seu estudo de 1866, escreve:

> "The science of Probability occupies at present a somewhat anomalous position. It is impossible, I think, not to observe in it some of the marks and consequent disadvantages of a *sectional* study. By a small body of ardent students it has been cultivated with great assiduity, and the results they have obtained will always be reckoned among the most extraordinary products of mathematical genius. But by the general body of thinking men its principles seem to be regarded with indifference or suspicion. Such persons may admire the ingenuity displayed, and be struck with the profoundity of many of the calculations, but there seems to them, if I may so express it, an *unreality* about the whole treatment of the subject. To many persons the mention of Probability suggests little else than the notion of a set of rules, very ingenious and profound rules no doubt, with which mathematicians amuse themselves by setting and solving puzzles." [Venn, 1866]

No século e meio que se passou desde esta publicação, o uso da ciência da probabilidade se tornou uma obrigatoriedade no meio científico, principalmente sob a forma da inferência estatística. Em qualquer periódico renomado da área, é virtualmente impossível publicar um artigo experimental que não mencione alguma propriedade estatística a respeito das amostras coletadas ou de resultados experimentais obtidos. Por muitos anos, a estatística vigente nas análises biológicas teve uma forte inspiração frequentista, com testes de hipóteses e valores p sendo requeridos para publicações. No entanto, existe uma ampla e duradoura crítica à forma como essa estatística é utilizada nas ciências biológicas e na área de medicina [Barber and Ogle, 2014; Burnham et al., 2011; Gardner and Altman, 1986; Ioannidis, 2005].

Muito da crítica decorre da interpretação simplista assumida por muitos cientistas a respeito do valor p reportado sobre um problema, mas as suas raízes devem ser traçadas mais profundamente em uma falta de consciência dos cientistas sobre as interpretações a respeito da natureza das probabilidades que são pressupostas por

uma escola de pensamento, mas raramente examinadas profundamente. De outro lado, análises recentes sugerem que o tamanho dos efeitos em várias áreas da ecologia são tão pequenos que estes dificilmente serão corretamente detectados pelos procedimentos tradicionais [Jennions and Moeller, 2003]. Em áreas como a ecologia do comportamento, o poder dos testes estatísticos é muito pequeno, e esse problema é acentuado quando são utilizados métodos (como o método de Bonferroni) para corrigir o valor p reportado. A solução para esses problemas pode não estar ligada a desenvolver novos métodos ou procedimentos estatísticos, mas na análise crítica e racional dos pressupostos por trás dos testes frequentemente utilizados.

Novamente, citamos John Venn:

> "Students of Philosophy in general have thence conceived a prejudice against Probability, which has for the most part deterred them from examining it. As soon as a subject comes to be considered 'mathematical' its claims seem generally, by the mass of readers, to be either on the one hand scouted or at least courteously rejected, or on the other hand to be blindly accepted with all their assumed consequences. Of impartial and liberal criticism it obtains little or nothing." [Venn, 1866]

Embora o debate sobre a natureza da probabilidade tenha sido fervorosamente promovido por matemáticos e estatísticos, este não alcança, muitas vezes, o cientista que faz uso do ferramental da probabilidade para resolver um problema particular. Assim como o debate matemático era ignorado por estudantes de filosofia à época de Venn, hoje o debate filosófico e epistemológico é ignorado por estudantes da área de ciências e cientistas. Isso pode ser parcialmente atribuído ao aumento histórico na especialização de currículos - o que diminui o entusiasmo e desempenho dos estudantes em áreas não relacionadas à sua especialidade [Popham, 1986].

Para poder discorrer sobre a estimação de incertezas e a atribuição de probabilidades a diferentes resultados de um modelo, é necessário examinar com atenção o que entendemos sobre probabilidade. O conceito por trás da palavra probabilidade vai embasar não apenas o que entendemos pela palavra probabilidade em si, mas também traz um significado para os conceitos de incerteza, inferência e simulação, que serão de importância para a discussão a seguir.

Embora o debate surja como uma questão sobre a natureza das probabilidades, vamos encontrar as divergências práticas em torno de questões de inferência. Um problema simplesmente de atribuição de probabilidades sobre um experimento mental, que seja bem formulado, poderá ter a mesma solução, não importando a afiliação filosófica de quem o resolve. Esse problema pode ser, por exemplo, qual a probabilidade de uma moeda justa cair com a cara para cima em pelo menos dois lançamentos de três; qual a chance de pelo menos dois alunos em uma classe de trinta fazerem aniversário no mesmo dia; qual a probabilidade de uma única observação de uma distribuição normal cair no intervalo de x até $x + dx$. As escolas vão divergir principalmente no procedimento adotado para conectar essas experiências mentais com problemas vindos de observações reais.

Esse texto é resultado de uma investigação breve sobre a história dos conceitos de probabilidade, inferência, e incerteza, e sua relação mais geral com visões

alternativas de ciência. Ele foi motivado por uma percepção de que a vasta maioria dos textos introdutórios em estatística se foca em uma escola de pensamento, alinhada com as visões do autor, sem uma análise crítica aprofundada das suas relações complexas com as demais. A separação da estatística entre as escolas frequentista e bayesiana, muito repetida em textos introdutórios, esconde uma riqueza de detalhes, envolvendo contradições importantes dentro de cada escola e convergências nos pensamentos entre escolas diferentes. Essa separação também apaga a importância de definições que não se encaixam em nenhuma das alternativas, como a lógica probabilística de Carnap e o apelo à verossimilhança enquanto única base de inferência. A presente exposição é, por força, limitada e fragmentária; muitos pontos mais finos da argumentação foram deixados de fora para que o texto não se tornasse proibitivamente longo. É um lugar comum afirmar que cada seção aqui poderia ser expandida para um livro; algumas delas na verdade são esforços para resumir livros já escritos.

1.2 Contra uma teoria axiomática

O ensino moderno da probabilidade é, em larga escala, baseado na axiomatização proposta por Andrey Kolmogorov, matemático soviético muito prolífico em diversas áreas da ciência. Uma primeira visão dos axiomas da probabilidade, difundida em livros-textos e outros textos introdutórios, pode ser escrita como [Freedman et al., 2007; LeBlanc, 2004; Morettin and Bussab, 2009]:

Não-negatividade A probabilidade de um evento é um número real não-negativo: $P(E) \geq 0$.

Unitariedade A probabilidade de um evento certo é 1: $P(\Omega) = 1$.

Aditividade A probabilidade de eventos mutuamente exclusivos é aditiva: $P(U \cup V) = P(U) + P(V)$.

Uma abordagem puramente axiomática para a probabilidade, embora seja útil sob uma perspectiva estritamente matemática, como para resolver problemas resultantes do cálculo da probabilidade associada a um experimento mental, é insuficiente ao ser aplicada ao conhecimento científico de experimentos empíricos, por não levar a uma conclusão sobre no que *consiste* uma probabilidade. A proposição desses axiomas não leva a nenhuma conclusão sobre problemas reais: não há como conectar uma proposição teórica derivada dos axiomas da probabilidade a um problema prático sem assumir, em algum momento, um significado para a palavra probabilidade que não está contido nos axiomas. Uma posição que pode ser tentadora é presumir que probabilidade assume *qualquer* significado coerente com os axiomas: essa posição, no entanto, também falha por não restringir o tipo de proposição ao qual probabilidades podem ser associadas (veja a seção 1.13).

Outro problema da visão estritamente axiomática é que alguns problemas decorrentes da axiomatização podem passar despercebidos. Considere a troca do axioma

de aditividade pelo seguinte axioma de aditividade completa (ou σ-aditividade)[1]:

Aditividade completa Qualquer conjunto contável de eventos mutuamente exclusivos satisfaz:

$P(E_1 \cup E_2 \cup \cdots) = \sum_{i=1}^{\infty} P(E_i)$

Esse axioma é uma generalização natural do axioma de aditividade a conjuntos infinitos (contáveis) de eventos, e é amplamente aceito por algumas escolas de probabilidade. No entanto, tome um sorteio de um número racional aleatório entre 0 e 1, com distribuição uniforme. Já que existem infinitos números racionais em qualquer intervalo intervalo, a probabilidade de sortear cada número racional individualmente é zero. Por outro lado, a probabilidade de sortear *qualquer* número (ou seja, a probabilidade do evento certo), que é 1, pelo axioma da unitariedade, deve ser também 0, pois é a soma de infinitos termos iguais a zero.

Há outro ponto relevante contra uma interpretação puramente axiomática: nem toda quantidade que satisfaz os axiomas de Kolmogorov pode ser tratada como uma probabilidade. Massa e volume normalizados (ou seja, medidos em uma escala de 0 a 1) satisfazem os axiomas, mas não podem ser considerados probabilidades.

Para encontrar uma definição de probabilidade que permita conectar os axiomas a questões práticas do uso de probabilidades na ciência, vamos repassar a opinião de alguns dos grandes pensadores sobre o tema nas próximas seções.

1.3 Interpretação clássica

A teoria matemática da probabilidade remonta aos séculos XVI e XVII, nos quais Girolamo Cardano, Galileu Galilei, Blaise Pascal e Pierre Fermat desenvolveram métodos para resolver problemas envolvendo combinações de resultados em jogos de dados e outros jogos de azar.

Essa origem, juntamente com a facilidade e a universalidade de problemas envolvendo jogos de azar, explicam porque tantos textos introdutórios sobre estatística empregam exemplos envolvendo rolar dados e tirar cartas de baralhos. Embora a simplicidade destes exemplos ajude a transmitir os conceitos de probabilidade com facilidade, isso causa muitas vezes a sensação de que as regras da probabilidade servem apenas para os casos simples, e que estas não se aplicam, ou não se adequam totalmente aos problemas de natureza complexa com os quais a biologia e a ecologia têm de enfrentar. De fato, é muito difícil justificar que questões como "qual a chance de que uma população de jararaca-ilhoa entre em extinção nos próximos dez anos?" sejam embasadas nas mesmas ideias que "qual a chance de uma moeda jogada para cima dar coroa?". É mais natural tratar a primeira pergunta como um questionamento sobre o nosso conhecimento atual dos processos biológicos e da condição de vida dos indivíduos dessa espécie do que um questionamento sobre um processo físico simples que pode ser repetido um grande número de vezes.

[1] A axiomatização de Kolmogorov usa a aditividade completa, seguindo o trabalho anterior de Émile Borel. Para uma discussão das contribuições de Kolmogorov, Borel, Cantelli, Fréchet, Von Mises e muitos outros, veja [Shafer and Vovk, 2003]

O tamanho desta disparidade entre os exemplos simples e os problemas mais práticos faz com que muitas vezes textos sobre probabilidades empreguem definições conflitantes sobre probabilidades: às vezes, elas são usadas para representar algo pessoal e subjetivo, como um nível de confiança; outras vezes são apresentadas como a razão entre diferentes contagens.

A primeira tentativa influente de definir formalmente o conceito de probabilidade vem em 1814, com o trabalho de Pierre Simon de Laplace[2]:

> "La théorie des hasards consiste à réduire tous les évènemens du même genre, à un certain nombre de cas également possibles, c'est-à-dire tels que nous soyons également indécis sue leur existence, et à determiner le nombre de cas favorables à l'évènement dont on cherche la probabilité. Le rapport de ce nombre à celui de tous les cas possibles, est la mesure de cette probabilité"
>
> ("A teoria da probabilidade consiste em reduzir todos os eventos de um mesmo tipo a um certo número de casos igualmente possíveis, isso é, tal que sejamos igualmente indecisos sobre a sua ocorrência, e a determinar o número de casos favoráveis ao evento ao qual buscamos a probabilidade. A razão deste número para o número de todos os casos possíveis é a medida desta probabilidade") [Laplace, 1814]

Para Laplace, todos os eventos que presenciamos seriam resultados de leis físicas imutáveis, e uma inteligência superior, dotada do conhecimento do estado do universo em um dado instante, poderia prever todos os eventos futuros sem qualquer incerteza. No entanto, nosso conhecimento limitado, tanto do estado do universo quanto das leis que o regem, faz com que não possamos fazer previsões para um grande número de sistemas. O estudo das probabilidades se coloca, então, como um apoio ao nosso poder de realizar previsões sobre o universo, enquanto não possuímos o conhecimento das leis e estados necessária para realizar previsões certeiras.

Uma dificuldade com essa definição de Laplace é que a demarcação de "eventos igualmente prováveis" deve ser feita com base em argumentos que não partam da ideia de probabilidade; caso contrário, podemos incorrer em raciocínios circulares. Por exemplo, se uma moeda tem 2 lados perfeitamente simétricos, e a lançamos de forma a não privilegiar um dos lados, podemos dizer que o número de eventos totais é dois, e o número de eventos favoráveis a tirar coroa é um. Não há qualquer motivo para supor que um dos lados seja mais propenso a cair para cima que o outro, visto a moeda ser perfeitamente simétrica; logo, a probabilidade de tirar coroa em um lançamento é de $\frac{1}{2}$. Aqui, o ato de atribuir iguais chances aos dois lados da moeda vem de um argumento físico de simetria. Isso é dizer que, para Laplace, a ciência da probabilidade deve se basear em leis físicas do mundo natural, e não no nosso estado presente de conhecimento sobre o mundo. A probabilidade é, desta forma, algo que existe objetivamente. Esta interpretação faz com que as

[2]Há uma série de trabalhos mais antigos na área, mas estes ou se preocupavam mais com questões operacionais do cálculo da probabilidade, como o *Ars conjectandi* de Jacques Bernoulli (1713) ou não tiveram grande influência na comunidade científica, como o *Liber de ludo aleae* de Cardano, escrito em 1525 e publicado apenas em 1663.

leis da probabilidade, estritamente, só possam ser aplicadas para sistemas simples, para os quais leis simples e com poucas premissas possam ser formuladas, como é o caso de baralhos bem embaralhados ou dados perfeitos[3]. Laplace, no entanto, não parece ter levado essa definição à suas últimas consequências, já que ele discute no tratado citado acima problemas referentes a fazer inferências sobre testemunhos judiciais, nos quais a testemunha poderia mentir, com uma certa probabilidade, ou ter se equivocado, com outra probabilidade. No entanto, nenhuma explicação convincente é dada sobre como medir essas probabilidades dentro do paradigma clássico.

Presumindo que as probabilidades estão corretas, o método usado por Laplace é conhecido como "método das probabilidades inversas", e tem uma alta importância na história da estatística. Deve-se notar que o argumento de Laplace se assemelha ao raciocínio desenvolvido pelo rev. Thomas Bayes em seu trabalho publicado postumamente em 1763. Esse trabalho não será discutido a fundo aqui, por ser muito convoluto em seus detalhes e notação, e não oferecer uma visão clara sobre a natureza das probabilidades. No entanto, vamos considerar o impacto que esse trabalho teve sobre o pensamento estatístico na seção 1.10.

1.4 Interpretação subjetivista

Uma visão alternativa é dada por Augustus de Morgan, conhecido principalmente por suas leis em lógica proposicional, no seu tratado Formal Logic, de 1847. Para de Morgan, a única certeza que podemos ter é a de nossa própria existência. Este conhecimento não é passível de ser refutado. No entanto, qualquer outra proposição feita deve ser acompanhada por um *grau de conhecimento* subjetivo:

> "It will be found that, frame what circumstances we may, we cannot invent a case of purely objective probability. I put ten white balls and ten black ones into an urn, and lock the door of the room. I may feel well assured that, when I unlock the room again, and draw a ball, I am justified in saying it is an even chance that it will be a white one. If all the metaphysicians who ever wrote on probability were to witness the trial, they would, each in his own sense and manner, hold me right in my assertion. But how many things there are to be taken for granted! Do my eyes still distinguish colours as before? Some persons never do, and eyes alter with age. Has the black paint melted, and blackened the white balls? Has any one else possessed a key of the room, or got in at the window, and changed the balls? We may be *very sure*, as those words are commonly used, that none of these things have happened, and it may turn out (and I have no doubt will do so, if the reader try the circumstances) that the ten white and ten black balls will be found, as distinguishable as ever, and unchanged. But for all that, there is much to be assumed in reckoning upon such a result, which is not so objective

[3]Considere, por exemplo, uma moeda viciada que dá coroas com 75% de frequência. A definição de probabilidade de Laplace ainda sugere que a probabilidade de coroa desta moeda deve ser um caso sobre dois casos possíveis, ou seja, $\frac{1}{2}$.

> (in the sense in which I used the word) as the knowledge of what the balls were when they were put into the urn." [de Morgan, 1847]

A teoria de probabilidades segundo de Morgan, portanto, lida com graus de conhecimento subjetivos. Assim, ao perguntar para uma pessoa comum qual é a chance de que um lançamento de moeda resulte em cara, essa pessoa pode responder $\frac{1}{2}$, e esta é a medida de probabilidade correta, dada a informação que ele possui sobre o problema. Uma outra pessoa, que sabe que esta moeda é viciada, ou que é capaz de jogar a moeda de forma a privilegiar um resultado, pode dar outra resposta completamente diferente, e ainda assim estará correta. Para de Morgan, o passo essencial na construção de uma probabilidade consiste em *medir* o grau de certeza que temos em uma proposição. Algumas proposições são fáceis de medir, como "dois mais dois são quatro", "um lançamento de uma moeda justa vai dar coroa", e "dois mais dois são cinco", aos quais podemos atribuir facilmente as probabilidades de 1, $\frac{1}{2}$ e 0. Essas medidas, juntamente com o senso de que algum evento é mais ou menos provável que outro, podem ser usados para construir uma escala subjetiva de probabilidades. Como essa visão sobre probabilidades parte exclusivamente do conhecimento que temos, e não de uma realidade objetiva, a probabilidade que queremos determinar não é algo que exista à parte no mundo; ela existe somente enquanto descrição de um processo mental.

Um empecilho ao uso dessa interpretação de probabilidades se dá quando percebemos que, quanto mais complexo é o problema que queremos resolver, mais imperfeita se torna nossa sensação de conseguir ordenar a probabilidade referente a diferentes resultados de um experimento. É fácil apreender que, se eu jogar duas moedas para cima, a chance de que ambas resultem em cara é menor do que a chance de que um único lançamento resulte em cara. É consideravelmente mais difícil encaixar nessa ordenação a frase "ao escolher uma letra aleatória da obra completa de Shakespeare, ela será E ou A"; mas esse caso é passível de uma análise exaustiva. No entanto, questões como "qual a probabilidade de que o signo zodiacal de uma pessoa afete suas características, como sociabilidade ou perseverança?", embora perfeitamente válidas para a visão de de Morgan, apresentam grande dificuldade para serem respondidas nesse paradigma.

A visão subjetiva de probabilidades permite ainda que frases cotidianas, como "é provável que chova hoje" ou "existe uma grande chance de encontrarmos o José hoje", sejam tratadas pelo mesmo ponto de vista que os jogos de azar. Mas embora essa interpretação seja de grande utilidade por permitir a generalização do conceito de probabilidade para problemas mais gerais do que os sistemas físicos simples contemplados pela interpretação clássica, ela não garante que o valor encontrado para a probabilidade de um evento seja coerente para diversas pessoas. Isso irá embasar a crítica de que uma visão subjetivista de probabilidade pode ser perfeitamente adequada para embasar uma decisão pessoal, mas falhar como uma ferramenta de avanço da ciência, entendida como um corpo teórico objetivo e independente das opiniões de cada cientista individualmente.

1.5 Interpretação frequentista

A próxima crítica à visão de Laplace vem de uma série de pensadores dentre os quais se destacam George Boole e John Venn. Este último, em seu trabalho de 1866, postula que a probabilidade de um evento se refere à frequência relativa com a qual esse evento se repetiria em uma série infinita de experimentos, pelo que sua visão será conhecida como frequentista:

> "We may define the probability or chance (the terms are here regarded as synonymous) of the event happening in that particular way as the numerical fraction which represents the proportion between the two different classes in the long run (...). This assumes that the series are of indefinite extent, and of the kind which we have described as possessing a fixed type. If this be not the case, but the series be supposed terminable, or regularly or irregularly fluctuating, (...) the series ceases to be a subject of science. What we have to do under these circumstances, is to substitute a series of the right kind for the inappropriate one presented by nature, choosing it, of course, with as little deflection as possible from the observed facts. This is nothing more than has to be done, and invariably is done, whenever natural objects are made subjects of strict science." [Venn, 1866]

A probabilidade de um dado evento poderia então ser definida como o limite da razão entre o número de eventos favoráveis sobre o número de experimentos realizados, com o número de experimentos realizados indo para o infinito. Com n_x sendo o número de eventos favoráveis e n_t sendo o número de experimentos realizados:

$$P(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t} \tag{1.1}$$

Venn também critica a visão subjetivista de de Morgan, sob o argumento de que uma avaliação individual está sujeita a um enorme número de fontes de erro:

> "Our conviction generally rests upon a sort of chaotic basis composed of an infinite number of inferences and analogies of every description, and there moreover distorted by our state of feeling at the time, dimmed by the degree of our recollection of them afterwards, and probably received from time to time with varying force according to the way in which they happen to combine in our consciousness at the moment."[Venn, 1866]

Venn propõe um conceito de probabilidade que é independente da pessoa que vai medir a probabilidade. A probabilidade de um experimento físico será a mesma para qualquer pesquisador que realizar os mesmos passos para medi-la. Isso tem consequências para o uso da probabilidade na ciência e também na moral: convém lembrar que um dos casos em que Laplace e Bernoulli empregam probabilidades é no apoio ao julgamento de réus. Uma medida de probabilidade que independa de quem a fornece vai fornecer um peso argumentativo maior durante um julgamento do que

uma probabilidade que está necessariamente atrelada ao estado de conhecimento de um indivíduo particular.

A definição frequentista leva a um problema claro: não temos acesso a infinitas experiências para determinar o valor de uma dada probabilidade. A série infinita postulada por Venn precisa, para qualquer fim prático, ser identificada com uma realização parcial observável. No entanto, essa realização parcial pode ser identificada com infinitas séries diferentes, algumas das quais convergentes, outras divergentes. O pesquisador frequentista precisa, assim, tomar uma decisão pessoal sobre qual série utilizar. Ao embasar esse modelo em princípios físicos (dois lados de uma moeda caem com igual chance), ele estará se aproximando da concepção física de Laplace; ao escolher qualquer outra série, ele estará transferindo a subjetividade da medida de probabilidade para a escolha do modelo.

Nenhum critério objetivo garante que uma destas séries potenciais é aquela com "menor desvio possível em relação aos fatos observados"; a Lei dos Grandes Números [Loève, 1977] vai embasar uma determinada escolha, que a média amostral de uma variável aleatória converge quase certamente para sua esperança. Embora essa lei forneça uma justificativa para pensar na probabilidade como uma média a longo prazo, é importante frisar que ela não implica a interpretação frequentista: a atualização de crenças Bayesiana vai levar ao mesmo resultado com amostras suficientemente grandes.

Mesmo quando temos acesso a um grande conjunto de dados para estimar o valor de uma determinada probabilidade, esse conjunto pode ser suficientemente heterogêneo para que nossa conclusão se torne incorreta. Como escreve Venn,

> "At the present time the average duration of life in England may be, say, forty years; but a century ago it was decidedly less; several centuries ago it was presumably very much less (...). Let us assume that the regularity is fixed and permanent. It is making a hypothesis which may not be altogether consistent with fact, but which is forced upon us for the purpose of securing precision of statement and definition."[Venn, 1866]

Desta forma, o pesquisador que decide medir uma probabilidade à partir de dados reais precisa, necessariamente, supor que haja certa regularidade no fenômeno medido dentro da escala de tempo e espaço na qual o estudo é levado a cabo. Essa suposição não deixa de ser uma idealização do mundo real: ou seja, para a visão frequentista, a probabilidade só corresponde a uma entidade tangível e objetiva dentro de um mundo idealizado no qual seja possível realizar exatamente o mesmo experimento repetidas vezes.

Essa percepção levará críticos do frequentismo a considerar que essa visão não é tão objetiva quanto se pretende, ao que os frequentistas respondem que a abstração é inescapável na ciência. Nas palavras de Edwards,

> "Though the notion of a random choice is not devoid of philosophical difficulties, I have a fairly clear idea of what I mean by 'drawing a card at random'. That the population may exist only in the mind, an abstraction possibly infinite in extent, raises no more (and no less) alarm than the infinite straight line of Euclid. I am only prepared to

> use probability to describe a situation if there is an analogy between the situation and the concept of random choice from a defined population." [Edwards, 1972]

Essa postura é coerente com a visão frequentista defendida por Richard von Mises, irmão do economista Ludwig von Mises, que procede à axiomatização dos chamados coletivos, ou *Kollektivs*, sequências de tamanho infinito obedecendo fortes condições de convergência[Von Mises, 1941]. Para von Mises, a ciência da probabilidade deve identificar situações nas quais essas condições de convergência se aplicam: jogos de azar e física teórica, por exemplo. O cálculo das probabilidades nestas situações específicas poderia embasar, com um certo grau de confiança, a aplicação de probabilidades para casos análogos. O trabalho de von Mises, embora mal recebido no momento da sua publicação original, foi instrumental para a axiomatização de Kolmogorov[Shafer and Vovk, 2003], e tem recebido prestígio por autores contemporâneos.

Para finalizar a seção, a concepção frequentista, apesar de ter pontos de contato com o pensamento laplaciano, apresenta uma ruptura com a metodologia de probabilidades inversas. Para um frequentista, uma probabilidade é algo objetivo e mensurável, de forma que não faz sentido atribuir uma probabilidade a uma hipótese. Tampouco faz sentido tratar uma hipótese estatística como uma sentença escolhida aleatoriamente de uma população de sentenças, algumas das quais são verdadeiras. Desta forma, a escolha da definição frequentista de probabilidade implica abandonar o método das probabilidades inversas para comparar o mérito de diferentes hipóteses. Essa percepção vai levar ao desenvolvimento da escola de testes de significância e à proposição do conceito de verossimilhança, como será descrito na seção 1.7.

1.6 Interpretação por propensões

Uma terceira interpretação para a natureza da probabilidade é dada pelo cientista e lógico norte-americano C.S. Peirce, cuja obra prolífica se estende sobre vários ramos da estatística moderna. O interesse de Peirce no estudo de probabilidades deve ser traçado na sua visão indeterminista, segundo a qual eventos não se sucedem de maneira estritamente causal, mas são influenciados por um elemento de chance. Peirce baseia essa visão na constatação de que "the mind is not subject to 'law' in the same rigid sense that matter is (...). There always remains a certain amount of arbitrary spontaneity in its action" [Peirce, 1892].

A oposição de Peirce a às outras escolas filosóficas ocorre, então, em um nível mais fundamental do que dicotomia entre subjetivistas e frequentistas: enquanto para os autores anteriores, notadamente Laplace, a realidade física é determinística, para Peirce a própria realidade é aleatória. Enquanto Laplace propõe uma visão de ciência que visa aproximar nosso grau de conhecimento da onisciência - quando então a probabilidade deixaria de existir - Peirce abandona esse programa em favor de uma ciência que deve lidar constantemente com a incerteza.

Se a experiência no mundo natural é intrinsecamente atormentada pela sorte, ela não pode ser base para um sistema de conhecimento baseado estritamente na

dedução. Portanto Peirce irá desenvolver um sistema de argumentação baseada em probabilidades, incluindo o argumento de Dedução Provável Simples [Fetzer, 1993]:

1. Uma proporção p de As são Bs
2. X é um A
3. Portanto, X é um B com probabilidade p

Peirce é certamente um objetivista, rejeitando qualquer utilidade para probabilidades pessoais; mas sua visão sobre a conexão entre o mundo físico e a probabilidade é uma questão que o afastará da visão frequentista.

Peirce ataca ainda o argumento frequentista segundo o qual a probabilidade é definida como sendo a frequência de ocorrência porque, em larga escala, os dois coincidem. Para Peirce, a concordância entre essas duas medidas, que ocorre em uma escala muito grande, não pode ser usada como argumento para *definir* probabilidade como uma frequência em todos os casos. Para contrastar as duas visões, note que, para um frequentista, a probabilidade de um evento específico não pode ser definida. "Qual a probabilidade de que um lance qualquer de moeda dê cara?" é uma pergunta que tem resposta no paradigma frequentista, enquanto "qual a probabilidade de que *o próximo lance de moeda que eu fizer* dê cara?" (e qualquer outra pergunta sobre probabilidades de caso único) é uma pergunta que não pode ser respondida, já que a probabilidade só é definida sobre uma série infinita de experimentos. Para poder responder perguntas como esta, Peirce, lança mão de uma propriedade física intrínseca descrita como "would-be":

> "I am, then, to define the meanings of the statement that the *probability* that if a dice be thrown from a dice box it will turn up a number divisible by three, is one third. The statement means that the dice has a certain 'would-be'; and to say that it has a certain 'would-be' is to say that it has a property, quite analogous to any *habit* that a man might have. Only the 'would-be' of the dice is presumable as much simpler and more definite than the man's habit as the dice's homogeneous composition and cubical shape is simpler than the nature of the man's nervous system and soul." (C. S. Peirce, CP 2.664, in [Fetzer, 1993])

A interpretação de "would-bes" representa uma quebra fundamental com a interpretação frequentista, sem deixar de caracterizar probabilidade como uma propriedade objetiva, e independente de um observador subjetivo. C. S. Peirce trata o "would-be" como uma característica primitiva, que será indiretamente conectada com a experiência. Aceitar a existência de "would-bes", embora possa causar espanto ao cientista moderno, acostumado com a visão frequentista, não é diferente de aceitar outras propriedades primitivas presentes nas teorias físicas, como massa, carga elétrica ou *spin*.

Peirce também é um pioneiro no desenho de experimentos, dando grande ênfase à criação de desenhos amostrais aleatorizados. Nessa linha, ele vai argumentar contra o uso de ferramentas estatísticas desenvolvidas sob hipótese de randomização em estudos que não a apresentem, como estudos observacionais [Stigler, 1978].

Uma interpretação mais recente, de surpreendente similaridade com a dada por Peirce, é proposta independentemente por Karl Popper, sob o nome de *propensão* [Popper, 1959], o que irá fazer com que filósofos contemporâneos passem a se referir à interpretação de Peirce pelo mesmo nome [Miller, 1975]. A visão de probabilidades como propensões, advogada por Peirce ou Popper é, no entanto, criticada por não prover nenhuma base operacional para o cálculo de probabilidades à partir da definição de propensidades[Hájek, 2012]. É inconclusivo se esta visão pode ser considerada uma teoria de interpretação de probabilidades, pois uma interpretação deve possuir uma determinada estrutura, como cita Peter Milne:

> "Interpretations in the literal sense (...) have sufficient intrinsic mathematical structure that one can derive the characteristics of probability from them."[Milne, 1993]

Por outro lado, autores mais recentes como Donald Gillies e J.H. Fetzer têm advogado o uso do termo "propensão" para qualquer teoria de probabilidades objetiva e não baseada em frequências observadas. Desta forma, embora a visão de propensões tenha recebido pouca atenção em meios científicos ela tem tido novo interesse nos últimos anos em círculos filosóficos [Gillies, 2000]. O interesse continuado em interpretações de probabilidade que permitam a probabilidade do caso único está, desde o trabalho de Popper, ligado ao estudo da física quântica, o que é natural por dois motivos. O primeiro é que há evidências de que a natureza da física quântica é intrinsecamente probabilística [Gudder, 1988], ao contrário de fenômenos macroscópicos para os quais a descrição probabilística pode ser vista como uma aproximação dada a falta de informações completas. O segundo é que o estudo de problemas quânticos leva naturalmente a estruturas matemáticas semelhantes às probabilidades, mas com propriedades distintas, como as quasi-probabilidades, úteis no estudo de ótica quântica, as quais podem não satisfazer o axioma de aditividade de Kolmogorov e ter regiões de probabilidade negativa [Mandel and Wolf, 1995].

1.7 A formalização da inferência

A escola frequentista será de grande influência para a formulação da teoria de testes de hipóteses de R. A. Fisher, que se lança no começo da década de 1920 sobre o problema de formalizar a inferência estatística, ou seja, o processo de extrair conclusões a partir de dados obtidos em uma amostra. Baseando-se nos trabalhos de Karl Pearson e William Sealy Gosset (conhecido por seu pseudônimo Student), Fisher realiza um passo fundamental dessa formalização diferenciando os conceitos de população e amostra:

> "It is customary to apply the same name, *mean*, *standard deviation*, *correlation coefficient*, etc., both to the true value which we should like to know, but can only estimate, and to the particular value at which we happen to arrive by our methods of estimation."[Fisher, 1922]

Nesta perspectiva, Fisher propõe que precisamos diferenciar uma *população*, hipoteticamente infinita, na qual os objetos se dividam em duas classes de acordo

com uma característica de interesse, de uma *amostra* finita, que será nossa base para tirar conclusões. Por exemplo, examinando o lançamento de um dado, a população de interesse se constitui numa sequência infinita de rolagens de dado, na qual distinguimos uma característica, como o fato de uma rolagem resultar no número 5. O valor da probabilidade de que uma rolagem resulte em 5 *na população* pode ser visto como um parâmetro da distribuição teórica de resultados para essa população infinita, portanto trata-se de um número fixo, embora a princípio desconhecido; já a proporção de lançamentos realizados que resulta em 5 *em uma dada amostra* é um estimador desse parâmetro, a partir da amostra, e seu valor está sujeito a variações conforme repetimos a amostragem diversas vezes. Qualquer função calculada sobre uma dada amostra é conhecida então como uma *estatística* da amostra. A população idealizada corresponde a um *modelo* da realidade, descrito por uma função matemática. Assim, Fisher explicita o fato de que a inferência estatística, em sua visão, depende da construção de um modelo matemático abstrato.

Em seu trabalho de 1922, Fisher divide a tarefa da inferência estatística em 3 classes de problemas: problemas de especificação, ou seja, a determinação da função matemática que define a população; problemas de estimação, ou seja, o cálculo de estatísticas sobre a amostra que estimem corretamente os parâmetros populacionais; e problemas de distribuição, ou seja, estimações realizadas sobre a distribuição das estatísticas calculadas sobre a amostra. Fisher discorre brevemente sobre o problema de especificação, se concentrando nos problemas de estimação e distribuição. Sobre especificação, ele escreve:

> "We may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested *a posteriori* [by an objective criterion of goodness of fit]. For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts."[Fisher, 1922]

A economia de palavras sobre especificação não deve sugerir que este é um problema de menor importância: pelo contrário, uma especificação incorreta vai inutilizar toda análise posterior. Fisher, no entanto, se debruça neste trabalho sobre uma teoria matemática, e parece reconhecer que as escolhas que levam à *escolha* de um modelo requerem deliberações que extrapolam suas fronteiras.

Presumindo uma escolha adequada de modelo, ou seja, de função matemática que descreva as propriedades da população, a tarefa da inferência estatística se concentra em estudar o pequeno número de parâmetros desse modelo que a descrevem completamente. Esse estudo corresponde aos problemas de estimação e distribuição.

Fisher descreve as seguintes propriedades desejáveis para um estimador. Embora as suas definições tenham sido melhor formalizadas posteriormente, as ideias presentes ainda são de grande valia:

1. Consistência: "when applied to the whole population the derived statistic should be equal to the parameter"

2. Eficiência: "that statistic is to be chosen which has the least probable error"

3. Suficiência: "the statistic chosen should summarize the whole of the relevant information supplied by the sample"

É importante frisar que estas definições se dão em relação a um único problema de estimação. Quando coletamos uma amostra de uma população, de distribuição normal e variância conhecida e perguntamos "qual é o melhor estimador para a média populacional", a média amostral se mostra uma estatística suficiente. Por outro lado, se a distribuição da qual estamos amostrando é uma distribuição gama, de parâmetros α e β desconhecidos, a média amostral *não é* uma estatística suficiente. Desta forma, não é possível responder a pergunta "a média amostral é uma estatística suficiente?" sem a especificação do modelo que está sendo utilizado.

Para este estudo, no entanto, o ponto mais importante do trabalho de Fisher é a descrição de uma quantidade conhecida como verossimilhança (likelihood), que será fundamental para contrastar as teorias de inferência. Embora a verossimilhança já tivesse sido identificada por C.S. Peirce [Peirce, 1883], a exposição de Fisher sobre ela é considerada um passo fundamental para sua aceitação entre estatísticos europeus.

Fisher reconhece a necessidade de definir uma nova quantidade - que não seja uma probabilidade - para identificar dentre um conjunto de hipóteses, quais encontram maior suporte das evidências:

> "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."[Fisher, 1922]

Para melhor compreender essa grandeza, vamos usar um exemplo citado por Laplace [Laplace, 1814], que se ocupa da proporção de batismos (usado para representar o número de nascimentos) de meninos e meninas em Paris nos anos de 1745 a 1784. Laplace tem a hipótese, corroborada por dados coletados em diversas outras cidades, de que nascem mais meninos do que meninas em uma proporção de 22 para 21. Coletando a informação de que houve 393386 nascimentos de meninos e 377555 nascimentos de meninas no período, e presumindo o modelo binomial para os nascimentos, a verossimilhança da hipótese levantada por Laplace é dada pela probabilidade atribuída ao valor 393386 pelo modelo binomial com tamanho $393386 + 377555 = 770941$ e parâmetro $22/(21 + 22) \approx 0.511$. Essa probabilidade é um número da ordem de 10^{-5}. Ao calcular a verossimilhança da hipótese de que não há preponderância de nenhum gênero, ou seja, a probabilidade de sucesso é de 0.5, encontramos o valor de 10^{-74}, muitas ordens de grandeza inferior ao anterior. Portanto, os dados indicados por Laplace fornecem uma evidência mais forte para a hipótese de que os nascimentos de meninos são mais prováveis na proporção de 22 para 21.

Laplace não formaliza esse argumento em seu livro, sem dúvida pela dificuldade de realizar as contas indicadas acima sem a ajuda de computadores. Mas é importante notar que, enquanto Laplace poderia desenvolver esse raciocínio em termos das *probabilidades* das hipóteses conflitantes, Fisher nota que as verossimilhanças calculadas acima não se comportam como probabilidades, no sentido de não estarem sujeitas às leis da probabilidade (veja na seção 1.2).

1.8 Abordagens para testes de hipóteses

A estrutura mental desenvolvida por Fisher durante a década de 1920 vai levar também a um procedimento conhecido como teste de significância, e que será alvo de uma intensa controvérsia entre ele e a dupla formada por Jerzy Neyman e Egon Pearson, durante a década de 1930. Em uma terminologia que vai influenciar muito do pensamento em inferência estatística do século XX, Fisher usa as expressões *modelo estatístico* como uma função que descreve a população de interesse, sendo definida por um conjunto de parâmetros; e *hipótese* como uma afirmação a respeito do valor desses parâmetros. O modelo contém, assim, todas as suposições sobre o problema a respeito das quais os dados não irão discriminar, enquanto a hipótese contém a afirmação que será julgada pela aplicação do teste. Dada uma hipótese, é possível calcular a distribuição que uma certa estatística teria, se um mesmo experimento fosse repetido infinitas vezes.

O procedimento de teste de significância, segundo Fisher, requer a escolha de um modelo apropriado e a especificação de uma hipótese (conhecida como hipótese nula) para a qual é possível calcular a distribuição da estatística de interesse (veja detalhes no exemplo abaixo). Após realizar um experimento, a estatística calculada sobre a amostra é comparada com a distribuição teórica sob a hipótese nula, e é calculada a probabilidade do desvio entre a estatística calculada e a sua esperança ser *tão grande ou maior* do que o encontrado.

Em um exemplo simples, queremos saber se uma moeda é justa ou não lançando-a 20 vezes. É improvável que o número de caras seja exatamente 10, mesmo para uma moeda justa. Como o resultado de cada lançamento é uma de duas respostas (cara ou coroa), e a probabilidade de cada lançamento é suposta constante e independente, o modelo escolhido para representar o problema pode ser uma binomial (Determinar que o modelo binomial é adequado corresponde ao problema da especificação, segundo Fisher). A hipótese nula é a de que o parâmetro da binomial é de 0.5. Nesse caso, podemos calcular a distribuição esperada do número de caras, e veremos que em 82% dos experimentos realizados, a diferença entre o valor esperado de dez caras e o resultado obtido será de 1 ou mais; em 26%, de 3 ou mais; e em 4%, de 5 ou mais. Ao examinar o resultado de um experimento, podemos então *rejeitar* a hipótese nula, ou seja, entendemos que a moeda não é justa, ou *deixar de rejeitá-la*, ou seja, não encontramos evidência suficiente de que esta hipótese está errada.

Neyman e Pearson fazem uma crítica ao método descrito por Fisher, visto por eles como incompleto, e estabelecem uma metodologia para escolha entre diversas hipóteses concorrentes, muitas vezes sumarizadas em duas: a nula e a alternativa. Um procedimento de teste de significância no qual uma única hipótese sobre o valor de um parâmetro é privilegiada em detrimento de todo o espectro de outras respostas possíveis é falho. Como o valor de parâmetros é uma variável contínua, Neyman e Pearson argumentam (numa linha muito semelhante à usada por argumentos Bayesianos):

> "If x is a continuous variable, then any value of x is a singularity of relative probability equal to zero. We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory

> of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis." [Neyman and Pearson, 1933]

Neyman e Pearson propõem um procedimento formal para delimitar quando devemos aceitar e quando devemos rejeitar a hipótese nula em favor de uma hipótese alternativa, dividindo o resultado desse processo em quatro classes: podemos aceitar uma hipótese verdadeira ou falsa, ou podemos rejeitar uma hipótese verdadeira ou falsa. O ato de rejeitar a hipótese nula verdadeira é conhecido como erro de tipo I, e o ato de aceitar uma hipótese nula falsa é conhecido como erro de tipo II. O conceito do erro de tipo I encontra paralelo na teoria de Fisher, enquanto o erro de tipo II é ausente, e será contestado por Fisher em diversas bases. Esse procedimento consiste na construção prévia de uma região crítica, tal que, se os dados observados se encontrarem nesta região, a hipótese nula é rejeitada em favor da alternativa; caso contrário, a hipótese nula é aceita.

A construção dessa região crítica é feita de forma a tolerar um valor para a probabilidade de erro de tipo I, normalmente em 0.05. Um resultado central do trabalho de Neyman e Pearson foi a demonstração de que, entre todas as possíveis regiões críticas, a região ótima, no sentido de minimizar a probabilidade de erro de tipo II, é aquela construída pela razão entre as verossimilhanças [Neyman and Pearson, 1933].

Convém notar que esse resultado é verdadeiro apenas para hipóteses ditas simples, ou seja, hipóteses que atribuem um único valor de probabilidade para cada possível observação. Enquanto Neyman e Pearson obtêm resultados parciais para testes envolvendo hipóteses compostas, outros autores vão eliminar essas hipóteses das suas considerações. Edwards, por exemplo, afirma que:

> "The class of hypotheses we call 'statistical' is not necessarily closed with respect to the logical operations of alternation ('or') and negation ('not'). For a hypothesis resulting from either of these operations is likely to be composite, and composite hypotheses do not have well-defined statistical consequences, because the probabilities of occurrence of the component simple hypotheses are undefined. For example, if p is the parameter of a binomial model, about which inferences are to be made from some particular binomial results, '$p = \frac{1}{2}$' is a statistical hypothesis, because its consequences are well-defined in probability terms, but its negation, '$p \neq \frac{1}{2}$', is not a statistical hypothesis, its consequences being ill-defined."[Edwards, 1972]

Fisher critica duramente o procedimento de Neyman-Pearson por uma série de motivos, entre os quais o engessamento do procedimento, que levaria a uma aceitação ou rejeição de hipóteses acriticamente:

> "It is a fallacy (...) to conclude from a test of significance that the null hypothesis is thereby established (...). In an acceptance procedure, on the other hand, acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid down in advance; no *thought* is given to the particular case, and the tester's

> state of mind, or his capacity for *learning*, is inoperant. By contrast, the conclusions drawn by a scientific worker from a test of significance are *provisional*, and involve an intelligent attempt to *understand* the experimental situation."[Fisher, 1955]

É curioso que um autor que tenha atacado tão intensamente os conceitos de probabilidade subjetiva seja também tão crítico de uma abordagem que visa minimizar as decisões subjetivas a serem tomadas por meio de uma estruturação rígida do procedimento de teste de hipóteses.

Enquanto o debate entre Fisher, Neyman e Pearson dividia os proponentes da inferência frequentista, a visão subjetivista da probabilidade foi gradualmente abandonada em círculos científicos. O interesse nessa escola de pensamento irá retornar à partir da década de 1960, com o trabalho dos chamados Bayesianos, mas essa linha de pensamento tem pontos importantes de contato com uma escola independente: a probabilidade necessária de W.E. Johnson e Rudolf Carnap.

1.9 Lógica indutiva

Tanto de Morgan como Boole, contados entre os fundadores da interpretação subjetivista e frequentista da probabilidade, se ocuparam do estudo da probabilidade como um subconjunto da lógica formal, que constituía um interesse mais amplo para ambos. No mesmo espírito, o trabalho de Rudolph Carnap, à partir da década de 1940, vai ligar novamente o estudo da probabilidade às suas raízes dentro da lógica. A notação usada originalmente por Carnap favoreceu sua comunicação com filósofos e logicistas, mas dificultou seu acesso por estatísticos, de forma que vamos adaptar alguns termos e notações seguindo textos mais modernos ([Zabell, 2009] e [Fitelson, 2007]), que combinam a notação de Carnap com a de W.E. Johnson, que forneceu resultados semelhantes de forma independente 20 anos antes de Carnap.

Para esta linha de pensadores, a questão da definição da palavra "probabilidade" não deve ser a de decidir quais acepções dessa palavra são corretas ou incorretas, mas sim quais são satisfatórias, e se existe alguma definição mais satisfatória que as demais. A pergunta "o que é a probabilidade?" vai receber diferentes respostas dependendo de para quem for formulada, assim como a pergunta "baleias são peixes?" recebia uma resposta afirmativa em tempos bíblicos e recebe uma negativa de cientistas modernos. É equivocado pensar que os biólogos "descobriram que baleias não são peixes": o que mudou foi o próprio conceito de peixe.

Carnap, primeiramente, reconhece que a palavra "probabilidade" é usada em dois grupos diferentes de significados, e portanto separa os conceitos de *probabilidade*$_1$ e *probabilidade*$_2$, o primeiro se referindo a uma medida de confirmação, enquanto o segundo é uma medida de frequência. Enquanto os frequentistas identificam os dois conceitos como se referindo ao mesmo objeto, outras escolas de pensamento vão conectar esses conceitos de forma menos direta.

Seguindo o trabalho de Wittgenstein, Carnap trabalha com a *probabilidade condicional* de uma proposição lógica H dada uma proposição E, $c(H, E)$, e interpreta essa grandeza como a medida com a qual a evidência E corrobora uma hipótese

H [Zabell, 2009][4]. Essa visão será desenvolvida de forma a construir uma visão de probabilidade como uma necessidade lógica:

> "The truth conditions for a probability statement are logical or semantic: they are built into the language in the same way that the truth conditions for logical entailment are built into the language. Given the statement 'It will rain tomorrow', and given a body of evidence E, there is exactly one real number r, determined by logical conditions alone, such that the probability of 'It will rain tomorrow', relative to the body of evidence E, is r." [Kyburg, 1974]

Carnap provê ainda uma distinção entre três formas de confirmação: classificatória (a evidência confirma ou não a hipótese), quantitativa (com qual intensidade a evidência suporta a hipótese) e relacional (a evidência confirma uma hipótese mais do que outra)[Carnap, 1962]. Vamos retornar a essa tipologia de confirmação na seção 1.14 para contrastar a abordagem de verossimilhança com a abordagem Bayesiana.

O argumento de Johnson e Carnap é que observações passadas de um processo multinomial embasam a probabilidade de observações futuras, baseados nos postulados da permutação e da combinação. Assim, temos uma base para regras de sucessão como a desenvolvida por Laplace sem a necessidade de invocar uma ignorância *a priori* sobre o valor de um parâmetro. No entanto, a aplicação desses princípios para problemas gerais em inferência necessita de uma extensão desses resultados para considerar não apenas observações futuras idênticas, mas observações *similares*. A definição dessa similaridade e seu uso operacional para problemas de inferência levam a problemas matemáticos extremamente complexos, e uma teoria geral de inferência baseada em probabilidades necessárias ainda não foi alcançada (veja Zabell [2009] para um histórico mais completo).

Ainda, o uso de uma medida de confirmação para embasar a construção de conhecimento será contestada por filósofos da ciência, a começar por Karl Popper [Popper, 1963]. Popper aponta que algumas teorias pseudocientíficas, como a astrologia, podem ser confirmadas pela experiência, mas nunca falsificadas. Ele propõe, então, que apenas a falsificação de hipóteses pode ser utilizada para o avanço do conhecimento científico.

1.10 Probabilidade Bayesiana moderna

Embora uma grande parcela da comunidade científica tenha adotado uma visão frequentista durante a primeira metade do século XX, tomando algum dos lados do debate entre Fisher e Neyman-Pearson, muitos autores continuaram seguindo a interpretação subjetivista da probabilidade, mais ou menos semelhante à advogada por de Morgan. Entre eles, Frank P. Ramsey, Dennis Lindley, Harold Jeffreys, Bruno de Finetti e Leonard J. Savage podem ser apontados como os principais responsáveis por um interesse continuado na interpretação subjetiva, que, próximo à década

[4]Embora os autores citados usem h e e em minúsculo, a notação foi alterada para não haver confusão com o número neperiano $e = 2.7182...$

de 1960, passou a ser o principal embasamento para uma classe de métodos de inferência conhecidos como Bayesianos. Para Ramsey e Savage, a probabilidade surge como um sistema de preferência racional, enquanto para de Finetti, ela representa uma chance justa em uma aposta. As visões são razoavelmente intercambiáveis, e embora difiram no grau de empiricismo advogado, ambas levam logicamente aos axiomas padrão da probabilidade[5]

A inferência Bayesiana é devida ao rev. Thomas Bayes, que demonstrou o teorema que leva seu nome, e ao já mencionado marquês de Laplace, que formulou a metodologia das probabilidades inversas. Em notação moderna, o teorema de Bayes se escreve simplesmente [Morettin and Bussab, 2009]:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1.2}$$

Aqui, $P(A|B)$ indica a probabilidade de A condicional a B, ou seja, a probabilidade de que A seja verdade dado que B é verdade. Este teorema decorre diretamente dos axiomas usuais da probabilidade, e é aceito por todas as escolas de interpretação para *atualizar* informações de probabilidade à luz de um novo conhecimento. Um exemplo de uso muito comum em livros introdutórios é o de exames clínicos. A probabilidade de que uma dada pessoa tenha uma doença, sem informações adicionais, pode ser estimada grosseiramente pela prevalência da doença na população. Quando esta pessoa se submete a um exame clínico de propriedades conhecidas[6], nós podemos usar esse novo conhecimento para *atualizar* a nossa estimativa de que esta pessoa de fato tenha esta doença.

A contribuição relativa de Bayes e Laplace aos sucessos e fracassos da teoria Bayesiana é um tópico controverso entre os que estudam a história da probabilidade [Zabell, 2009]. Fisher, por exemplo, considera o argumento com o qual Bayes estabelece uma *priori* uniforme como um comentário a respeito do problema específico que Bayes estava estudando, enquanto Laplace verá esse argumento como aplicável a outras classes de problemas [Laplace, 1814] - em particular, Laplace argumenta que podemos introduzir probabilidades iguais a hipóteses sobre as quais não conhecemos nada. Assim, enquanto o Teorema de Bayes é aceito por todas as escolas de interpretação, o argumento de Laplace é, em essência, aceito pela escola subjetivista [Aldrich, 2008] e completamente rejeitado pela frequentista.

Talvez a formalização mais convincente da teoria subjetivista de probabilidades seja a dada por de Finetti no seu argumento sobre *Dutch books*. Um *Dutch book*, ou "aposta holandesa" é um conjunto de apostas que garante um ganho para um apostador, qualquer que seja o resultado observado após o jogo. Suponha que um evento vai ser observado (seja uma corrida de cavalos, uma luta, etc), e um agente pode apostar uma quantia em dinheiro para cada possível resultado desse evento, baseado na sua avaliação subjetiva da probabilidade de cada resultado, e recebe um prêmio correspondente ao resultado que de fato ocorrer. A relação

[5]Para uma visão mais aprofundada das diferenças filosóficas entre de Finetti e os demais subjetivistas, veja [Galavotti, 1989].

[6]Nesse caso, é necessário especificar a sensibilidade do teste, ou seja, a proporção de pessoas doentes que recebem um resultado positivo pelo teste, e a especificidade do teste, ou seja, a proporção de pessoas saudáveis que recebem um resultado negativo pelo teste

entre as probabilidades e os prêmios é dita *incoerente* se uma "aposta holandesa" puder ser feita a favor ou contra o agente (ou seja, garantindo lucro ou perda sempre), e *coerente* caso contrário. De Finetti mostra que toda avaliação coerente de probabilidades subjetivas implica os axiomas da probabilidade [de Finetti, 1937].

Já a teoria de inferência Bayesiana pode ser vista, resumidamente, como uma estrutura de atualização das probabilidades subjetivas frente a novas evidências. O teorema de Bayes conecta a probabilidade *a priori* com a probabilidade *a posteriori* através do uso da verossimilhança dos dados. No caso de duas hipóteses concorrentes $H1$ e $H2$, para explicar um evento observado E, o teorema de Bayes diz que:

$$\frac{p(H1|E)}{p(H2|E)} = \frac{p(E|H1)}{p(E|H2)} \frac{p(H1)}{p(H2)} \tag{1.3}$$

Onde o termo $\frac{p(E|H1)}{p(E|H2)}$ é justamente a razão das verossimilhanças. As grandes questões que precisam ser resolvidas são: como construir uma *priori* a partir do conhecimento prévio (que será retomada na seção 1.13), e principalmente, como construir *prioris* na ausência de informação sobre o problema. Fisher aponta que *presumir a indiferença* em relação a um parâmetro p depende da forma matemática do modelo empregado: tomando $\sin\theta = 2p - 1$, uma *priori* "plana" em relação a p privilegiará alguns valores de θ e vice-versa [Fisher, 1922].

A proposição da escola Bayesiana foi acompanhada de um debate intenso, nos quais os argumentos eram não raramente personalizados. Ronald Fisher foi um árduo crítico de Laplace, escrevendo que "We can know nothing of the probability of hypotheses or hypothetical quantities" [Fisher, 1921], e "The theory of inverse probability is founded upon an error, and must be wholly rejected" [Fisher, 1925], ao mesmo tempo em que propunha uma metodologia de estimação fiducial, que permitia a inferência absoluta sobre hipóteses sem *prioris* especificadas para uma classe de "problemas bem-formulados". O método fiducial foi fundamentalmente abandonado nos anos seguintes à sua morte, sendo chamado por Savage de "a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs" [Savage, 1961], e é considerado como refutado por I. J. Good [Good, 1992]. No entanto, os argumentos de Fisher alcançaram várias vezes uma grande lucidez e perspicácia, e a formulação moderna da teoria Bayesiana deve muito à crítica constante feita por ele sobre os pontos mais sutis da sua lógica[7]. Ao mesmo tempo que propunha o argumento fiducial, Fisher também defendia a inspeção e interpretação da função de verossimilhança apenas, em uma abordagem que será retomada mais recentemente.

Com a axiomatização e formulação moderna da interpretação subjetivista, cuja história é intimamente ligada à da teoria econômica de utilidade (veja por exemplo [Friedman and Savage, 1948] e [Pfanzagl, 1967]), a escola Bayesiana de pensamento ganhou maior aceitação nos meios científicos. Savage escreve sobre essa formulação moderna:

> "Personal probability can be regarded as part of a certain theory of coherent preference in the face of uncertainty. This preference theory is normative; its goal is to help us make better decisions by exposing to

[7] O argumento fiducial continua sendo altamente controverso entre filósofos e cientistas, sendo defendido por exemplo em [Hacking, 1965] e na segunda edição de [Edwards, 1972].

> us possible incoherences in our attitudes toward real and hypothetical alternatives." [Savage, 1967]

Essa exposição é reminiscente das palavras de de Morgan:

> "I throw away objective probability altogether, and consider the word as meaning the state of the mind with respect to an assertion (...). 'It is more probable than improbable' means in this chapter 'I believe that it will happen more than I believe that it will not happen. Or rather 'I *ought* to believe, &c.', for it may happen that the state of mind which *is*, is not the state of mind which should be. D'Alembert believed that it was *two* to *one* that the first head which the throw of a halfpenny was to give would occur before the third throw; a juster view of the mode of applying the theory would have taught him it was *three* to *one*. But he *believed* it, and thought he could show reason for his belief: to him the probability *was* two to one. But I shall say, for all that, that the probability *is* three to one; meaning that in the universal opinion of those who examine the subject, the state of mind to which a person *ought* to be able to bring himself is to look three times as confidently upon the arrival as upon the non-arrival."[de Morgan, 1847]

Savage, ainda, propõe a teoria subjetiva de probabilidade como uma resposta ao problema da indução, ou seja, como podemos justificar que nossa experiência passada possa ser usada para prever eventos futuros:

> "The theory of personal probability [prescribes] exactly how a set of beliefs should change in light of what is observed. It can help you say, 'My opinions today are the rational consequence of what they were yesterday and of what I have seen since yesterday.' In principle, yesterday's opinions can be traced to the day before, but even given a coherent demigod able to trace his present opinions back to those with which he was born and to what he has experienced since, the theory of personal probability does not pretend to say with what system of opinions he ought to have been born. It leaves him, just as Hume would say, without rational foundation for his beliefs of today (...). That all my beliefs are but my personal opinions, no matter how well some of them may coincide with opinions of others, seems to me not a paradox but a truism (...). If there is a rational basis for beliefs going beyond mere coherency, then there are some specific opinions that a rational baby demigod must have. Put that way, the notion of any such basis seems to me quite counter intuitive." [Savage, 1967]

1.11 Críticas e secções do pensamento Bayesiano

Uma crítica da escola frequentista pode ser resumida como: se existirem chances físicas associadas a um evento e probabilidades subjetivas, não haveria motivo para

supor que ambas serão iguais. Há diferentes respostas para esse problema entre os proponentes da teoria Bayesiana, que essencialmente caracterizam as diversas vertentes do Bayesianismo moderno.

Por um lado, dadas algumas condições sobre a construção da probabilidade subjetiva de um ator racional, pode ser demonstrado que as chances físicas de um evento são iguais às probabilidades subjetivas às quais ele deve chegar; embora alguns autores discordem sobre quais são as condições razoáveis para tomar como axioma. Anscombe e Aumann, por exemplo, escrevem sobre um problema compondo processos do tipo roleta (nas quais as chances físicas de cada resultado são conhecidas) com processos do tipo corridas de cavalo (nas quais a chance física é desconhecida) que:

> "In this case the subjective probability of any outcome is equal to the [physical] chance associated with that outcome. Since the two are equal, it does not matter much which word or symbol we use. The chance refers to the phenomenon, the probability refers to your attitude towards the phenomenon, and they are in perfect agreement"[Anscombe and Aumann, 1963]

Contraste-se com essa visão a posição anti-realista advogada por de Finneti, para quem a probabilidade não precisa ser racionalmente justificável:

> "The subjective theory (...) does not contend that the opinions about probability are uniquely determined and justifiable. Probability does not correspond to a self-proclaimed 'rational' belief but to the effective personal belief of anyone"[de Finetti, 1951]

Neste sentido, a posição de Anscombe, que privilegia um certo valor racional para a probabilidade de um evento, se afasta daquela defendida por de Finetti, e se aproxima da lógica probabilística de Rudolph Carnap e do Bayesianismo objetivo advogado por E. T. Jaynes [Jaynes, 1968], que se utiliza do princípio da máxima entropia para determinar *prioris* plenamente objetivas.

Outra crítica diz respeito ao fato de que, enquanto subjetiva, a probabilidade percebida por um sujeito pode não corresponder à probabilidade anunciada por ele. Esse problema é abordado pela teoria de *Scoring Rules*, desenvolvida por de Finetti e Savage durante a década de 1970 [Lindley, 1982], cujo desenvolvimento mostra métodos para garantir coerência entre ambas. Dennis Lindley expande esses resultados para mostrar que:

> "Let a person express his uncertainty about an event E, conditional upon an event F, by a number x and let him be given, as a result, a score which depends on x and the truth or falsity of E when F is true. It is shown that if the scores are additive for different events and if the person chooses admissible values only, then there exists a known transform of the values x to values which are probabilities. In particular, it follows that values x derived by significance tests, confidence intervals or by the rules of fuzzy logic are inadmissible. Only probability is a sensible description of uncertainty." [Lindley, 1982]

A escola Bayesiana também é criticada ao assumir uma postura subjetivista, que não seria compatível com o desenvolvimento da ciência, definida como a construção de conhecimento racional objetivo. Essa oposição, a bem da verdade, é mais reveladora de uma visão sobre a natureza da ciência do que uma crítica propriamente ao paradigma Bayesiano: para um Bayesiano, o julgamento de um especialista treinado permite uma relação confiável com natureza, ainda que não objetiva. Frise-se que mesmo avaliações de probabilidade pessoais e subjetivas podem estar abertas ao estudo objetivo, uma posição que remonta ao trabalho de C.S. Peirce [Stigler, 1978]. Outro fator importante para responder essa crítica é a maior percepção moderna de que a escola frequentista, ao ser obrigada a escolher um modelo mental de referência, é menos objetiva do que se propõe. Savage escreve que "the Bayesian approach is more objectivistic than the frequentist approach in that it imposes a greater order on the subjective elements of the deciding person." [Savage, 1961].

1.12 O problema da parada opcional

Nas seções anteriores, encontramos a proposição de duas escolas de inferência: a frequentista e a bayesiana, e críticas a esta segunda. Nesta seção e na próxima vamos nos debruçar sobre críticas feitas sobre a escola frequentista, em dois problemas que expõem dificuldades fundamentais em sua aplicação - e justificam a procura por uma terceira escolha.

Uma crítica muito veemente aos proponentes da inferência frequentista é dada pelo problema da parada opcional, ou "optional stopping". Suponha que um cientista realizou uma série de 12 ensaios de Bernoulli, ou seja, experimentos aleatórios com dois possíveis resultados (sucesso e falha), independentes e de mesma probabilidade. Após obter 3 sucessos, o cientista decide testar a hipótese nula de que a probabilidade de sucesso é de 0.5. Pelo paradigma frequentista, esse teste depende da maneira como o experimento foi projetado. Se o cientista planejava fazer 12 experimentos, a probabilidade de observar 3 ou menos sucessos sob a hipótese nula é de $\left(\binom{0}{12} + ... + \binom{3}{12}\right)(\frac{1}{2})^{12} = 7.3\%$. No entanto, se o cientista decidiu continuar realizando experimentos até encontrar o terceiro sucesso, a probabilidade de chegar a doze tentativas é de $1 - \left(\frac{1}{2}^3 + \binom{3}{2}\frac{1}{2}^4 + ... + \binom{10}{2}\frac{1}{2}^{11}\right) = 3.3\%$. Ou seja, dependendo da intenção original do cientista, os mesmos dados podem servir para refutar ou não a hipótese nula.

Para evitar esse tipo de paradoxo, a sabedoria convencional dos estatísticos frequentistas afirma que um cientista *não pode* examinar os próprios dados enquanto os coleta. Toda a análise estatística deve ser feita posteriormente à coleta. Desta forma, o cientista que realizou os 12 testes e encontrou um p-valor de 7.3%, próximo do nível crítico de 5%, é proibido de continuar a coleta de dados. Richard Royall, citando essa prática, diz:

> "Subsequent observations, no matter how consistent and convincing, can never justify a claim of statistical significance at his target level (...). Finding that the early partial results represent evidence that is only fairly strong precludes the possibility that the evidence in the final results might be quite strong. Does this make sense?" [Royall, 1997]

Alguns estatísticos veem neste problema "um uso ingênuo de p-valores" [Good, 1992], ao lembrar que, sob um ponto de vista Fisheriano, um valor p de 7.3% não leva a qualquer juízo sobre a validade ou não da hipótese nula. Outros, como Richard Royall, consideram esta situação uma indicação clara de que o paradigma frequentista leva a contradições inerentes, e portanto deve ser abandonado. A discordância entre as conclusões frequentistas e a inferência julgada aceitável por seus opositores pode ser traçada na incoerência entre o procedimento de teste de hipóteses e o princípio da verossimilhança. Utilizado desde a década de 1930 em caráter razoavelmente intuitivo Neyman and Pearson [1933], o princípio é demonstrado por Allan Birnbaum em 1962, e afirma que toda a inferência feita a partir de um experimento deve ser baseada na função de verossimilhança, sendo portanto irrelevante o desenho experimental [Birnbaum, 1962]. No exemplo discutido acima, a função de verossimilhança do experimento é a mesma, não importando a intenção do pesquisador:

$$\mathcal{L}(p|x) \propto p^3(1-p)^9 \tag{1.4}$$

A razão de verossimilhanças, critério já utilizado por Fisher e Neyman-Pearson, entre os valores de $p = 0.5$ e $p = 3/12$ é de aproximadamente 5, oferecendo uma evidência moderada a favor do valor $p = 3/12$. Dentro do paradigma Bayesiano, onde o princípio da verossimilhança é aceito, essa razão de verossimilhanças deve ser usada para atualizar as probabilidades que cada hipótese tem *a priori*. Como veremos na seção 1.14, proponentes da verossimilhança como base da inferência vão questionar o uso explícito de *prioris*, atendo-se apenas ao valor da razão de verossimilhanças.

1.13 O problema da classe de referência

Em seu trabalho de 1922, Fisher escreve:

> "The framing by means of a model is located at the beginning of the statistical treatment of a problem of application. The postulate of randomness thus resolves itself into the question, 'Of what population is this a random sample?' which must frequently be asked by every practical statistician"[Fisher, 1922]

Fisher volta a essa questão em 1955, ao contrastar seu método de teste de hipóteses com o procedimento de Neyman-Pearson:

> "The root of the difficulty of carrying over the idea from the field of acceptance procedures to that of tests of significance is that, where acceptance procedures are appropriate, the source of supply has an objective reality, and the population of lots, or one or more, which could be successively chosen for examination is uniquely defined; whereas if we possess a unique sample in Student's sense on which significance tests are to be performed, there is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard

> our sample as belonging; so that the phrase 'repeated sampling from the same population' does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination." [Fisher, 1955]

A mesma questão se manifesta no exemplo mais frequente de aplicação do teorema de Bayes (brevemente discutido na seção 1.10): dado o resultado positivo em um teste clínico, qual a probabilidade de que o paciente tenha a doença? A resolução desse problema passa por identificar que:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\bar{D})P(\bar{D})} \tag{1.5}$$

Onde $P(+)$ é a probabilidade do teste dar positivo e $P(D)$ é a probabilidade de um indivíduo ter a doença. Este último termo vai ser interpretado como a *priori* por um Bayesiano, ou mais geralmente como a prevalência da doença na população. A aparente simplicidade desse exemplo esconde a *escolha* de uma população: devemos considerar a prevalência na cidade onde o indivíduo mora, ou no país? Indivíduos de ambos os gêneros devem ser considerados? A prevalência deve levar em consideração a faixa etária, a renda, o hábito esportivo, alimentar ou uso de substâncias pelo indivíduo? Como cita Henry Kyburg:

> "Probability theorists and statisticians have tended to take one of two tacks in relation to this problem: either they deny the significance of probability assertions about individuals, claiming that they are meaningless; or they open the door the whole way, and say that an individual may be regarded as a member of any number of classes, that each of these classes may properly give rise to a probability statement, and that, so far as the theory is concerned, each of these probabilities is as good as another. (...)
>
> Some philosophers have argued that this is a merely pragmatic problem, rather than a theoretical one (...). But calling a problem practical or pragmatic is no way to solve it." [Kyburg, 1974]

Essa questão, então, perpassa a formulação do problema de inferência para ambas as principais escolas. De certa forma, podemos ver que os problemas de ambas as escolas se concentram não na interpretação do dado observado, mas sim no *background* contra o qual ele vai ser comparado. Isso inspira a questão: será que é possível desenvolver um programa de inferência focado na observação, sem precisar recorrer a classes externas de referência?

1.14 A escolha da verossimilhança

A grandeza conhecida como verossimilhança, defendida por Fisher a partir de 1922, viu grande uso nas últimas páginas. Por um lado, ela é a base sobre a qual se assenta o teste de hipóteses de Neyman-Pearson. Por outro, ela é a ponte entre a *priori*

e a *posteriori* Bayesianas. No final dos anos 1960, o uso da verossimilhança como base da inferência ganha vida independente, e em grande parte devido ao livro de I. Hacking em 1965 sobre a lógica da inferência, e ao trabalho de A.W.F. Edwards em 1972, é possível considerar essa postura como uma escola independente de pensamento, denominada às vezes Verossimilhantismo. Mais recentemente, essa escolha vai ser veementemente apoiada por Richard Royall e Elliott Sober. Esta escola de pensamento se baseia na conjunção do princípio da verossimilhança, conforme demonstrado por Birnbaum, e na lei da verossimilhança, enunciada por Hacking.

Birnbaum deriva o princípio da verossimilhança a partir de dois princípios mais facilmente aceitos[8]: o princípio da suficiência e o princípio da condicionalidade. Informalmente, estes princípios afirmam que "dados que não agregam informação são irrelevantes para a inferência", e "experimentos que poderiam ser realizados mas não foram são irrelevantes para a inferência". Desses princípios, Birnbaum deduz o princípio da verossimilhança, informalmente, "resultados amostrais possíveis mas não observados são irrelevantes para a inferência", de forma que questões como o desenho amostral não podem influenciar a inferência realizada:

> "The likelihood principle: If E and $E\prime$ are any two experiments with the same parameter space, represented respectively by density functions $f(x, \theta)$ and $g(y, \theta)$; and if x and y are any respective outcomes determining the same likelihood function; then $Ev(E, x) = Ev(E\prime, y)$. That is, the evidential meaning of any outcome x of any experiment E is characterized fully by giving the likelihood function $cf(x, \theta)$ (which need be described only up to an arbitrary positive constant factor), without other reference to the structure of E." [Birnbaum, 1962]

Já a lei da verossimilhança é enunciada por Hacking como:

> "If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence." [Hacking, 1965]

A divergência entre a escola da verossimilhança e a inferência frequentista se dá pois esta última rejeita o princípio da verossimilhança (embora aceite tanto a lei da verossimilhança quanto os princípios da suficiência e condicionalidade). Por outro lado, a escola da verossimilhança vai divergir da inferência Bayesiana (que aceita o princípio da verossimilhança) pela interpretação da lei da verossimilhança.

A escola da verossimilhança pode ser vista como herdeira intelectual da abordagem de Neyman-Pearson, cujo ponto de partida para atacar o teste de significância de Fisher foi reconhecer a necessidade lógica da comparação entre diferentes hipóteses. Enquanto o teste de significância privilegia uma hipótese nula, a abordagem de Neyman-Pearson, tal qual a de Edwards e Royall, vai refutar essa abordagem como

[8]Essa derivação é controversa até o presente, sendo contestada por uma série de filósofos e estatísticos e defendida por outros (veja [Mayo, 2010] e [Gandenberger, 2012], respectivamente, para críticas e defesas atuais)

incompleta. Os dados recolhidos por um experimento podem indicar um alto ou baixo suporte para uma certa hipótese, mas isso não deve ser considerado evidência a favor ou contra essa hipótese sem que hipóteses alternativas sejam igualmente escrutinizadas. É sobre o uso da razão entre verossimilhanças de diferentes hipóteses, e não de qualquer estatística calculada sobre uma única hipótese, que Neyman, Pearson, Edwards e Royall vão assentar seu programa de inferência. O detetive Sherlock Holmes, personagem de Conan Doyle, famoso por seu raciocínio arguto, é célebre pela frase "How often have I said to you that when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth?". Ou seja, a probabilidade ou improbabilidade de uma única hipótese, vista isoladamente, não deve ser usada para concluir sobre sua veracidade.

A divergência entre as abordagens se dá fundamentalmente com a *ação* que o cientista deve tomar após examinar a razão de verossimilhanças. Neyman-Pearson argumentam que esse valor deve nortear um processo de teste de hipóteses, que leva ao *comportamento* de aceitar uma hipótese em detrimento das demais. Royall e Edwards realizam a separação conceitual entre (1) o grau de certeza que temos sobre a hipótese, (2) a força da evidência que um conjunto de dados confere a uma hipótese em detrimento de outras, e (3) o curso de ação tomado após verificar tais quantidades. Enquanto a escola Bayesiana vai relacionar a força de evidência (2) com o grau de certeza (1), e a proposição de Neyman e Pearson vai identificar a força de evidência (2) com o curso de ação (3), a proposta dos verossimilhantistas é de que a inferência estatística se concentra *apenas* na força da evidência. O curso de ação tomado pode levar em conta uma multiplicidade de outros fatores, representados por *priors* e funções de perda na teoria de decisões. Laplace já havia separado esses conceitos com o problema do número de juízes necessário para condenar um prisioneiro: A decisão, além de levar em conta as probabilidades de condenar um inocente ou perdoar um culpado, deve levar em conta se a pena será uma multa ou a morte.

Um exemplo que ilustra as diferentes concepções de inferência e suas consequências é dado pelo seguinte experimento mental: tome um baralho ordinário e vire a primeira carta, para encontrar um ás de espadas. Suponha agora duas hipóteses concorrentes: H_N de que o baralho é normal e H_A de que o baralho é um engodo, composto por 52 ases de espadas. A verossimilhança de H_N é $\mathcal{L}(H_N|A\spadesuit) = \frac{1}{52}$, contra $\mathcal{L}(H_A|A\spadesuit) = 1$. Enquanto o procedimento de Neyman-Pearson diz "escolha a hipótese que leva à maior verossimilhança", Royall e Edwards vêem nessa situação apenas uma *evidência* favorável a H_A sobre H_N. Para transformar essa evidência em um curso de ação, é necessário incorporar mais informação ao problema. Ainda, para transformar essa evidência em um grau de certeza, é necessário incorporar o nosso conhecimento *a priori* sobre a situação: enquanto nossa intuição diz que um baralho de engodo deve ser muito raro, e portanto o baralho examinado *provavelmente* é normal, um marciano, que não tem a mesma *priori*, deve achar muito natural que todas as cartas tenham a mesma figura, afinal todas tem o mesmo verso.

Em particular, ao dissociar a *força de evidência* da *tomada de decisão*, Royall e Edwards estão solicitando ao cientista que divulgue seus dados na forma de razões de verossimilhança, não transformadas em valores-p ou *posterioris*, para que seus pares possam ter clareza de qual é a evidência apresentada. Ao recuperar o papel

do cientista como tomador de decisões subjetivas, a abordagem da verossimilhança volta a se encontrar com o pensamento de Fisher, que critica Neyman-Pearson por um programa que é adequado ao processamento industrial de lotes, mas não à produção de conhecimento científico. Neste aspecto, Royall e Edwards ecoam as palavras de Fisher e de I. J. Good:

> "We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions." [Fisher, 1955]

> "If a Bayesian is a subjectivist, he will know that the initial probability density varies from person to person and so he will see the value of graphing the likelihood function for communication." [Good, 1976]

Porém, se a tomada de decisão deve ser feita com base na razão de verossimilhanças e na especificação de uma probabilidade *a priori*, no que o paradigma da verossimilhança difere das escolas Bayesianas? Ou ainda: será que a inferência baseada em verossimilhanças é uma "inferência Bayesiana sem *prioris*"? Para responder isso, é importante retomar a formulação da lei da verossimilhança:

> "If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is **evidence supporting** A **over** B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence." [Hacking, 1965]

Embora seja possível considerar o paradigma Bayesiano como um paradigma verossimilhantista *sensu lato*, ele utiliza a razão das verossimilhanças como meio para construir quantias absolutas, a probabilidade *a posteriori* de A e a probabilidade *a posteriori* de B. Um verossimilhantista *sensu stricto*, ao contrário, verá na razão das verossimilhanças o objeto final do seu estudo. O paradigma da verossimilhança, dessa forma, se opõe à transformação de uma quantidade de suporte relacional entre A e B em quantidades não-relacionais de suporte para cada hipótese [Fitelson, 2007]. Embora Royall e Edwards forneçam razões fortes para aceitar essa lei como premissa, ela não é uma necessidade lógica, e não decorre de princípios mais básicos - ao contrário do princípio da verossimilhança. Note-se aqui que o texto de Royall não provê nenhuma definição do que seja evidência ou suporte estatístico, enquanto Edwards provê apenas uma lista de propriedades desejáveis a uma medida de suporte, demonstrando que a razão de verossimilhanças é coerente com estas propriedades. Hacking, por outro lado, usa os axiomas de [Koopman, 1940] para embasar sua definição de suporte, à qual a lei da verossimilhança é adicionada axiomaticamente - e de forma coerente. A posição de todos estes autores indica que a aceitação da lei da verossimilhança equivale a uma definição de evidência.

A interpretação verossimilhantista da evidência recebe diversas críticas. Uma, no entanto, é especialmente contundente: Edwards propõe que definir a natureza da probabilidade é um problema irrelevante, dado que a evidência estatística é

formulada por razões de verossimilhança, e não por probabilidades. No entanto, a verossimilhança é, por definição, *uma função proporcional a uma probabilidade*, de forma que sem uma interpretação para a probabilidade não podemos ter uma interpretação para a verossimilhança. A teoria verossimilhantista corre, portanto, o risco de se assentar sobre uma areia movediça filosófica.

1.15 A abordagem por seleção de modelos

A escola de verossimilhança vai ganhar especial força devido ao desenvolvimento de um procedimento geral de escolha de modelos, baseado no resultado de Akaike relacionando o número de parâmetros livres em um modelo e a log-verossimilhança com a perda de informação de cada modelo [Akaike, 1974]. O procedimento de escolha de modelos vai retomar o problema da especificação de Fisher (conforme descrito na seção 1.7) como um problema possível de ser tratado dentro da metodologia estatística.

A proposta do AIC (Akaike Information Criterion) é baseada na divergência de Kullback-Leibler[9] [Kullback and Leibler, 1951], definida como a diferença entre duas distribuições de probabilidade, P e Q; supondo que P é a distribuição verdadeira, a divergência de Q a partir de P é a medida de informação perdida ao aproximar P usando Q.

O uso da divergência de KL na inferência estatística espelha a definição de Fisher de uma população infinita, para a qual existe uma lei de distribuição desconhecida, com parâmetros desconhecidos. Da mesma forma que na inferência Fisheriana, a amostra disponível será usada em comparação com esse modelo idealizado. A relação entre a distribuição real para KL e a realidade é a mesma que a relação entre a população Fisheriana e a realidade: após aceitar essa construção teórica, a qualidade do ajuste ("goodness of fit") deve ser verificada entre os fatos disponíveis e o modelo mental, e não mais com a realidade.

Como definida acima, a divergência de KL não pode ser medida ou estimada, pois ela depende da distribuição verdadeira, suposta desconhecida. O resultado fundamental de Akaike é mostrar que uma fórmula simples (conhecida como AIC) aproxima assintoticamente sem viés a divergência de KL relativa esperada dentro de um conjunto estabelecido de modelos. Ou seja, para um grande número de pontos amostrais, a log-verossimilhança máxima de cada modelo ($\log(\mathcal{L}(\hat{\theta}|x))$) aproxima a diferença de divergência de KL entre os modelos de forma enviesada, e a correção para este viés é aproximadamente igual ao número de parâmetros livres no modelo (K)[Burnham and Anderson, 2004]. Assim, podemos calcular o AIC como:

$$AIC = 2K - 2\log(\mathcal{L}(\hat{\theta}|x)) \tag{1.6}$$

O valor do AIC de um único modelo não possui nenhum significado: ele apenas pode ser usado *em comparação* com o AIC de outro modelo proposto para o mesmo

[9]Muitas vezes referida incorretamente como "distância de Kullback-Leibler", a divergência de Kullback-Leibler não apresenta em geral propriedades de uma métrica: em particular, a divergência entre P e Q não é igual à divergência entre Q e P

conjunto de dados. É usual calcular a diferença entre o AIC de cada modelo e o AIC mínimo encontrado:

$$\Delta_i = AIC_i - AIC_{min} \tag{1.7}$$

Este valor pode ser usado para comparar diversos modelos dentro do mesmo arcabouço conceitual no qual se realiza a estimativa dos parâmetros. O melhor modelo, dentro do conjunto escolhido, terá $\Delta_i = 0$, enquanto os demais modelos poderão ser ordenados em relação a esses valores. Essa ordenação permite que todos os modelos plausíveis sejam usados para a realização de inferências ou previsões (veja uma descrição completa em [Burnham and Anderson, 2002]).

O uso do número de parâmetros K para "penalizar" modelos com mais parâmetros é intuitivamente atraente, por se aproximar de um princípio de parcimônia. Proponentes anteriores da abordagem de verossimilhança já haviam notado que o modelo com melhor verossimilhança para qualquer conjunto de dados é o de criação individual[10], mas o uso do AIC fornece um motivo sólido e teoricamente bem embasado para preferir os modelos com menos parâmetros livres. Ainda, o uso de seleção de modelos permite que cada modelo comparado se relacione com uma hipótese biologicamente interpretável, ao contrário do uso de testes de significância, nos quais a hipótese nula dificilmente tem relevância biológica. Estes motivos, aliados ao fácil uso do AIC, fizeram com que o uso de seleção de modelos tenha visto interesse crescente na ecologia e outras áreas da biologia [Johnson and Omland, 2004].

Outra medida usada em ordenação de modelos é conhecida como Bayesian Information Criterion (BIC) [Schwarz et al., 1978], o que levou a um debate na literatura da área sobre qual critério deve ser usado [Weakliem, 1999]. Embora os argumentos a favor de cada critério estejam fora do escopo deste texto, é importante notar que o debate não pode ser enquadrado como proponentes do Bayesianismo *versus* proponentes do frequentismo/verossimilhança, já que tanto o AIC pode ser deduzido dentro do paradigma Bayesiano pelo uso de uma "*priori* esperta" [Burnham and Anderson, 2004], como o BIC pode ser deduzido de uma perspectiva de teoria da informação usando uma "verossimilhança substituta" [Stoica and Selen, 2004].

1.16 A incerteza e a produção de conhecimento

Concluímos esse capítulo com a ponderação de que a divisão rotineiramente feita entre as ditas "estatística frequentista" e "estatística Bayesiana" esconde o fato de que há, na verdade, uma variedade de correntes filosóficas que, baseadas em idéias distintas sobre a natureza das probabilidades, chegam a conclusões distintas sobre assuntos variados da epistemologia à metodologia de inferência. A escola frequentista é fundamentalmente dividida entre a modelagem de Fisher e a rigidez de Neyman-Pearson, enquanto a escola Bayesiana é dividida em uma miríade de escolas, desde subjetivistas anti-realistas como de Finetti até objetivistas como Jaynes.

[10]Intuitivamente, é muito fácil ajustar perfeitamente 150 pontos de dados usando um modelo com 150 parâmetros livres. Este modelo é conhecido como "modelo de criação individual", pois ele é equivalente a supor que cada ponto amostral foi gerado independentemente, ou "alternativa de verossimilhança maximal".

Essa taxonomia, no entanto, não acomoda a visão de positivistas lógicos como Rudolf Carnap, que ao lado de Karl Pearson atacarão tanto a visão subjetiva de probabilidade quanto o modelo de população infinita dos frequentistas [Lenhard, 2006; Zabell, 2009]. Tampouco é suficiente para enquadrar o trabalho de Peirce sobre propensões, ou a visão agnóstica de Edwards, posteriormente expandida por Royall e Sober e complementada por Akaike, que propõem a razão de verossimilhança como única base lógica para a inferência.

É necessário que as divergências entre os diversos autores citados sejam interpretadas não na sua práxis, que representa o nível metodológico, como costuma acontecer em círculos mais inclinados à ciência do que à filosofia, e sim nos seus fundamentos, em cada suposição elementar feita sobre os elementos constituintes da teoria probabilística e de inferência. Possivelmente, essas suposições devem ser mapeadas em distintos elementos da filosofia da ciência. Esta seção não pretende trazer nenhuma resposta sobre isso, mas apenas elencar algumas das principais questões que devem ser consideradas.

Ao discutir as correntes objetivistas e subjetivistas de interpretação, deixamos de lado dois pontos muito importantes (mas cujo tratamento adequado é vasto demais para esta nota): o que significa "objetividade", e porque objetividade é uma virtude a ser buscada na ciência. Em seu livro de 2007, Daston e Galison trazem um panorama das diferentes visões sobre objetividade que a comunidade científica apoiou ao longo da história [Daston and Galison, 2007]. Em particular, é importante a distinção entre a objetividade mecânica (a natureza deve ser mensurável de formas que independam do observador) da objetividade estrutural (a ciência deve ser comunicável a partir de estruturas invariantes - leis, relações lógicas, equações). No início do século XX, surge ainda a figura do julgamento treinado, que pode ser entendido como a idéia de que os dados científicos precisam do julgamento de um profissional treinado para fazerem sentido. Apesar de usar a mesma palavra, Venn, Fisher, Neyman e Savage podem estar se apoiando em conceitos muito diversos.

Outra divergência entre os diversos autores diz respeito à construção de conhecimento através da observação, o que podemos enquadrar dentro da classe de problemas da indução. Se observamos que o sol sempre nasce no leste, isso pode ser usado para estabelecer como fato que o sol nascerá sempre no leste? A conexão feita entre "eu observei esse fato no passado" e "eu prevejo que esse fato se repetirá no futuro" não encontra justificativa lógica. Nossa própria convicção de que a indução deve funcionar para estabelecer fatos é baseada em uma indução. Um argumento geralmente associado a esse problema é o problema das esmeraldas "verzuis"[11], proposto por Nelson Goodman em 54 [Goodman, 1983], e extensamente debatido por Hacking [1965], Kyburg [1974], W.C. Salmon, B. Fitelson, entre outros: um objeto é verzul se ele é verde antes de um determinado tempo t e azul após esse instante[12]. Supondo que t seja o presente momento, as duas frases a seguir devem ser igualmente justificadas:

1. Todas as esmeraldas que eu observei até agora são verdes, logo a próxima esmeralda que eu observar deve ser verde

[11] Traduzido livremente a partir do termo *grue*

[12] Mas veja [Quine, 1970] para uma definição diferente do problema, junto com uma proposta de solução.

2. Todas as esmeraldas que eu observei até agora são verzuis, logo a próxima esmeralda que eu observar deve ser verzul (ou seja, azul!)

O problema de confirmar uma hipótese a partir de observações experimentais não se trata apenas de decidir quais formas de confirmação são adequadas[13], mas também de decidir como construir as hipóteses - Goodman, por exemplo, propõe uma divisão entre predicados *projetáveis* e *não-projetáveis*. Embora seja possível descartar exemplos construídos artificialmente como tolos, o problema de determinar se hipóteses científicas seriamente propostas são ou não projetáveis pode ser muito complexo. Vários trabalhos científicos publicados em periódicos são informalmente criticados por serem análogos à anedota do atirador texano, que primeiro atira várias vezes na lateral do estábulo, e depois desenha um alvo no maior agrupamento de acertos.

Finalmente, há paralelos entre as diferentes abordagens de inferência estatística e questões ligadas ao problema da demarcação de Karl Popper:

> "It is easy to obtain confirmations, or verifications, for nearly every theory - if we look for confirmations.
>
> Confirmations should count only if they are the result of *risky predictions*; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory - an event which would have refuted the theory.
>
> Every *good* scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is.
>
> A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice." [Popper, 1963]

O paradigma de testes de significância de Fisher tem paralelos com esta visão: uma teoria científica é aceitável enquanto a evidência contrária não é suficiente para refutá-la. Uma dada hipótese, para Fisher, nunca é comprovada - portanto nos situamos de maneira análoga à posição de Popper, para quem o conhecimento é sempre provisório. No entanto, na forma como teste de hipóteses Fisheriano é frequentemente utilizado, a hipótese de interesse não é diretamente testada - para investigar a existência de um fenômeno X, testamos a hipótese nula de inexistência de X. Neste paradigma, rejeitar a inexistência de X não implica em aceitar a existência de X.

Uma abordagem Bayesiana ou verossimilhantista, por outro lado, provê uma forma de testar diretamente teorias conflitantes e de rejeitar teorias incompatíveis com as observações experimentais. Isso tem importância central no problema de testar teorias cuja própria formulação é estatística:

> "Those theories which are under a probabilistic form (thermodynamics, statistical [mechanics?], etc.) can never be refuted (...): after all, even an event with probability 0 can well occur. According to Popper we

[13] Vide seção 1.9 para uma primeira tipologia das formas de confirmação

> then have to take a methodological decision and consider a probabilistic theory T refuted if a certain event occurred which, as far as T is concerned, would otherwise be extremely unlikely." [de Finetti, 2010]

Reforçamos que a discussão metodológica na aplicação da estatística às ciências práticas deve se fazer lado a lado com a uma discussão epistemológica e filosófica sobre a verdadeira base do pensamento estatístico. Nesse sentido, a proposição de métodos baseados na verossimilhança no presente trabalho não trata de rejeitar o uso de testes de hipótese e procedimentos frequentistas e bayesianos em bases pragmáticas, e sim na construção de ferramentas e na aparelhagem do pesquisador com métodos que não encontrem as dificuldades filosóficas às quais o frequentismo leva, nem pressuponham as escolhas nas quais a escolha o bayesianismo implica.

Chapter 2

Parameter space exploration: a synthesis

2.1 Introduction

There is a growing trend in the use of mathematical modeling tools in the study of many areas of the biological sciences. The use of models is essential as they present an opportunity to address questions that are impossible or impractical to answer in either in purely theoretical analyses or in field or laboratory experiments, and to identify the most important processes which should then be investigated by experiments. One compelling example is made by the Individual Based Models (IBM), which represent individuals that move and interact in space, according to some decision-making rules. These models permit a great level of detail and realism to be included, as well as linking multiple levels of complexity in a system.

On the other hand, more realistic models employ a vast selection of input parameters, from temperature and rainfall to metabolic and encounter rates, which may be difficult to accurately measure. Moreover, one may be interested in estimating how much predictions for a model fitted in one place or to one species may be extrapolated to different places or species. While variations in some of those parameters will have negligible impact on the model output, other parameters may profoundly impact the validity of a model's predictions, and it may be impossible to determine *a priori* which are the most important parameters. A *naïve* approach would consist on the evaluation of the model at all possible combinations of parameters, however this would require a prohibitive number of model runs, specially considering that a single run of those models may take days to complete. Our challenge then consists in providing the best estimates for the importance of the several parameters, requiring the least number of model runs.

Some models are either expressed or can be reasonably approximated by analytical functions. Such is the case for the matrix population models, for which a wide range of analytical tools are available to examine the uncertainty and sensitivity of parameters[Caswell, 1989]. In the general case, however, there is no possible formulation of the model in a closed equation, so analytical methods are not possible.

The disciplines of uncertainty and sensitivity analysis have been developed in the context of the physical sciences and engineering, and have been greatly developed

in the 1980 and 1990 decades [Archer et al., 1997; Florian, 1992; Helton et al., 2005; Helton and Davis, 2003; Huntington and Lyrintzis, 1998; Iman and Conover, 1982, 1987; Kleijnen and Helton, 1999; McKay and Beckman, 1979; Morris, 1991; Saltelli, 2004; Saltelli et al., 1999; Smith, 2002; Ye, 1998].

More recently, these analyses have been successfully applied to biological models, in order to explore the possible outcomes from the model output, estimate their probability distribution and the dependency of the output on different combinations of parameters, and to assess which parameters require more experimental effort in order to be more confidently estimated. This kind of parameter space exploration is considered a fundamental step prior to using the model in management decisions [Bart, 1995].

One approach to the parameter space exploration, which will be described here, is to generate samples from the parameter space, run the model with these samples, and analyze the qualitative or quantitative differences in the model output. In this context, the sampling of the parameter space may be regarded as a bridge between the modeling of the system and the inference problem of acquiring information about the whole parameter space, having only access to a subset of that information. This inference can be done in the light of any of the statistical schools.

Section 2.2 will present the sampling techniques, with emphasis on the Latin Hypercube method, while section 2.3 will present some tools for the quantitative analyses. We should emphasize that the analyses tools are not coupled with the sampling techniques used: one can, in principle, use the sampling techniques described and other analyses tools, or apply the analyses described here to a more general class of sampling methods. We also present two examples of the sampling and analysis used in section 2.4 and 2.5. Then, we briefly review some relevant research papers which have used such techniques in the exploration of ecological models in section 2.6.

2.1.1 Parameter spaces

In order to better pose our questions, we need first to discuss some properties of the parameter space (or PS for short) of our models.

The parameters (or inputs) are quantities $x_1, x_2, \cdots, x_m$ which will be used to run the model. In our discussion, we will assume that all the x_i are real valued. These quantities are unknown, and one first challenge is to determine which set of values better fit a model to the available data, which is the subject of linear and nonlinear estimation.

However, the same model may be parametrized in different ways, as discussed by Ross [1990]. For example, in population ecology, the logistic growth equation may be represented with two parameters r and K as:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \tag{2.1}$$

However, the same equation may be written as

$$\frac{dN}{dt} = \alpha N + \beta N^2 \tag{2.2}$$

Here, the two parameters are $\alpha = r$ and $\beta = -r/K$. While the first equation is far more commonly used in the biological context, both are equivalent, and using one or the other model is simply a matter of choice.

There are many other ways of writing this equation, and one of special interest when trying to fit real data is in terms of orthogonal polynomials, such as:

$$\frac{dN}{dt} = \theta_0 + \theta_1(x - \bar{x}) + \theta_2((x - \bar{x})^2 - (\overline{x - \bar{x}})^2) \tag{2.3}$$

Where θ_0 can be calculated from θ_1 and θ_2. This more complicated equation has several numerical advantages over the previous, as the parameters θ_1 and θ_2 can be estimated with much more accuracy, and will not be correlated (as is the case with α and β, as well as r and K). However, these parameters are hard to interpret in biological terms.

These different equations illustrate the existence of *interpretable* (Eq. 2.1), *defining*, or algebraic (Eq. 2.2) and *computing* (Eq. 2.3) parameters. Most of the times, it would be preferable to estimate the values that best fit some data by using computing parameters, and then to transform them to interpretable parameters in order to present the results.

Also, it should be mentioned that the parameter space may be constrained. This will have an impact on some of the available sampling and analysis techniques. The simplest constraint is requiring some parameter to be positive or negative. Also, there may be combinations of values that are meaningless. For example, if we model a community with N individuals and S species, the number of individuals and species, considered on their own, may be any positive number. However, it is clear that the number of species may not be bigger than the number of individuals, which imposes the condition $S \leq N$. This condition is called a *constraint*, and limits the values that the parameter vector may assume.

If we consider the m-dimensional space consisting on all possible combination of values for the parameters, our parameter space will be the subset of this space that respects all our constraints. For example, consider that the parameters we are interested are two angles of a triangle. In this case, the sum of the angles must be less than 180 degrees, $a_1 + a_2 < 180\circ$. Clearly, this parameter space is not square, in the sense that, if we define the ranges of the variables a_1 and a_2 independently as $(0, 180)$, not all combinations of parameters will be meaningful. What can be done in this case is to create a new parameter $\hat{a_1}$, defined as

$$\hat{a_1} = \frac{a_1}{180 - a_2} \tag{2.4}$$

This new parameter varies between 0 and 1, and all combinations of $\hat{a_1}, a_2$ are points from our parameter space. Now, care must be exercised after applying such transformations in order to preserve the marginal distributions from the original variables, as exemplified on figure 2.1.

Another related concept that should not be confused with the constraints is the correlation between variables. For example, acidic soils are likely to have a lower cation exchange capacity (CEC), and more alkaline soils are likely to have a larger CEC [Sparks and Sparks, 2003]. Thus, those variables are *correlated*. Correlations

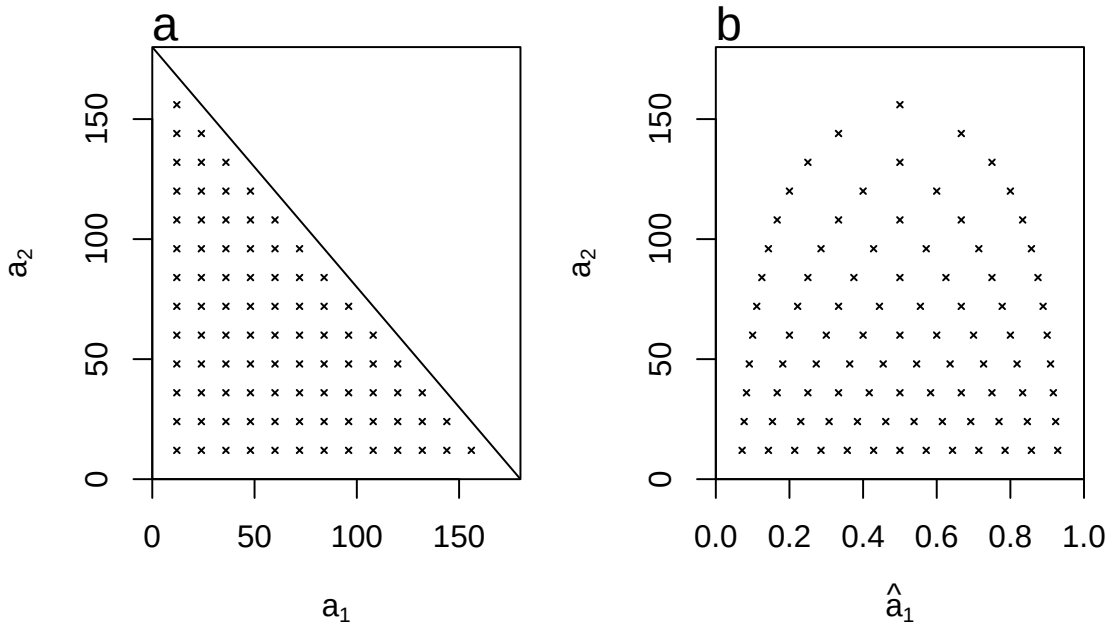


Figure 2.1: a. The constrained parameter space considered, with the line representing $a_1 + a_2 = 180$. The symbols represent a uniform sample taken from the space. b. The transformed parameter space $\hat{a_1}, a_2$ (see Eq. 2.4), showing the same sampled points.

have a profound impact on some analysis, however, they are difficult to measure, and data on correlations are not generally available on the literature. We will return to questions related to correlations in parameter spaces in section 2.2.2.

2.1.2 Applications of parameter space exploration

Next, we turn our attention to the kind of problems we might want to address with the exploration of the parameter space. First, the simplest case is asking "is there a region of my parameter space where condition X holds?" This condition might be, for example, the extinction or coexistence of species, some pattern of distribution or abundance of species. We also might be interested in mapping where are these regions. In complex models, where several different regions might exist where the qualitative results of the models are very different, we may ask how many of these regions are there, as well as map the frontiers between them. For a detailed discussion of this approach, see the PSExplorer software[Tung and Lee, 2010].

Another class of problems arises when the model produces some quantitative response, and we are interested in determining the dependency of this response to the input parameters. For example, when modeling the dynamics of a population, we might want to know how the final population varies with each of the input parameters. In this context of quantitative analysis, the questions are divided in two classes: first, how much the variation of the input parameters is translated into

the total variation of the results, which is the topic of uncertainty analysis, and second, how much of the variation in the results can be ascribed to the variation of each individual parameter, which is the topic of sensitivity analysis [Helton et al., 2005; Helton and Davis, 2003]. We will present the techniques and results from both uncertainty and sensitivity analysis in section 2.3.

Also, the model which we wish to analyze can be any function of the input parameters. In particular, there are three classes of models that can be used. First, the model may be a complex mathematical function (for example, defined by a differential equation). Second, the model may be a simulation model, like an IBM. Third, the model may be the result of fitting a statistical model.

All these problems may be formulated in a general way, defining some response from the model $\mathbf{Y}$ as a function of the input parameter vector $\mathbf{x}$:

$$\mathbf{Y} = \mathbf{f}(\mathbf{x}) \tag{2.5}$$

In the equation 2.5, all the quantities are vectors, indicated by the boldface. Here, $\mathbf{x} = [x_1, x_2, \ldots, x_m]$ represent the parameters to the model $\mathbf{f}$, and $\mathbf{Y} = [y_1, y_2, \ldots, y_n]$ represent the some quantitative responses from the model. In some sections, we will discuss the response as a single value y, without loss of generality.

Each of the input parameters x_i is associated with a probability distribution $D_i(x)$, which represent our degree of knowledge about the values that x_i may assume (see figure 2.2 for examples); Berger [1985] provides a more detailed discussion).

Taken together, all the distributions D_i form the *joint probability distribution* of the parameters, $\mathbf{D}(\mathbf{x})$. This function takes into account not only the individual distribution of each parameter, but also all the correlation terms between them[1].

In very simple models, it may be possible to analytically deduce the behavior of the model response taken at each point of the joint distribution of parameters. In the general case, however, this is impossible, and a way of investigating the model is to choose some points from the joint distribution and analyzing the model at each point. Section 2.2 will present some strategies for choosing these points.

2.2 Sampling Techniques

There are several strategies that can be used to choose the samples from the parameter space that will be used as input to our model of interest. Here, we will present some of them, along with their limitations, to justify our choice for the Latin Hypercube Sampling, which we will describe in section 2.2.1.

One way of exploring the parameter space is by discretizing every distribution and running the model for every possible combination of values for all parameters. This is called full parameter space exploration, as done by Turchin and Hanski [1997], and although it possesses many advantages, it may become very costly in terms of computer time. In addition, the number of possible combinations increases exponentially with the number of parameter dimensions considered.

[1]For clarity, it should be noted that the joint probability distribution is *not* used explicitly in the methods discussed here. Only the marginal distributions and correlation terms, if necessary, are explicitly used.

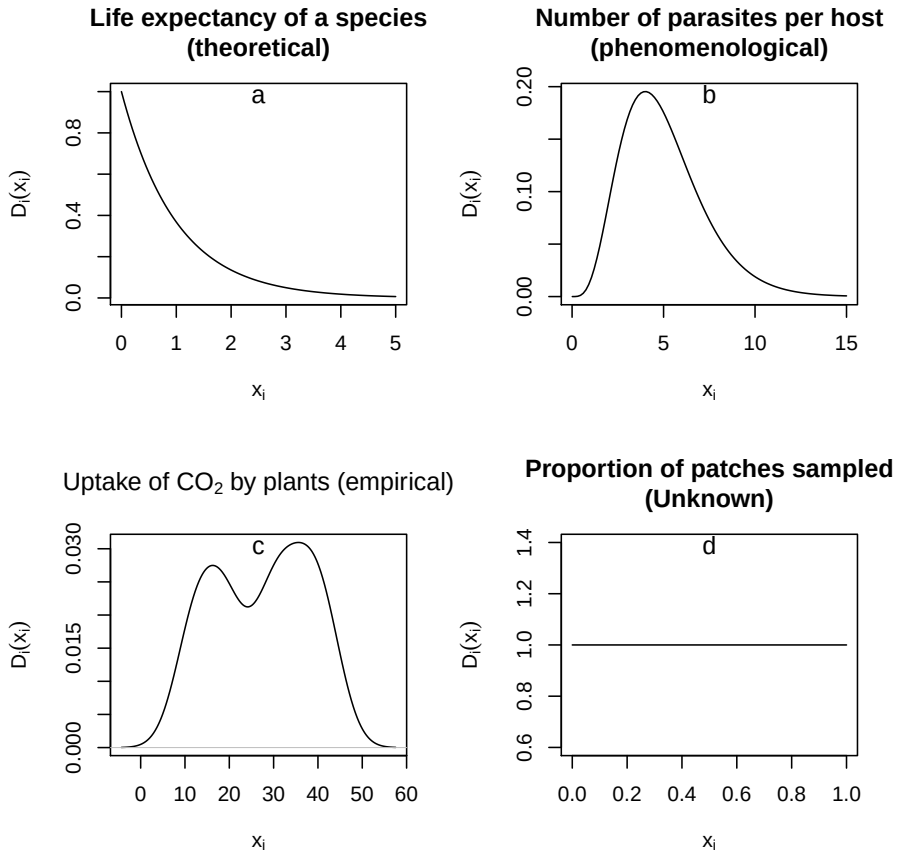


Figure 2.2: Four different possibilities for choosing the distributions D_i. Panel "a" shows the exponential distribution of life expectancy, which can be deduced from theory [Cole, 1954]. Panel "b" shows a gamma distribution, which can be used to model the number of parasites per host, but has no theoretical derivation [Bolker, 2008]. Panel "c" shows data from an empirical study on the CO_2 uptake from plants [Potvin et al., 1990], and panel "d" shows an example where no prior information can be used.

To circumvent the exponential increase in the number of samples, it is usual to explore the parameter space in the following fashion: holding all but one parameter constant, we analyze how the output of a model is affected by one parameter dimension at a time (as done by Yang and Atkinson [2008]). This analysis is referred to as individual parameter disturbance. This kind of analysis is, however, limited by the fact that the combinations of changed parameters may give rise to complex and unexpected behaviors. Often, algorithms based on individual parameter disturbance are used as a first step in order to discriminate between parameters that may have a substantial impact on the output, and parameters that are less relevant (but see [Morris, 1991] for an alternative).

Another viable option would be to chose N random samples from the entire space, in order to analyze both the effect of each parameter and the combined effect of changing any combined number of parameters. This sampling scheme is called random sampling, or Monte Carlo sampling, and has been applied to many biological models [Letcher et al., 1996]. One important feature of the Monte Carlo sampling is that its accuracy does not depend on the number of dimensions of the problem [MacKay, 2003].

Stratified sampling strategies, which are a special case of Monte Carlo sampling, consist in strategies for choosing these random samples while, at the same, making sure that each of the subdivisions (or *strata*) of the distribution are well represented. As shown by McKay and Beckman [1979], the estimates of statistical properties (such as the mean or the variance) of the model output are better represented by stratified random sampling than by simple random sampling (see figure 2.3 for examples). As we shall see in the next session, the Latin Hypercube sampling is a practical and easy to understand stratified sampling strategy.

Another class of Monte Carlo methods that should be mentioned here is the Markov Chain Monte Carlo (MCMC), which is also used on similar analyses [MacKay, 2003]. This methods consists in generating a sequence of points $\{\mathbf{x}^{(t)}\}$ from the parameter space whose distribution *converges* to the joint probability distribution $\mathbf{D}(\mathbf{x})$, and in which each sample $\mathbf{x}^{(t)}$ is chosen based on the previous $\mathbf{x}^{(t-1)}$. MCMC methods perform better than LHS methods for estimating the distribution of the model responses, however, they require a number of model runs which is orders of magnitude higher than LHS requirements.

2.2.1 Latin Hypercube: Definition and use

In this section, we describe the Latin Hypercube Sampling, and show how it can be used to efficiently solve the questions posed in section 2.1. We also discuss what are the available methods for obtaining the LHS.

Firstly, let us define, in the context of statistical sampling, what is a Latin Square:

Definition If we divide each side in a square in N intervals, and then take samples from the square, the resulting square will be called Latin if and only if there is exactly one sample in each row and each column.

A Latin Hypercube is simply the generalization of the Latin Square to an arbitrary number of dimensions m. It should be noted, then, that the number of

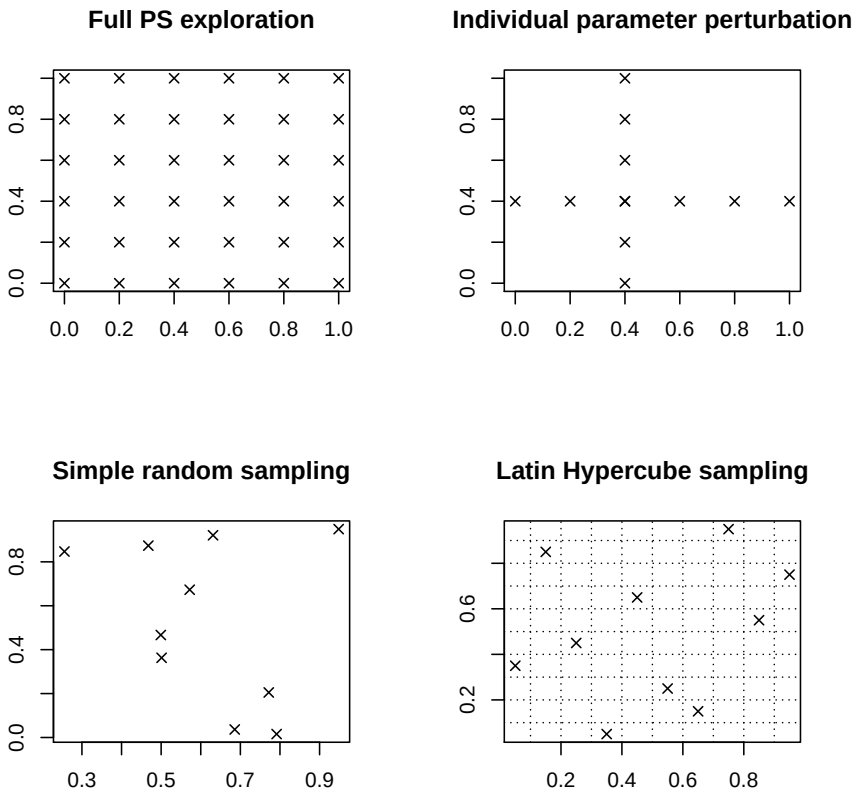


Figure 2.3: Illustration of four sampling methods. While the full parameter space exploration is clearly representative of the whole space, it requires a very large number of samples. The individual parameter perturbation chooses samples by holding one parameter constant and varying the other, and clearly cannot take into account interactions between parameters. The random sampling uses information about the whole parameter space with a small number of samples, but can oversample some regions while under sampling others. The Latin Hypercube (section 2.2.1) samples all the intervals with equal intensity.

samples N is fixed *a priori*, and does not depend on the number of parameters considered.

We will now construct the Latin Hypercube. Let's fix our attention in one parameter dimension i of the parameter space. The first step we should take is to divide the range of x_i in N equally probable intervals. In order to do so, we will turn our attention to the probability distribution of x_i, defined on section 2.1.1 as D_i. Recall that this probability distribution must be chosen in a way that represents our current understanding of the biology of the given system. This function might be estimated by an expert in the field, it might represent a data set from field or laboratory work, or in some cases it may be simply the broadest possible set of parameters, in some cases where the actual values are unknown or experiments are unfeasible (see fig. 2.2).

In possession of the distribution function D_i, we must sample one point from each equally probable interval. There are two approaches used here: it is possible to choose a random value from within the interval [McKay and Beckman, 1979], or instead, we can use the midpoint from each interval [Huntington and Lyrintzis, 1998]. As the statistical properties of the generated samples are very similar, we will use the second approach here.

The integral of the distribution function is called the cumulative distribution function $F_i(x)$. This function relates the values x that the parameter may assume with the probability p that the parameter is less than or equal to x. We will refer to the inverse of the cumulative distribution function, F_i^{-1}, as the quantile function of the parameter x_i, as it associates every probability value p in the range $(0, 1)$ to the value x such that $P(x_i \leq x) = p$. We divide the range $(0, 1)$ in N intervals of size $1/N$, and use this quantile function to determine the x values as the midpoints of each interval. Summarizing, we take the N points, represented as $x_{i,k}$, $k \in [1, N]$, from the inverse cumulative distribution $F_i^{-1}(x)$ as[2]:

$$x_{i,k} = F_i^{-1}\left(\frac{k - 0.5}{N}\right) \tag{2.6}$$

The samples from each dimension are subsequently shuffled, to randomize the order in which each value will be used (see example on figure 2.4). As the samples come from the distributions D_i, and are only reordered, their (marginal) distribution will remain that of D_i. However, the joint distribution of the parameters is still not well defined. In particular, this simple shuffling may result in some of the parameters to be positively or negatively correlated with each others, which might be undesirable. Some techniques have been developed to eliminate these correlation terms or to impose different correlations between the variables, and will be presented on section 2.2.2.

It should be noted that, in the mathematical literature, it is usual to refer to a somewhat different object as a Latin Square: this would be a square whose sides are divided in N intervals, and is filled with N different symbols, such that for each row and column there is exactly one occurrence of each symbol, as represented in figure 2.5.

[2]This formula is given for simplicity; see [Huntington and Lyrintzis, 1998] for an alternative with better numerical properties

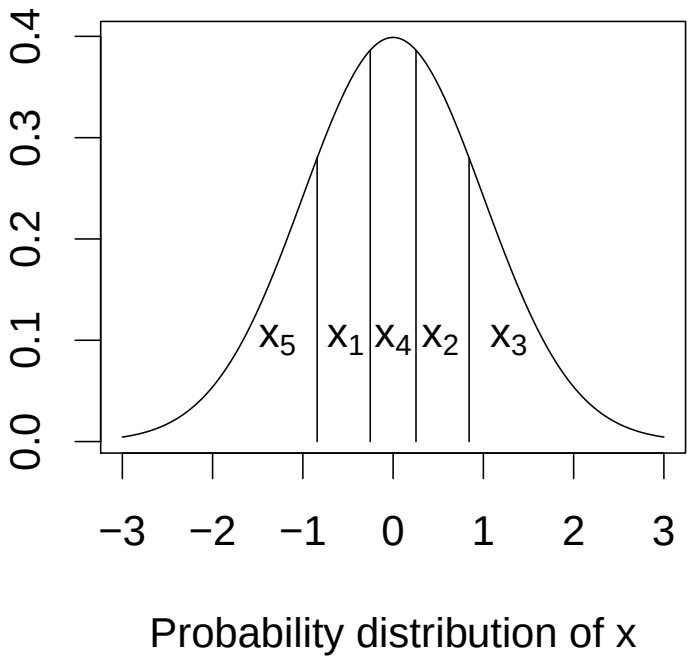


Figure 2.4: Sample normal probability distribution, with 5 samples collected from regions with same probability and shuffled. Note that the first sample correspond to the second interval, the second sample correspond to the fourth interval, and so on.

Figure 2.5: A stained glass window at the Caius College, Cambridge, showing a full Latin Square. Notice how there is only one occurrence of each color in each row and in each column.

2.2.2 Algorithms and extensions

As described above, the LH sampling generates an uniform distribution of samples in each parametric dimension. However, there is no guarantee that the correlation between two or more parameters will be zero, and the classical algorithm from McKay, described in the previous section, usually produces correlations as high as 0.3 between pairs of factors, which can difficult or even compromise further analyses. In this section, we will present one algorithm designed to take into account the correlation between the parameter variables [Huntington and Lyrintzis, 1998], using a single-switch-optimized sample reordering scheme. We will present the general case of prescribing a correlation matrix, and will also present results for the trivial case of zero correlation terms. Other methods have been proposed to address this problem [Florian, 1992; Steinberg and Lin, 2006; Ye, 1998], including methods that deal with higher-order correlation terms [Tang, 1998] using orthogonal designs and methods that resort to stochastic optimization based on simulated annealing [Vořechovskỳ and Novák, 2009]. These methods, however, either impose severe restrictions on the number of samples that must be chosen or are too computationally intensive.

In order to obtain the samples with prescribed correlation terms, we define as $C_{i,j}$ the desired $m \times m$ correlation matrix between the variables x_i and x_j, and denote by $C^*_{i,j}$ the current correlation between x_i and x_j.

The next step is done iteratively for each parameter dimension, starting with the second one. Suppose that the method has already been applied to $i = 1, 2, ..., l-1$, and we will apply it to $i = l$. The square sum of the errors in the correlations between x_l and the anterior parameters is given by

$$E = \sum_{k=1}^{l-1} \left(C_{l,k} - C^*_{l,k}\right)^2 \tag{2.7}$$

Afterwards, we calculate, for each pair of values sampled from the parameter dimension l, what would be the error in the correlation if they were switched. The pair that corresponds to the greater error reduction is then switched, and the procedure is repeated iteratively until the error is acceptably small.

Note on the existence of solutions

The problem of generating a sample with specified marginal distributions and correlation terms is tightly connected to the problem of generating samples from a multivariate distribution. This class of problems is very complex, and has received limited attention from the probabilist community, with the recent paper describing the exact construction of all feasible bivariate exponential distributions being considered a significant theoretical advance [Bladt and Nielsen, 2010].

One surprising result by Hoeffding [1940] (*apud* [Dukic and Marić, 2013]) is that the specified distribution function need not exist. For any pair of marginal distributions D_1 and D_2, there exist a maximum and a minimum correlation coefficients, ρ^+ and ρ^- such that there exists a joint probability distribution $\mathbf{D}$ with specified marginal distributions and correlation. For some marginal distributions,

such as the Gaussian, these values are $\rho^+ = 1$ and $\rho^- = -1$, so that any correlation term results in a valid distribution. On the other hand, if both D_1 and D_2 are exponential distributions with parameter $\lambda = 1$, the values change to $\rho^+ = 1$ and $\rho^- = 1 - \pi^2/6 \approx -0.65$.

In short, finding a Latin Hypercube with exponential marginal distributions and at least one correlation term of, for example, -0.9 is impossible, no matter which algorithm is employed. However important this result is for the theory of probability, it may not have strong consequences in practical applications. If the marginal distributions and correlation terms are chosen after theoretical considerations or real world data, the impossible probability distributions will not be attempted.

For recent developments in this area, see the work of Dukic and Marić [2013] and Huber and Marić [2014].

2.2.3 Stochastic models

When dealing with stochastic models, like several relevant individual based models (IBM), the questions presented become complicated by the fact that running the same model with exactly the same parameters might wield largely different results, both quantitative and qualitatively. In this scenario, we must be able to differentiate the variation in responses due to the variation of the parameters with the variation in response due to stochastic effects.

We will refer to the variation due to the input parameters as the epistemic uncertainty. This uncertainty arises from the fact that we do now know what are the correct values for a given parameter in a given natural system, and is related to the probability distributions D_i, presented in section 2.1.2. The variation in the behavior of the model which is caused by stochastic effects, for a fixed set of parameters, is called stochastic uncertainty, and is inherent to the model.

It is important to note that the two uncertainty components are impossible to disentangle in general stochastic models. This has prevented the general analysis of such models until recently. In recent years, studies have shown that the important parameters and their effects can be correctly identified by running such models repeatedly for the same input variables and then averaging the output [Segovia-Juarez et al., 2004], given that the following conditions are respected:

- Sample sizes should be large, relative to the stochastic uncertainty.
- The output values should be unimodal, that is, the output values for a given parameter choice should be clustered around a central value.
- The correct analysis tools should be used (as will be discussed on session 2.3).

2.2.4 Measuring the concordance with increasing sample size

We will now turn our attention to the problem of determining the optimal number of model runs we should apply in order to provide a good estimate of which are the relevant parameters for a given model. One way of proceeding is by systematically

increasing the number N of model runs and applying any of the sensitivity analysis techniques, which will be discussed on the following section. If the analyses indicate similar results for consecutive runs, we can presume that increasing the sample size will not yield major changes to the results.

All of the sensitivity analyses present us with a list of the parameters that have most influence in the model output. By comparing the resulting lists from two experiments, we can decide to stop increasing N when the lists are sufficiently similar. Our problem then is to determine how similar are two vectors of ranks. In principle, we could apply any distance function to those vectors. However, consider that 3 analyses indicated that the order of the most influential parameters is:

H1	=	1	2	3	4	5	6
H2	=	1	2	3	6	4	5
H3	=	2	3	1	4	5	6

By using standard distances (like Spearman's rho or Kendall's tau), we will see the same difference between H1 and H2 and between H1 and H3. On the other hand, in the context of determining the most influential parameters, we would be inclined to see H1 and H2 as more similar than any of them to H3, as the first two preserve the ordering of the three first parameters.

Iman and Conover proposed a correlation coefficient for this problem called Top-Down Correlation Coefficient [Iman and Conover, 1987], which is based on Savage Scores. This coefficient, know as TDCC, was extensively used for sensitivity analyses [Marino et al., 2008]. Another measure of concordance proposed more recently is the Symmetrized Blest Measure of Association (SBMA) [Genest and Plante, 2003]. Recent research suggests that estimates for SBMA produce a smaller standard error than TDCC without the assumption that there is no correlation between the variables [Maturi and Elsayigh, 2010]. Thus, we propose using SBMA as a measure of concordance between analyses from different sample sizes. Defining the ranks from the first sample as R_i and the ranks for the second sample as S_i, the estimator for the SBMA is:

$$\xi_n = -\frac{4n+5}{n-1} + \frac{6}{n^3-n}\sum_{i=1}^{n} R_i S_i \left(4 - \frac{R_i + S_i}{n+1}\right) \tag{2.8}$$

For the techniques that may output negative values, for instance negative correlations, the SBMA must be applied on the absolute values. Otherwise, the parameters which present strong negative effects will be ranked very low, and will not be taken into account by the SBMA.

We will apply SBMA for the PRCC technique (discussed on section 2.3) on the example on section 2.4.

2.3 Quantitative output analysis

2.3.1 Uncertainty analysis

The first question we would like to answer, in the context of quantitative analysis, is what is the probability distribution of the response variable y given that we know

the join probabilities of the input parameters $\mathbf{x}$ (see definitions in section 2.1.2), which is the subject of uncertainty analysis [Helton and Davis, 2003].

This can be done by fitting a density curve to the output y or an empiric cumulative distribution function (ecdf). If there is any theoretical reason to believe that the distribution of y should follow one given distribution, it is possible to fit this function to the actual output data and estimate the distribution parameters. If the joint distribution of the input parameters correspond to the actual probability of some natural system to exhibit some given set of parameter values (as opposed to the case where we have no biologically relevant estimates for some parameters), the estimate represented by the density and ecdf functions approaches the actual distribution that the variable y should present in nature. This functions may be used, for example, to provide confidence intervals on the model responses.

However, this is only the case when the input variables are uncorrelated or when enough correlation terms have been taken into account. Smith [2002] provides an example where ignoring the correlation terms leads to inaccuracies on the estimation of confidence intervals.

The next reasonable step is to construct and interpret scatterplots relating the result to each input parameter. These scatterplots may aid in the visual identification of patterns, and although they cannot be used to prove any relationship between the model response and input, they may direct the research effort to the correct analyses. There are extensive reviews of the use of scatterplots to identify the important factors and emerging patterns in sensitivity analyses [Kleijnen and Helton, 1999].

We will present here some quantitative analyses tools, aimed at identifying increasingly complex patterns in the model responses. It should be stressed that no single tool will capture all the relations between the input and output. Instead, several tools should be applied to any particular model.

2.3.2 Sensitivity analysis

The question of "what is the effect of some combination of parameters to the model output" may be answered by testing the relation between the parameters and outputs. There are extensive reviews about detecting these relations after generating samples with Latin Hypercubes [Kleijnen and Helton, 1999; Marino et al., 2008], so we will give just a brief overview. We will first note that the methods used must take into account the variation of all the parameters. For example, instead of calculating the correlation between the result and some parameter, partial correlation coefficients should be used, which discount the effect of all other parameters.

The classical approach to the sensitivity analysis, based on the frequentist school of hypothesis testing, consists in classifying the relations between the results and the input parameters, in order of increasing complexity, as:

- Linear relation, which can be tested with the Pearson partial correlation coefficient. Is is usual to test the significance of this linear relation by a t-test [Freedman et al., 2007].

- Monotonic relation, which can be tested with the Spearman partial correlation coefficient, also referred to as Partial Rank Correlation Coefficient, or PRCC. This measure is a robust indicator of monotonic interactions between y and x_i, and is subject to significance testing [Marino et al., 2008].
- Trends in central location, for which the Kruskal-Wallis test may be applied [Kleijnen and Helton, 1999].
- Trends in variability, for which the FAST method and Sobol' indexes may be used in order to partition the model variability [Archer et al., 1997; Saltelli, 2004; Saltelli et al., 1999].

Subsections 2.3.2 to 2.3.2 will provide some mathematical background for each method, and section 2.4 will present examples of use of those tests. We should stress here that the application of one method is not enough to draw conclusions about the relations between the input and output variables, as these techniques test different hypotheses, and have different statistical powers. Instead, every model should be analyzed by a combination of techniques, preferably one for each category outlined here.

Linear relation

Under the hypothesis of independence between the central location and dispersion of the model responses, the most straightforward relationship between y and x_i is the linear, represented by $y \sim x_i$. This is case if, every time x_i is increased, y increases by approximately the same amount. The Pearson correlation coefficient is the commonly used measure to test for a linear correlation:

$$\rho_{yx_i} = \frac{\sigma_{yx_i}}{\sigma_y \sigma_{x_i}} \tag{2.9}$$

Where σ_a is the variance of a and σ_{ab} is the covariance between a and b. The correlation coefficient is a measure of the predicted change in y when x_i is changed one unit, relative to its standard deviations, and, as such, approaches ± 1 when there is a strong linear relation between the variables. The square of ρ, usually written as R^2, measures the fraction of the variance in the output that can be accounted for by a linear effect of x_i. Is is usual to test the significance of this linear relation by a t-test [Freedman et al., 2007].

Other than examining the individual relationships between the parameters and the output, we can investigate the joint effect of several x_i, as $y \sim x_1 + x_2 + \cdots + x_m$. In this case, the multiple R^2 represent the fraction of the variance on the output due to linear effects of all the x_i considered.

However, a measure of ρ close to zero does not mean that no relationship exists between y and x_i - for instance, $x^2 + y = 1$, $x \in [-1, 1]$ presents $\rho = 0$, so clearly other methods might be needed.

The Partial Correlation Coefficient (PCC) between x_i and y is the measure of the linear effect of x_i on y after the linear effects of the remaining parameters have been discounted. In order to calculate the PCC, first we fit a linear model of x_i as a function of the remaining parameters:

$$\hat{x}_i \sim x_1 + x_2 + \cdots + x_{i-1} + x_{i+1} + \cdots + x_m \tag{2.10}$$

A corresponding model is done with y:

$$\hat{y} \sim x_1 + x_2 + \cdots + x_{i-1} + x_{i+1} + \cdots + x_m \tag{2.11}$$

The PCC is calculated as the correlation between the residuals of these two models:

$$PCC(y, x_i) = \rho\left((y - \hat{y}), (x_i - \hat{x}_i)\right) \tag{2.12}$$

Monotonic relation

Let us refer to each value of y as y_k and each value of x_i as x_{ik}. The rank transformation of y, represented by $r(y_k)$ can be found by sorting the values y_k, and assigning rank 1 to the smallest, 2 to the second smallest, etc, and N to the largest. The rank of x_{ik}, $r(x_{ik})$, can be found in a similar way.

If there exists a strictly monotonic relation between y and x_i, that is, if every time x_i increases, y either always increase or always decreases by any positive amount, it should be clear that the ranks of y and x_i present a linear relationship: $r(y) \sim r(x_i)$.

The correlation between $r(y)$ and $r(x_i)$ is called the Spearman correlation coefficient η_{yx_i}. The same analyses presented on section 2.3.2 can also be applied for the rank transformed data, including significance testing and multiple regression.

If the procedure described to calculate the PCC is followed on rank transformed data, that is, if y and x_i are rank transformed and fitted as linear models of the remaining parameters, the correlation between the residuals is called PRCC, or Partial Rank Correlation Coefficient. This measure is a robust indicator of monotonic interactions between y and x_i, and is subject to significance testing [Marino et al., 2008]. This measure will perform better with increasing N.

Trends in central location

Even if the relation between y and x_i is non monotonic, it may be important and well-defined. The case in which $y \sim x_i^2$, $x_i \in (-1, 1)$ is a common example. This relation may be difficult to visualize, and sometimes may not be expressed analytically. In these cases, the Kruskal-Wallis rank sum test may be used to indicate the presence of such relations [Kleijnen and Helton, 1999].

In order to perform the test, the distribution of x_i must be divided into a number N_{test} of disjoint intervals. The model response y is then grouped with respect to these intervals, and the Kruskal-Wallis test is used to investigate if the y values have approximately the same distribution in each of those intervals. A low p-value for this test indicates that the mean and median of y is likely to be different for each interval considered, and thus that the parameter x_i have a (possibly non monotonic) relationship with y.

The number of intervals N_{test} is not fixed as any "magical number", and may have a large impact on the test results. It is then recommended that this test

should be repeated with different values to obtain a more comprehensive picture of the interactions between x_i and y (fig. 2.6).

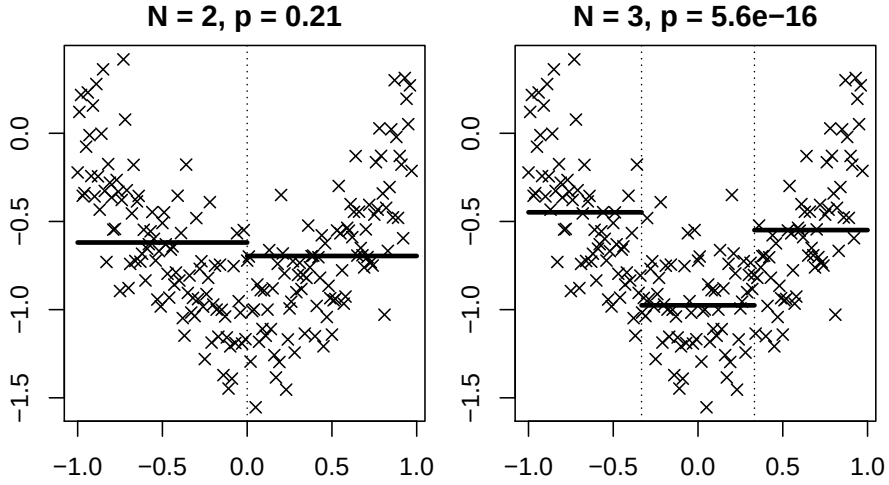


Figure 2.6: Example of application of the Kruskal-Wallis test on the same data set, which present a strong quadratic component, by dividing the range in 2 intervals (right) or 3 intervals (left). The dashed lines are the divisions between the intervals, and the strong horizontal lines are the sample means for each interval.

Trends in variability

Other than the central tendency of the results, their dispersal may be dependent on the input parameters. A classical approach which may be used to test whether the dispersal of the output is related to any input parameter is to divide the distribution of x_i into a number N_{test} of disjoint intervals and group the model response y with respect to these intervals, as done on the Kruskal-Wallis test. In this case, the ANOVA F statistic can be used to test for equality between the y conditional to each class [Kleijnen and Helton, 1999].

A similar approach, which we will use, is to employ the eFAST indexes [Saltelli, 2004; Saltelli et al., 1999], which is a variance decomposition method based on the FAST and Sobol' indexes. While the Sobol' indexes were described in 1969, in Russian, FAST was developed by Cukier et al. in 1973, and both are identical in all but one computation [Archer et al., 1997].

These methods estimate what fraction of the output variance can be explained by variation in each parameter x_i, which is called the *first-order sensitivity of* x_i or *main effect of* x_i. The method estimates as well the fraction which is explained by the higher-order interactions between x_i and all other parameters. The sum of all terms related to x_i is called the *total-order sensitivity of* x_i.

The eFAST method estimates the main effect of each parameter by choosing a periodic function $f_i(x_i)$ for each parameter, where the frequency ϕ_i of each function is distinct, and should, in theory, be incommensurable. Each of this functions is

sampled N_s times, and a Fourier analysis is applied to the model output. The Fourier coefficients at each frequency ϕ_i is related to the main effect of the variable x_i. The total order sensitivity of x_i is then calculated as the fraction of the variation which is not explained by the complimentary of x_i (that is, all parameters but this one).

There are two things that should be noted here about eFAST indexes. The first one is that the eFAST calculation does not involve the LHS sampling scheme, and may require more model evaluations. Also, this method produces small positive total-order sensitivity estimates even for parameters which do not play any role on the model output, as many numeric approximations are involved.

2.3.3 Bayesian alternatives

The Bayesian view of statistics present some alternatives to the techniques outlined in the previous sections. Beven and Binley described a procedure for the Bayesian updating of probabilities called GLUE, for Generalized Likelihood Uncertainty Estimation[Beven and Binley, 1992]. While the previously discussed methodologies are appropriate for purely exploratory analyses, the GLUE method is suited for problems in which one or more of the parameters of the model require calibration using the object of prediction. It is based on the notion that, for a given model response, there is always a set of models that will recreate it. This set is called *equifinal*.

Despite the widespread use and recognition that the GLUE method has received, it is subject to criticisms by not being formally Bayesian, and formal Bayesian approaches have been developed[Vrugt et al., 2009]. Another approach, based on the Metropolis algorithm, is provided by Kuczera and Parent [1998].

2.4 Case study 1: a structured model of *Euterpe edulis* populations

2.4.1 Model description

In this section, we demonstrate the uses and advantages of the methods outlined in the previous sections by performing sensitivity analyses on a density-dependent model of the tropical palm *Euterpe edulis* (commonly known as palmito juçara). All the data used here was extracted from Silva Matos *et al.* paper [Silva Matos et al., 1999], which compared a density independent matrix model of population growth with a density dependent model in which the recruitment of seedlings was affected by the number of seedings and adult trees. Silva Matos provided results, sensitivities and elasticities for the density independent model that can be compared to our findings, and results for mean and maximum values of the density dependent model - but unfortunately, their methods did not allow for a full sensitivity analysis of the density dependent model.

We have used the **R** language to perform the sampling and analysis, with the "pse" package, which implements the tools described in the previous sections. We

have also used code from the "sensitivity" package, which implements PRCC analysis (see section 2.3.2) and eFAST analysis (see section 2.3.2), among others. The "pse" package also implements the Huntington & Lyrintzis' algorithm to generate zero correlation between LHS samples (see section 2.2.1). All code used is freely available on the web.

The models analyzed are based on a Lefkovitch matrix with seven size classes. The matrix used on the density-independent model is

$$A = \begin{bmatrix} P_1 & 0 & 0 & 0 & 0 & 0 & F_7 \\ G_1 & P_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & G_2 & P_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & G_3 & P_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & G_4 & P_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & G_5 & P_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & G_6 & P_7 \end{bmatrix} \tag{2.13}$$

Here, P_i is the probability of a tree surviving and remaining in the same class (stasis), G_i is the probability of a tree surviving and growing to the next class, and F_i is the number of offspring produced per reproductive palm.

The dominant eigenvalue of this matrix is related to the predicted population growth rate. We considered the dominant eigenvalue for this matrix as the model output, as usually done on this modeling approach.

The density dependent model used the same matrix, but now the growth term of the first size class represented a decreasing function of the population density:

$$G_1 = \frac{G_m}{1 + aN_1} \exp\left(-\frac{z}{\rho} N_7\right) \tag{2.14}$$

Here, N_1 and N_7 represent the number of seedlings and adults per patch. The parameters G_m and a represent the maximum transition rate at low densities and the strength of reduction in G_1 with increasing seedling densities. The remaining parameters z and ρ represent the crown area of an adult tree and the plot size (which is fixed as $25m^2$), and their ratio is related to the reduction of recruitment due to the presence of adults, due to the fact that few seedlings are able to grow underneath the canopy of an adult.

As this model does not produce any static matrix, it is not meaningful to calculate any eigenvalue. Instead, the total population corresponding to the stable population distribution was used as model output.

A naïve approach to estimating the parameter sensitivities of this model would use the stasis, growth and fecundities given. However, this would yield erroneous results, as the probabilities of stasis and growth for a given class are not independent, as $P_i + G_i \leq 1$ for all classes. As discussed on section 2.1.1, we need to use an alternative parametrization for this model.

We will represent by s_i the probability of survival for each class, calculated as $s_i = P_i + G_i$, and by lowercase g_i the probability of growth, calculated as $g_i = (s_i - P_i)/s_i$. Using the notation for complementary probabilities $\overline{g_i} = 1 - g_i$, we can now write the Lefkovitch matrix as:

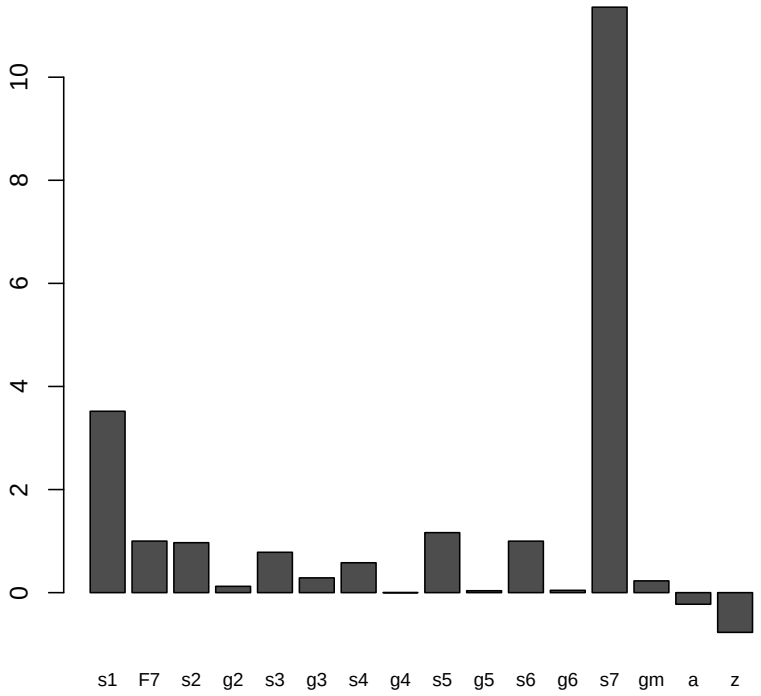


Figure 2.7: Analytical elasticities for the density-dependent model

$$A = \begin{bmatrix} s_1 \cdot \overline{g_1} & 0 & 0 & 0 & 0 & 0 & F_7 \\ s_1 \cdot g_1 & s_2 \cdot \overline{g_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & s_2 \cdot g_2 & s_3 \cdot \overline{g_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & s_3 \cdot g_3 & s_4 \cdot \overline{g_4} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_4 \cdot g_4 & s_5 \cdot \overline{g_5} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_5 \cdot g_5 & s_6 \cdot \overline{g_6} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_6 \cdot g_6 & s_7 \end{bmatrix} \tag{2.15}$$

The models have, respectively, 14 and 16 parameters. All analyses have been done with mean and standard deviation calculated from Silva Matos paper, assuming a normal distribution of parameters truncated at the $[0, 1]$ interval for probabilities, and on $[0, +\infty)$ for the other parameters. In the case of the density dependence parameters Gm, z and a, only the mean estimate was given on the paper, so conservative values were used for the standard deviations.

We have used the methods described in Caswell [2008, 2009, 2010] to estimate the analytical elasticities of the model in fig. 2.4.1.

2.4.2 Results

First, we have generated Latin Hypercubes consisting of all relevant variables for each model. Then, the models were run for each combination of parameters. By using the SBMA measure of concordance (section 2.2.4), we have determined that the sample size required for the density independent model is approximately 100, and between 300 and 500 for the density dependent (Table 2.1).

	Size	Independent	Dependent
1	50-100	0.87	0.51
2	100-200	0.86	0.49
3	200-300	0.92	0.89
4	300-400	0.93	0.82
5	400-500	0.91	0.87

Table 2.1: Comparison of PRCC analyses by sample size for both models

If we presume that the data collected is representative of our knowledge about each of these parameters, the probability distribution of the model responses can be seen as the probability that the real population of palms exhibit each value of the model output. Figure 2.8 shows these distributions, which suggest that the population is viable for the vast majority of parameters values in the parameter space considered. Also, the λ calculated from the density-independent model (mean 1.22, standard deviation 0.06), is very close to the value found by Silva Matos (mean 1.24 ± 0.06 se). Considering the density-dependent model, the median stable population predicted (6028 trees in each $25m^2$ plot), is comparable to, although higher than, the population actually measured by the study (1960 ± 560 trees per plot, mean and sd calculated over three years).

We have generated scatterplots between the result from the models and each independent parameter, in order to visually identify the relations between the inputs and outputs (figs. 2.9 to 2.12). It is clear from these scatterplots that the fecundity plays a major role on the population dynamics, and may be involved in non-linear interactions. Also, growth probabilities (g_i) have a greater impact on the model output than survival (s_i) on the density-independent model. This is to be contrasted with Silva Matos results, which show all of the elasticities to be approximately equal for all parameters. In the density-dependent model, the patterns are much more complex. Survival parameters seem to be more influent than growth, and the parameters reducing the recruitment (a and z) show a clear negative effect on the population size. However, there is evidence now for non-linear effects of the parameters, in particular s_1 and F_7.

These scatterplots show very high dispersion of values, mostly due to the fact that all parameters are being varied between runs. In order to investigate the effect of each parameter on the outputs discounting the effects of the others, we analyse the Partial Rank Correlation Coefficient (PRCC, fig. 2.13). The PRCC analysis for the density independent model is in agreement with our previous expectations, with F_7 being the most influential parameter, followed by growth probabilities. Survival probabilities follow with low correlations. The density dependent model

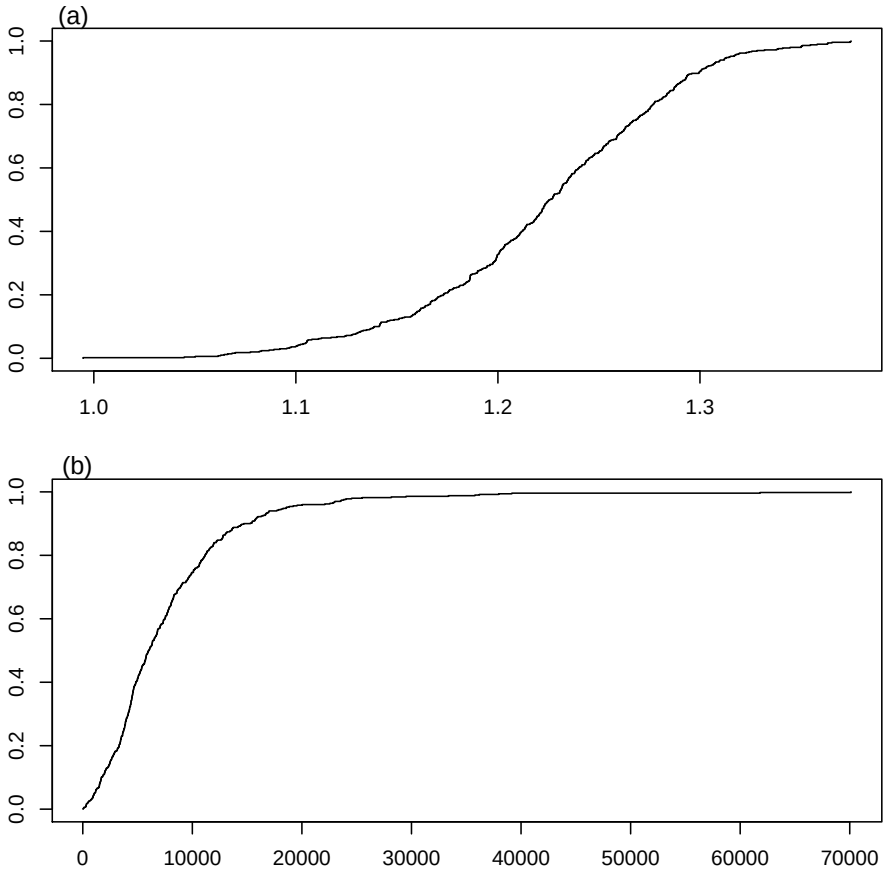


Figure 2.8: Empirical cumulative distribution functions (ecdf) for the density independent (a) and density dependent (b) models of population growth. In (a), the x axis represents the dominant eigenvalue, and the population is viable if $x > 1$. In (b), the x axis represents the total equilibrium population, and the population is viable if $x > 0$.

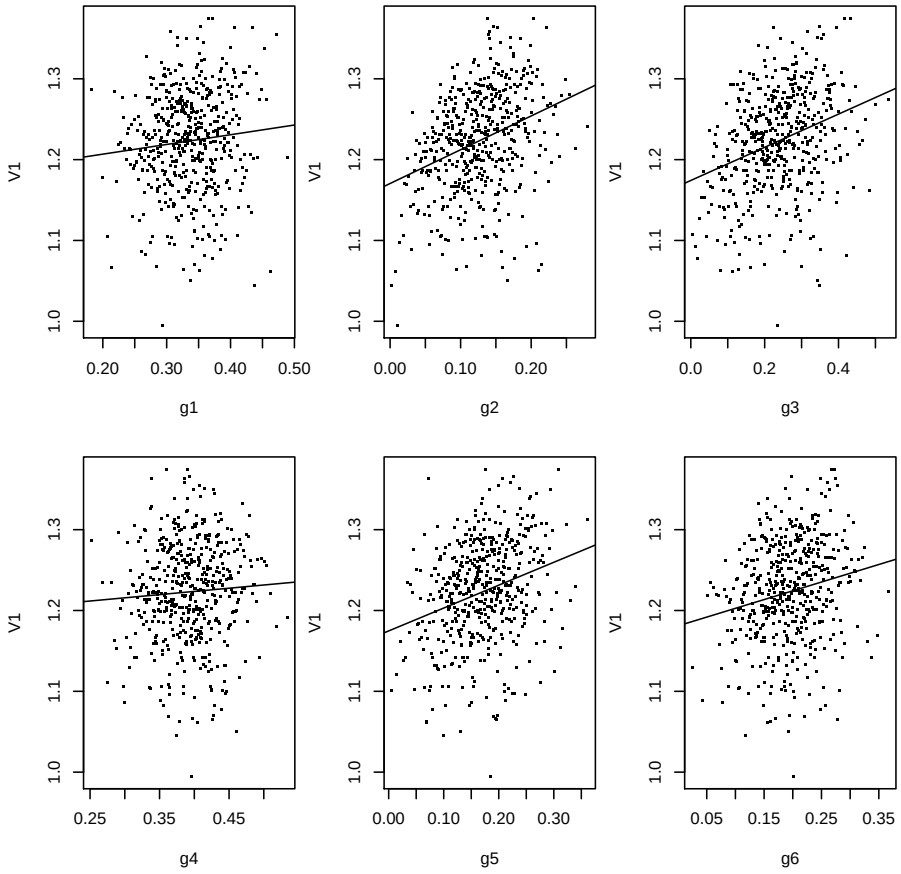


Figure 2.9: Scatterplots relating the value of the input parameters of growth to the λ calculated the output for the density independent model.

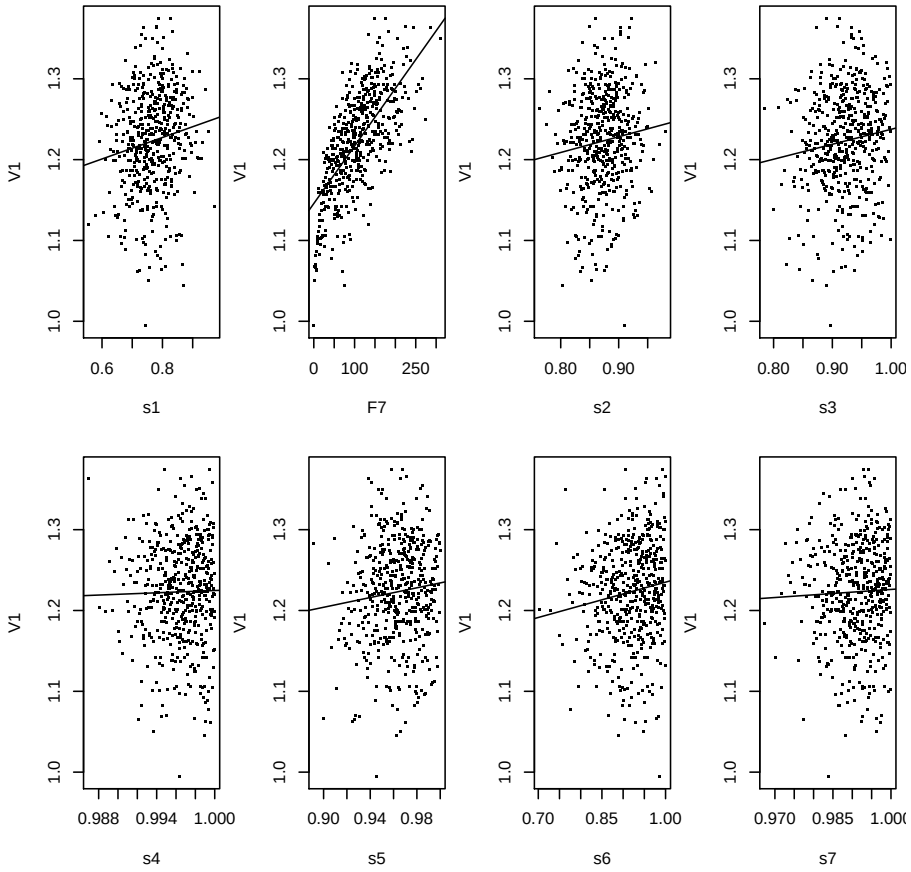


Figure 2.10: Scatterplots relating the value of the input parameters of survival and fecundity to the λ calculated the output for the density independent model.

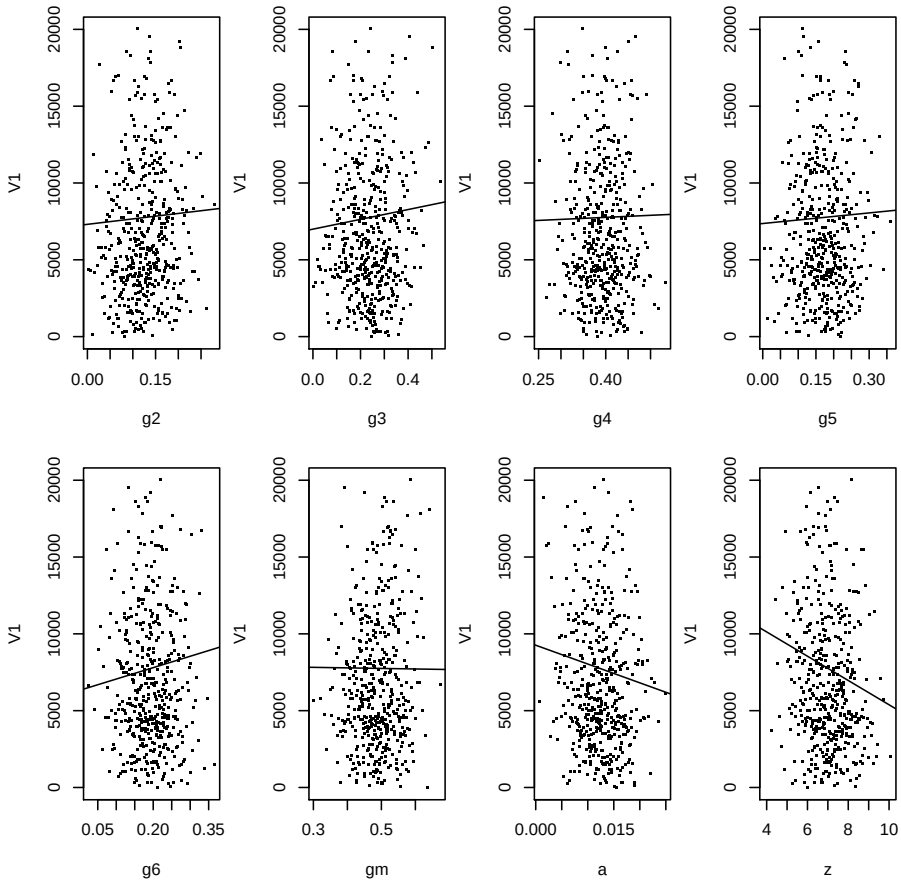


Figure 2.11: Scatterplots relating the value of the input parameter of growth and density-dependence to the output for the density dependent model.

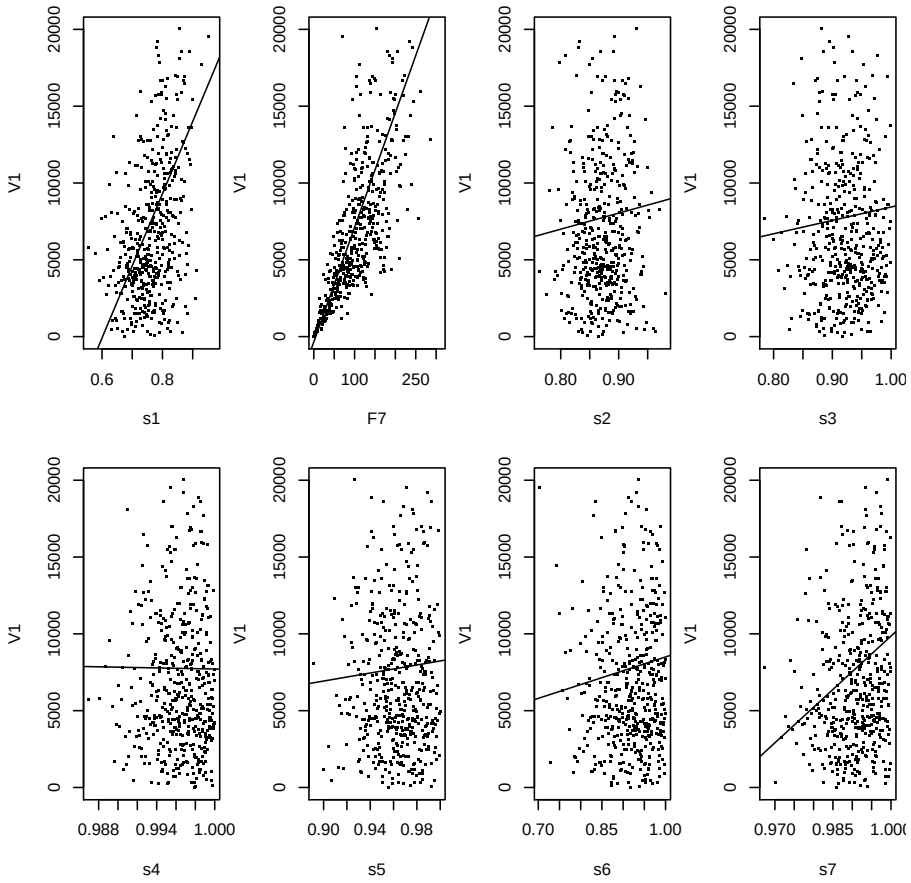


Figure 2.12: Scatterplots relating the value of each input parameter of survival and fecundity to the output for the density dependent model.

presents us with some surprises, as now the survival parameter for the smallest and largest size classes jumped to occupy the second and third largest positive PRCCs. The remaining parameters follow the new parameters a and z, which are strongly negatively correlated with the output.

It is interesting to contrast these results with the analytical analyses, presented on fig. 2.4.1. While all the elasticities have the same sign, and a comparable order, the most conspicuous difference between the two is the high importance given to s_7 (the survival of adults) given by the analytical analysis. This might be interpreted by noticing that this parameter is the one for which the collected data leaves the smallest margin of uncertainty.

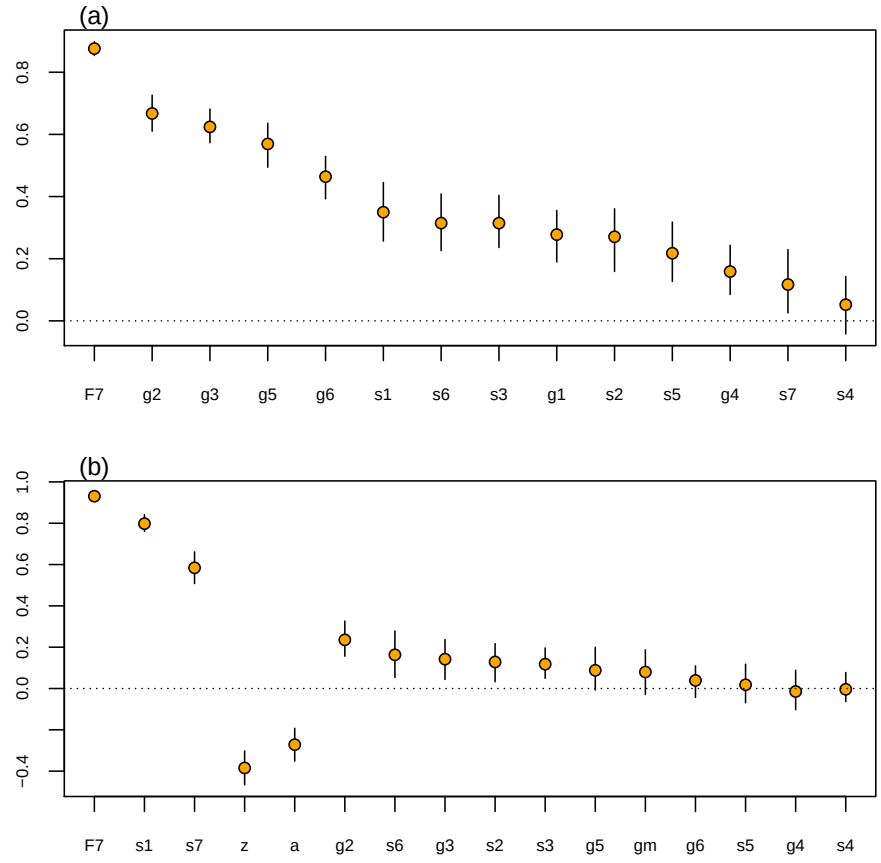


Figure 2.13: Partial Rank Correlation Coefficients for the density independent (a) and density dependent (b) models. The bars are confidence intervals, generated by bootstrapping 1000 times

The last analyses we present here are the decomposition of variance by Extended Fourier Amplitude Sensitivity Test (eFAST, fig. 2.14) and Sobol' methods. These analyses provides an estimation of the fraction of variation of model output that can be explained by the individual variation of each parameter (which we call first-order sensitivity, or main effect), along with the total variation caused by interaction between that parameter and others (total-order sensitivity). The interaction term

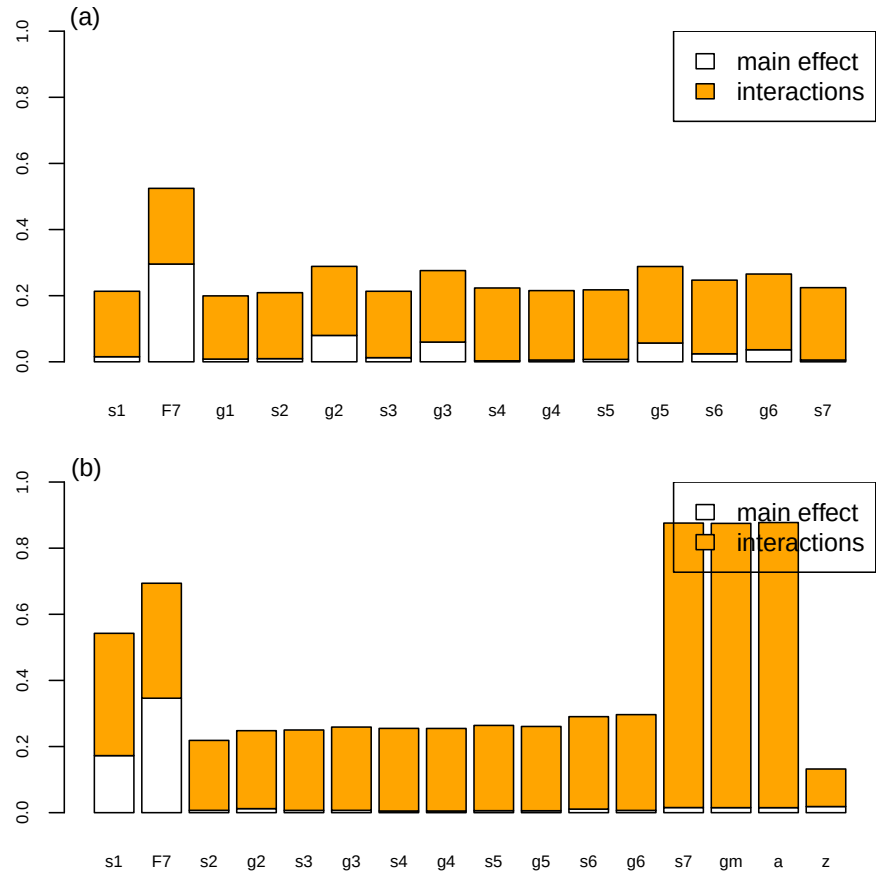


Figure 2.14: eFAST analysis for the density independent (a) and density dependent (b) models. The bars represent the first and total order estimates for the sensitivity of each parameter in the model output.

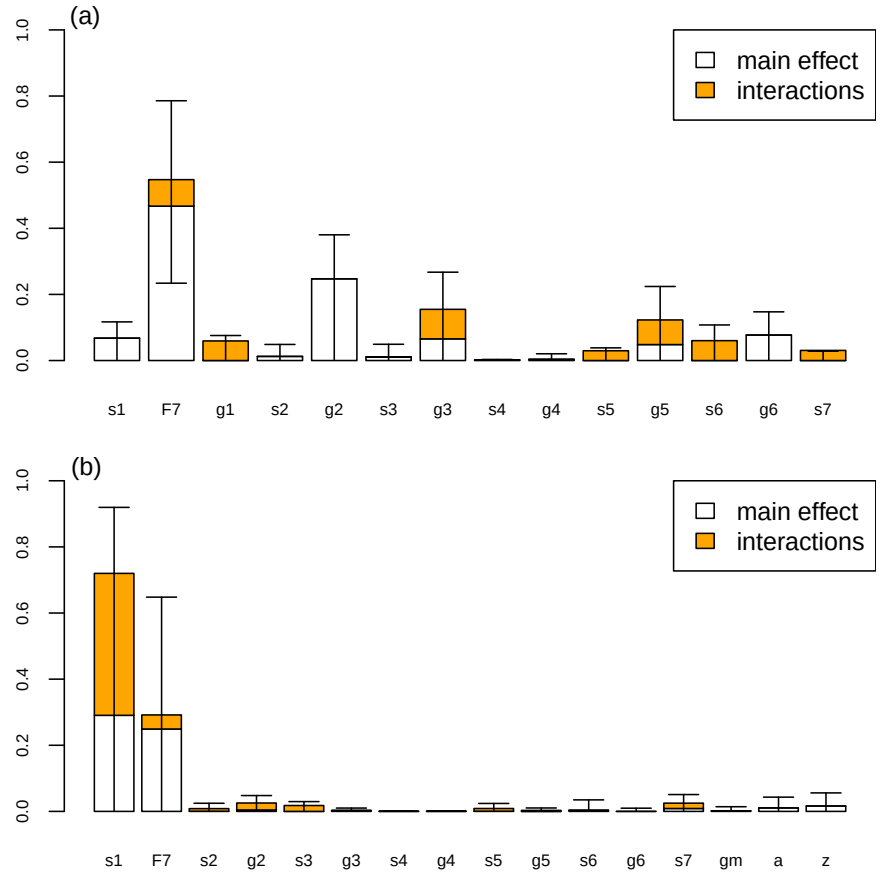


Figure 2.15: Sobol indexes for the density independent (a) and density dependent (b) models. The bars represent the first and total order estimates for the sensitivity of each parameter in the model output. Confidence intervals were generated by bootstrapping 1000 times.

for each parameter is the difference between its first and total order sensitivities. These analyses are substantially more intensive in terms of computer time than the previously mentioned. The eFAST analyses presented here required 7392 and 8448 runs of the simulations, respectively, for the independent and dependent cases. Table 2.2 shows the SBMA measure of concordance between different sizes. Note that N_s reported should be multiplied by the number of parameters in each model to obtain the total number of simulations executed. Also note that, while the main effect D_i converge for both model, the total order D_t indexes are still variable with large sample sizes. This difference in the results for several eFAST analyses, together with the differences between the eFAST and Sobol' results, hint at a numerical instability of this method. Sobol' indexes results are displayed on Fig. 2.15.

	Size (N_s)	Indep. D_i	Indep. D_t	Dep. D_i	Dep. D_t
1	66-132	0.78	0.85	-0.14	0.52
2	132-264	0.83	0.78	-0.00	0.75
3	264-528	0.97	0.93	0.95	0.66

Table 2.2: Comparison of eFAST analyses by sample size for both models

This analyses reveal that the output of the density independent model is mostly explained by first-order relations (which explain 62 % of the output variation, according to the eFAST analysis), with F_7 and the growth terms being the most important. The importance of the linear terms shouldn't come as a surprise, as the matrix growth model is a linear model. The density-dependent model, which has 66 % of the output variation predicted by linear terms, according to the eFAST analysis, exhibit more complex interactions between the terms, but while eFAST indicate the survival of adults and the terms associated with competition between seedlings as origins of these interactions, the Sobol' method points at higher-order terms involving the survival of seedlings.

2.4.3 Conclusions

The results from the uncertainty and sensitivity analyses presented show some of the advantages from the methodology described in this work that are unavailable to the usual framework used in ecological studies. First, we have been able to quantify the uncertainty in the asymptotic growth rate (related to λ) and the stable population size related to the uncertainty in the model inputs. Also, we provide a common framework to investigate side-by-side the linear and non-linear matrix models, from which we were able to point out the similarities and discrepancies between the models. Analyses based on matrix elasticities are also unable to investigate the role played by parameters not directly present on the matrix, as the size of the adult trees canopy z in our case. Finally, our approach allows the identification and quantification of relative importance of non-linearities and interactions between the input parameters in determining the model's outcome, and allows us to incorporate our previous knowledge about the system in specifying the range and distribution of each input parameter.

2.5 Case study 2: Non-linear structured model of *Tribolium* population

2.5.1 Model description

In this section, we present another example of performing uncertainty and sensitivity analyses to structured population models. This particular example was chosen because of its use by Hal Caswell in a series of papers published between 2008 and 2010, to illustrate the newly developed techniques of analytical sensitivity analyses. Although the theory which provides the tools to determine sensitivity and elasticity of linear matrix models has been established in the late 1970s, a consistent theory for the study of sensitivity of non-linear models (those in which the transition frequencies depend on the density or frequency or some size or age class) has only been fully developed in the recent years [Caswell, 2008, 2009, 2010].

In terms of the analytical analyses, the *sensitivity* of the result y in respect to a parameter x is given by the derivative $\frac{dy}{dx}$, and represents the additive effect that a small perturbation in x exerts over the result y. The *elasticity* of y in respect to x is given by $\frac{x}{y}\frac{dy}{dx}$, and represents the proportional effect of this perturbation.

To be able to compare the analytical results to a stochastic analysis based on Latin Hypercubes, we need to define how to estimate the elasticity using this methodology. The formula $\frac{x}{y}\frac{dy}{dx}$ needs to be adjusted in our context for two reasons: first, the derivative must be estimated with some numerical calculus, and second, the fraction $\frac{x}{y}$ is meaningless as there are no privileged points to base our analyses on. We propose the following formula as a candidate definition for the elasticity of y in relation to x, evaluated with a stochastic procedure:

$$\frac{\langle x \rangle}{\langle y \rangle} s_{yx} \tag{2.16}$$

Here, the brackets $\langle\rangle$ represent the average of a function, and s_{yx} is the linear partial correlation coefficient of y in relation to x. We stress that this is an arbitrary decision, but as we will show in the following subsection, it agrees with the analytical results for the investigated model.

The transition matrix for the *Tribolium* model is given by:

$$\mathbf{A}[\theta, \mathbf{n}] = \begin{bmatrix} 0 & 0 & b\exp(-c_{el}n_1 - c_{ea}n_3) \\ 1-\mu_l & 0 & 0 \\ 0 & \exp(-c_{pa}n_3) & 1-\mu_a \end{bmatrix} \tag{2.17}$$

Here, $\mathbf{n}(t)$ is the vector representing the beetle population, divided in three life stages: larvae, pupae and adult. The vector θ represents the model parameters, I.e., the vital rates used in the model.

The non-zero elements of this matrix, from left to right and from top to bottom, are:

- Adult fecundity, given by clutch size b times a term of cannibalism of eggs by adults (at a rate c_{ea}) and by larvae (at a rate c_{el});
- Maturation of larvae, reduced by base death rate μ_l;

- Pupae eclosion, reduced by cannibalism from adults at a rate c_{pa} (base pupae death rate is effectively zero);
- Permanence in the adult class, reduced by base adult mortality rate μ_a.

The best estimator for the parameter values, given by the original paper[Dennis et al., 1995], is

```
> b = 6.598
> cea = 1.155e-2
> cel = 1.209e-2
> cpa = 4.7e-3
> mua = 7.729e-3
> mul = 2.055e-1
```

Also, the original paper focuses on the metabolic equivalent of the beetle population from different life stages, which is given by $N_m(t) = \mathbf{c}^T\mathbf{n}(t)$, with $\mathbf{c}^T = (9, 1, 4.5)\mu lCO_2h^{-1}$. Thus, we will focus on $N_m(t)$, which is a scalar quantity.

Using the parameters given above, the model converges to a stable fixed point in which $N_m(t) = 1952$.

2.5.2 Elasticity analyses

The results of analytical elasticity analysis, following [Caswell, 2008] (and replicating the figure displayed in that paper), are displayed on figure 2.16[3].

A raise in the beetle's clutch size causes a positive change on the final value of $N_m(t)$. All other parameters have negative elasticities, with c_{ea} having the greatest impact.

After this analyses, we proceed to a stochastic exploration of the parameter space with the Latin Hypercube. All parameters are supposed to be normally distributed with a small dispersion (standard deviation of 1e-08). The result, presented on figure 2.17, shows that there is a good correspondence between both methods.

However, the analytical methods due to Caswell are limited to a very narrow neighborhood of the estimated parameter vector. If the measurement error is large enough that the linear approximation of the matrix becomes invalid, other methods are needed in order to study the sensitivity of the model. A purely analytical possibility is to take into account higher order derivatives of the matrix terms; however, that leads to a very fast growth of the complexity of the calculations. The stochastic approach to estimate sensitivity and elasticity of the parameters has the advantage of being performed in exactly the same way, regardless of how large is the uncertainty of the input parameters.

We have repeated the same analysis, but now with an uniform distribution of the parameters with very large ranges (from 0 to 1 in the rates, and from 2 to 12 in the clutch size), and have found out a much more complex figure, including non-linear and interaction terms between the parameters. Figure 2.18 shows the scatter plots, presenting a strong nonlinear response to the c_{ea} parameter and possibly complex

[3]The calculations are worked out step by step in the reference cited

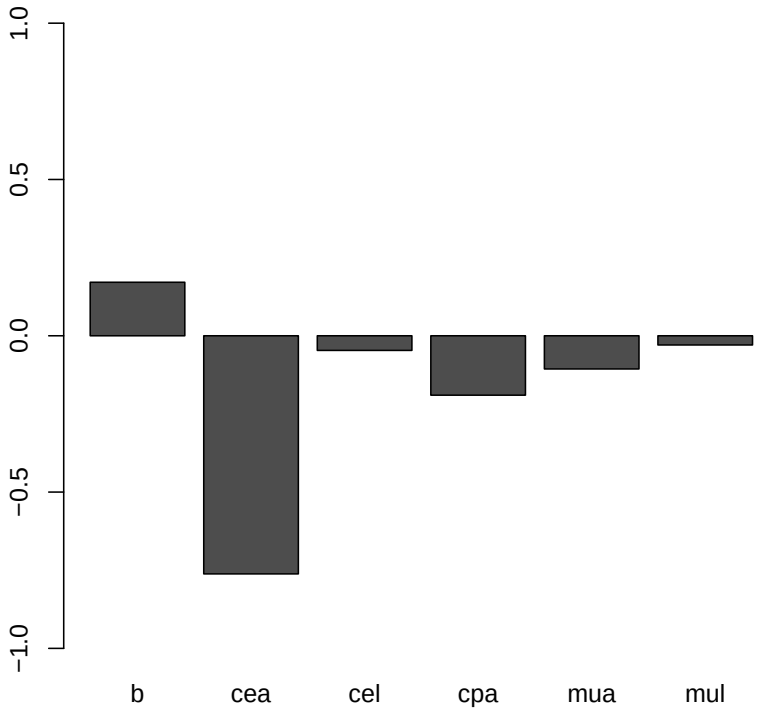


Figure 2.16: Analytical elasticity analysis for the structured population growth model of *Tribolium* beetles. Bars represent the elasticity of the metabolic equivalent of the equilibrium population in respect to each parameter.

interactions between parameters. In this analysis, we have excluded simulations where the population did not converge after 2000 time steps.

The corresponding elasticity analysis, now with a less restricted parameter space, are shown on figure 2.19. Even if no elasticity has changed its direction, all present a marked difference between the small and large perturbation scenarios. The biggest change occurs on parameter μ_l, which raises by 1509%.

Given this marked difference between the values, it is necessary to remember that this contradiction does not mean that one method is right while the other is wrong: each method is answering a different question.

2.6 Previous use of Latin Hypercube in ecology

The importance of sensitivity and uncertainty analyses in the development and use of ecological models is widely recognized. Searching for the terms "Sensitivity Analysis, Uncertainty Analysis or Parameter Space Exploration" in the Web of Knowledge reports 1199 papers in eight major journals (see fig. 2.20 and legend for details) since 1971, with 31650 total citations. However, most of these papers rely

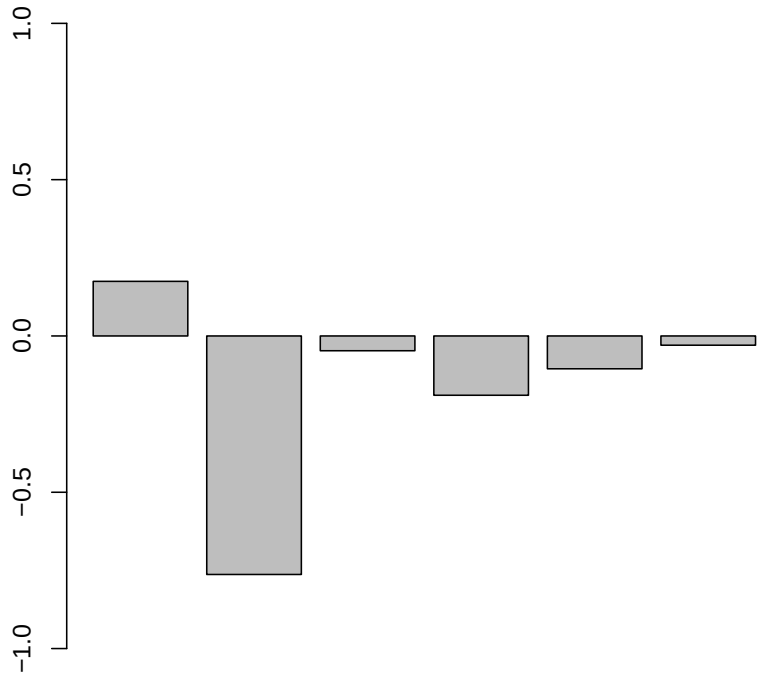


Figure 2.17: Stochastic elasticity analysis for the structured population growth model of *Tribolium* beetles, assuming small perturbations. Bars represent the elasticity of the metabolic equivalent of the equilibrium population in respect to each parameter.

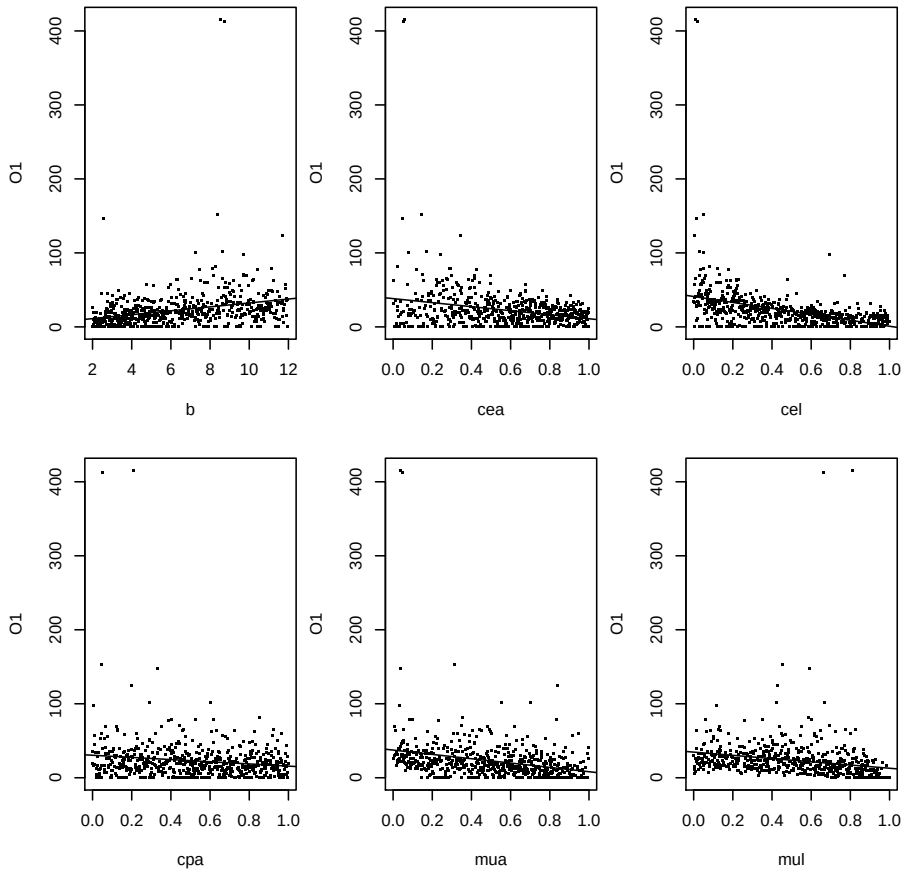


Figure 2.18: Scatterplots of the metabolic equivalent of the *Tribolium* beetle as a function of large changes in the input parameters for the population growth model

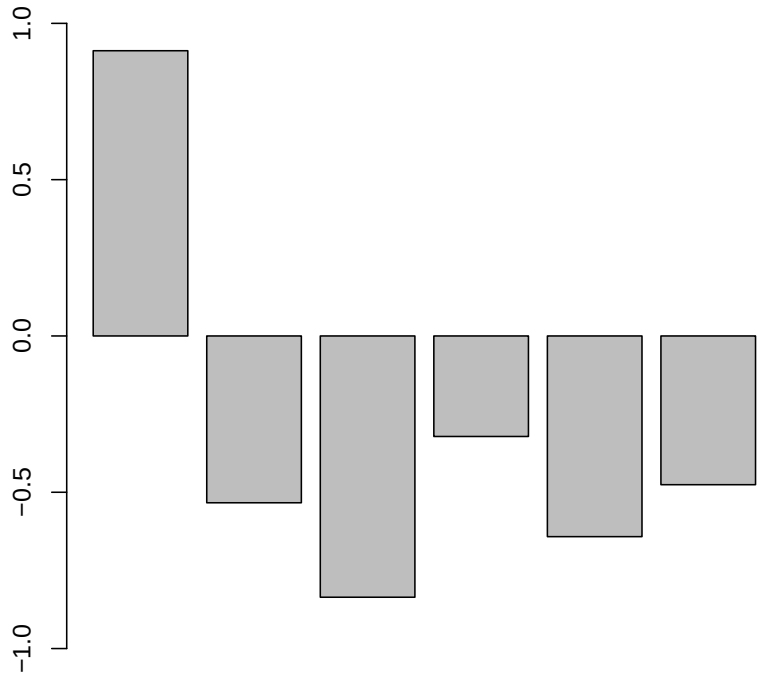


Figure 2.19: Stochastic elasticity analysis for the structured population growth model of *Tribolium* beetles, assuming large perturbations. Bars represent the elasticity of the metabolic equivalent of the equilibrium population in respect to each parameter.

on full and individual parameter space exploration, which, as discussed in section 2.2, are not optimal. When restricting those results with the keywords "Latin Hypercube, MCMC, Markov, Monte Carlo", just 120 works show up in the results. Of those, only 13 (about 10%) use Latin Hypercube Sampling [Berthaume et al., 2012; Confalonieri et al., 2010; Duchesne et al., 2003; Hamilton et al., 2010; Lovvorn and Gillingham, 1996; Marino et al., 2008; Meyer et al., 2007; Moore and Li, 2004; Nathan et al., 2001; Reed et al., 1984; Shirley et al., 2003; Tiemeyer et al., 2007; Xu et al., 2005]. There are also relevant examples of LHS use in other journals [Estill et al., 2012; Fisher et al., 2010; Thébault and Fontaine, 2010].

Also, many of these papers did not explicitly take into account the correlations between parameters. Those who did used mostly Iman and Conover's method [Iman and Conover, 1982].

These works have used Latin Hypercubes typically from 10 to 30 dimensions, but ranging from 6 to 143 [Berthaume et al., 2012], and the number of simulations ranged from 19 [Nathan et al., 2001] to 2000 [Tiemeyer et al., 2007]. Also, these works are from varied areas within ecology: applied plant ecology [Confalonieri et al., 2010; Tiemeyer et al., 2007], species richness [Hamilton et al., 2010], epidemiology [Shirley et al., 2003] and food chain analysis [Duchesne et al., 2003], stressing that the method is useful on varied problems.

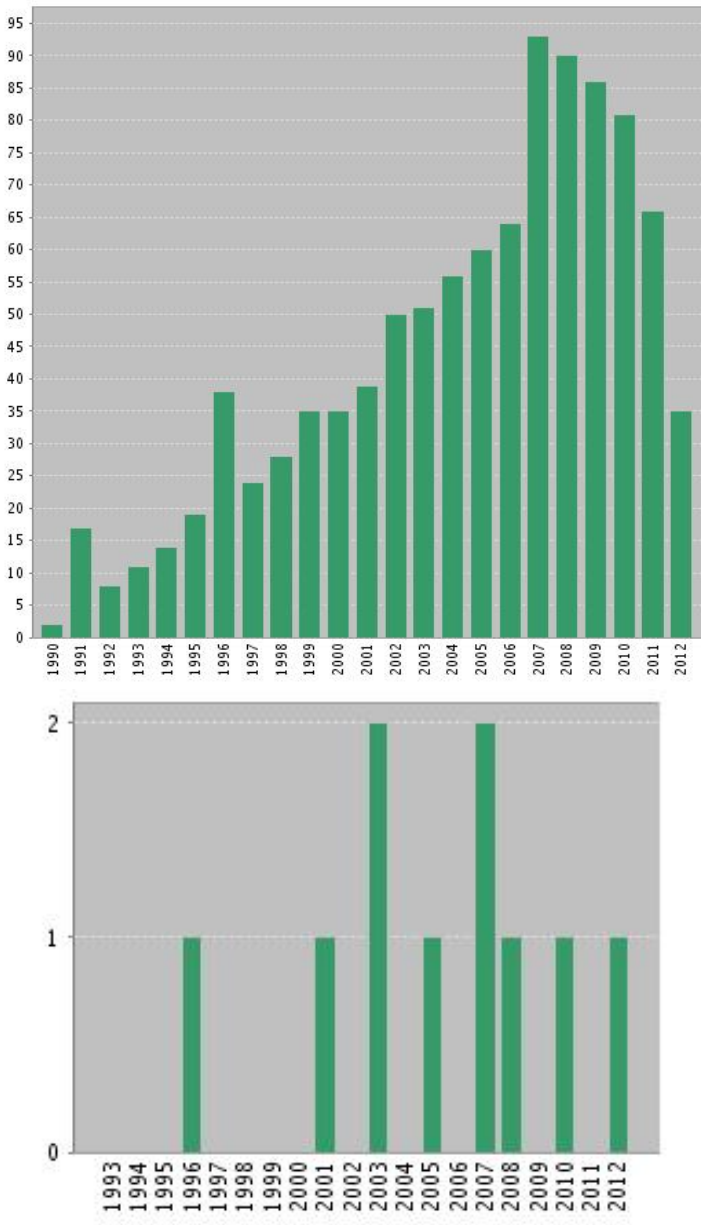


Figure 2.20: Top: Number of papers per year since 1990 containing the topics "Sensitivity Analysis, Uncertainty Analysis or Parameter Space Exploration" in the journals American Naturalist, Ecology, Journal of Ecology, Oikos, Oecologia, Ecological Modelling, Ecology Letters, Journal of Theoretical Biology as reported by Thomson Reuter's Web of Knowledge. Bot: Restriction of the search above to the keywords "Latin Hypercube". Search conducted 18.06.2012, 14h GMT.

Capítulo 3

Uma abordagem integrativa para análises de incerteza

3.1 Motivação

O desenvolvimento e utilização de modelos matemáticos e computacionais deve compreender as seguintes etapas: verificação, validação, calibração, análise de incerteza e análise de sensibilidade.

As duas primeiras etapas estão historicamente muito interligadas, e costumam ser referidas como V & V: a verificação consiste em garantir que o código computacional implemente o modelo desejado, enquanto a validação consiste em garantir que o modelo teórico é capaz de reproduzir os fenômenos reais desejados. Outra forma de pensar em V & V é que a validação se ocupa em determinar que o modelo está resolvendo as equações certas, enquanto a verificação se ocupa em determinar se o modelo está resolvendo as equações corretamente [Meisner, 2010]. Questões relacionadas a verificação de modelos não costumam ser formuladas estatisticamente, então não serão tratadas aqui.

A calibração consiste em ajustar os parâmetros implementados no código para melhorar a concordância entre o resultado obtido pelo modelo e aquele observado em uma determinada situação. A correta calibração de um modelo é considerada essencial para que o modelo possa ser considerado para fins preditivos [Meisner, 2010].

As duas últimas etapas já foram extensamente discutidas no capítulo 2, mas é importante frisar que é geralmente aceito que a calibração e as análises de incerteza e sensibilidade dependem fortemente da realização prévia de verificação e validação do modelo.

Durante o desenvolvimento e a análise de modelos matemáticos em ecologia, é comum separar a calibração por estimativa dos parâmetros (EP) da análise de incerteza e sensibilidade (AIS) do modelo. Os dois passos são feitos em seções diferentes dos artigos, discutidos em capítulos diferentes dos livros [Caswell, 1989], e chegam até a ser realizados por equipes diferentes. Em estudos que empregam modelos matriciais, a discussão dos resultados se concentra no ponto obtido por uma combinação de parâmetros considerada ótima durante as estimativas de parâmetros [Silva Matos et al., 1999], e a análise de incerteza é apresentada como um

procedimento separado e posterior à análise do resultado considerado como principal. Até mesmo a abordagem estatística usada em ambos os procedimentos pode ser incoerente.

Além disso, o conjunto de parâmetros usados na calibração do modelo pode ter pouca relação com as entradas individuais da matriz de projeção. Isso ocorre, por exemplo, quando séries temporais da população são utilizadas para estimar as taxas vitais: nesse caso, o responsável pela calibração vê o sistema como contagens de indivíduos, enquanto o responsável pela análise de incerteza tipicamente realiza a análise sobre as taxas vitais. Desta forma, a AIS é incapaz de apontar rumos para o planejamento de novos experimentos, apontando quais parâmetros devem ser alvo de maior esforço de coleta.

Essa falta de integração entre os passos necessários para o estudo de um modelo tem suas raízes no desenvolvimento das teorias de EP e AIS. Enquanto a estimação de parâmetros sempre foi tratada sob o ponto de vista de uma teoria estatística, a análise de sensibilidade de modelos matriciais foi desenvolvida como uma ferramenta analítica, baseada em expansões lineares das funções de interesse em torno de um ponto privilegiado. Dependendo do tipo de experimento utilizado para obter estimativas das taxas vitais, diferentes abordagens podem ser utilizadas para encontrar um conjunto de valores que melhor represente o estado de conhecimento que temos de uma população, entre abordagens frequentistas, bayesianas ou baseadas em verossimilhança. Já a teoria analítica de AIS, devida em grande parte aos trabalhos de Hal Caswell (por exemplo [Caswell, 1989]), procede rotineiramente por tomar um modelo já parametrizado da forma ótima, e estudar as derivadas de primeira ordem da resposta de interesse em relação a cada um dos parâmetros de entrada. Nessa formulação, a quantidade e a qualidade dos dados de entrada para o modelo são ignorados, a menos de suas médias, e medidas de sensibilidade são tomadas exclusivamente sobre a variação da resposta de interesse a perturbações infinitesimais. Desta forma, modelos parametrizados com dados com grande incerteza não apresentarão maiores medidas de sensibilidade que modelos feitos sobre dados robustos, assim como aumentar o esforço amostral não reduzirá necessariamente as medidas de incerteza. A análise de sensibilidade analítica, em última análise, diz respeito à estrutura do modelo matricial, e não ao procedimento completo desde a tomada de dados em campo até a formulação e execução do modelo. Isso pode levar a uma falsa confiança em estudos que apresentem uma estrutura matricial estável, portanto de baixas sensibilidades, mas dados tomados com muita incerteza.

Por outro lado, questões de validação dos modelos são frequentemente menosprezadas pela literatura da área, apesar de estarem intimamente conectadas com as questões relevantes de incerteza. Podemos dividir a incerteza de um modelo em três fontes principais: incerteza estrutural, incerteza de parâmetros (ou epistêmica) e incerteza estocástica. Enquanto a maioria das técnicas de AIS se concentram na segunda componente, a validação de modelos pode apontar qual o nível de confiança que podemos ter de que um dado modelo é o modelo correto para reproduzir o fenômeno visto.

A formulação estocástica da teoria de análise de incerteza e sensibilidade global, como descrita no capítulo 2, e interpretada dentro de uma abordagem pautada pelo princípio da verossimilhança, é capaz de contemplar da mesma forma todas as

fontes de incerteza descritas acima, gerando como resultado um quadro completo do nosso conhecimento a respeito de um sistema. Embora para sistemas muito simples as abordagens possam convergir, isso não se verifica para problemas e sistemas de maior complexidade. As principais vantagens de usar a formulação estocástica são:

1. A abordagem analítica é local (portanto, responde ao que acontece com perturbações infinitesimais) e depende das funções serem "suaves"na vizinhança, a estocástica é global e não tem essa limitação.

2. Uma abordagem estocástica (baseada ou não na verossimilhança) permite que a informação contida na variabilidade amostral seja utilizada para representar a incerteza sobre os parâmetros (veja seção 2.2).

3. O princípio de verossimilhança afirma que toda a informação contida nas amostras coletadas está contida na função de verossimilhança. Dessa forma, a análise de sensibilidade feita a partir da função de verossimilhança contém toda a informação obtida pela amostra e nenhuma informação além da obtida pela amostra - enquanto a abordagem analítica e outras formulações estocásticas podem, alternadamente, desperdiçar informações coletadas ou levar à falsa impressão de gerar respostas com maior precisão do que a informação coletada permite.

3.2 Uma função de suporte

Ao falar da análise de incerteza de um modelo matemático sob o paradigma da verossimilhança, a pergunta que estamos fazendo pode ser escrita como: "qual o suporte que existe para hipóteses concorrentes a respeito do resultado de um modelo, a partir de um conjunto de dados coletados?". Em um modelo de crescimento populacional, por exemplo, a pergunta se torna "qual o suporte que os dados coletados fornecem para a hipótese de que a população está crescendo ou estável, *versus* a hipótese de que a população está em declínio?".

Vamos considerar problemas onde $\mathbf{x}$ representa um vetor de dados obtidos de forma independente em um ou mais experimentos a partir de uma variável aleatória $\mathbf{X}$, tal que $P(\mathbf{X}=\mathbf{x}) = f(\mathbf{x};\theta)$. O parâmetro (ou vetor de parâmetros) θ é desconhecido e pode assumir valores em Θ. A verossimilhança de θ dadas as observações $\mathbf{x}$ é dada por $\mathcal{L}(\theta|\mathbf{x}) = f(\mathbf{x};\theta)$. A razão de verossimilhanças sob o mesmo conjunto de dados pode ser abreviada como $L(\theta_1,\theta_2) = \frac{\mathcal{L}(\theta_1|\mathbf{X})}{\mathcal{L}(\theta_2|\mathbf{X})}$.

Sabemos que essa pergunta, se formulada sobre os parâmetros de uma distribuição de probabilidade, é respondida através da função de verossimilhança. Também sabemos que transformações injetoras sobre parâmetros preservam as propriedades da função de verossimilhança, isso é: se um parâmetro θ está associado a uma função de verossimilhança $\mathcal{L}(\theta|\mathbf{x})$ e $\phi = f(\theta)$ é dada por uma função $f(\cdot)$ injetora, a verossimilhança de ϕ é dada simplesmente por $\mathcal{L}(\phi|\mathbf{x}) = \mathcal{L}\left(f(\theta)|\mathbf{x}\right)$ ([Edwards, 1972], sec. 2.5).

Nosso trabalho, então, é o de estender esse resultado para uma função genérica. Nosso argumento se baseia em uma função de um argumento, $\gamma = g(\theta)$, sendo que

a generalização para mais dimensões é trivial.[1]

O problema de definição de uma verossimilhança para análise de incerteza em modelos está, portanto, intimamente ligada à questão de hipóteses estatísticas compostas. O trabalho fundamental de Neyman e Pearson [Neyman and Pearson, 1933] estabelece uma justificativa matemática para o princípio da verossimilhança em problemas envolvendo hipóteses simples. O problema de hipóteses compostas é resolvido dentro do paradigma frequentista para alguns casos particulares, como o teste t de Student, que avalia a hipótese nula de que a média de duas populações é igual, sendo a variância um parâmetro desconhecido. Estatísticos frequentistas tratam o problema de hipóteses compostas através do máximo de verossimilhança obtido por qualquer de suas hipóteses simples componentes. Muito do trabalho nessa área se concentrará em encontrar aproximações para a distribuição dessa estatística [Wilks, 1938].

Sob esta inspiração, podemos construir uma função tentativa de suporte $\Psi^{\delta}(\gamma|\mathbf{x})$, definida por:

$$\Psi^{\delta}(\gamma|\mathbf{x}) = \sup_{g(\theta)=\gamma} \mathcal{L}(\theta|\mathbf{x}) \tag{3.1}$$

Essa função pode ser intuitivamente interpretada no sentido de equiparar a verossimilhança da hipótese composta com a da sua melhor hipótese simples componente. Alternativamente, vamos construir uma outra função a partir de $\mathcal{L}(\theta|\mathbf{x})$: $\Psi^{\star}(\gamma|\mathbf{x})$, definida por:

$$\Psi^{\star}(\gamma|\mathbf{x}) = \int_{g(\theta)=\gamma} \mathcal{L}(\theta|\mathbf{x})\, d\theta \tag{3.2}$$

Intuitivamente, podemos considerar que se dois valores de θ levam a um mesmo valor de γ, a equação 3.2 nos diz que o suporte para esse valor de γ seria a soma do suporte dado aos valores de θ. Enquanto a função 3.1 tem uma forte inspiração nos trabalhos frequentistas, a função 3.2 vem de uma inspiração Bayesiana.

Em geral, ao determinar regras para combinar o suporte de hipóteses simples para construir uma função de suporte para hipóteses compostas, estamos considerando funções da forma:

$$\Psi(\gamma|\mathbf{x}) = \int_{g(\theta)=\gamma} \mathcal{L}(\theta|\mathbf{x})\kappa(\theta)\, d\theta \tag{3.3}$$

Onde a função $\Psi^{\star}$ é obtida trivialmente com $\kappa(\theta) = 1$, e a função Ψ^{δ} é um caso-limite no qual $\kappa(\theta)$ se aproxima de uma função Delta de Dirac. Essa classe de funções não corresponde necessariamente a funções de suporte, que devem possuir as seguintes propriedades desejáveis[Edwards, 1972]:

1. *Transitividade*: Se H_1 tem melhor suporte que H_2 e H_2 tem melhor suporte que H_3, então H_1 deve ter melhor suporte que H_3.

[1] A notação usada aqui é mais compatível com a literatura estatística, enquanto na seção 2.1.2, definimos as entradas, resultados e o modelo em si como x, y e f, relacionados como $y = f(x)$, aqui podemos pensar nesses objetos como θ, γ e g, respectivamente.

2. *Aditividade em relação aos dados*: o suporte relativo entre duas hipóteses depreendido de uma observação deve ser facilmente combinável com o suporte relativo para as mesmas hipóteses depreendido de uma observação diferente.[2]

3. *Invariância a transformações injetoras dos dados.*

4. *Invariância a transformações injetoras dos parâmetros.*

5. *Relevância e consistência*: se uma hipótese for verdadeira, ela deve receber mais suporte do que hipóteses concorrentes no longo termo. O suporte relativo deve ser coerente entre diversos problemas: o mesmo valor de suporte relativo deve ter o mesmo significado.

6. *Compatibilidade*: uma medida de suporte deve ser facilmente usada para atualizar informações na forma de *prioris*, nos casos nos quais elas existam.

Assim, é fácil ver que a função $\Psi^\star$, por exemplo, não é uma função de suporte - já que esta não é aditiva em relação ao parâmetro:

> "No special meaning attaches to any part of the area under a likelihood curve, or to the sum of the likelihoods of two or more hypotheses (...). Although the likelihood function, and hence the curve, has the mathematical form of a [known] distribution, it does not represent a statistical distribution in any sense." [Edwards, 1972]

Trabalhos futuros podem examinar as propriedades desta classe de funções; examinar se elas se adequam aos requerimentos de uma função de suporte - e caso contrário, se existe uma formulação alternativa desses requisitos que as contemple; e investigar a relação entre a adoção dessas funções e as bases lógicas da inferência, em especial o princípio e a lei da verossimilhança. No presente trabalho, vamos nos concentrar na complexa relação entre a lei da verossimilhança e o problema das hipóteses compostas.

3.3 *Caveat* sobre o uso de estatísticas

O uso de amostras de uma função de suporte para embasar um procedimento de análise de incerteza, como proposto nas seções anteriores, deve ser feito de forma a levar em consideração a natureza das grandezas envolvidas. Vamos fornecer aqui um exemplo de aplicação ingênua e equivocada desse procedimento, que busca usar a média da função como uma medida da tendência central dos resultados do modelo.

Considere o modelo simples[3]

$$y = x \tag{3.4}$$

[2] Edwards usa "aditividade" sobre a log-verossimilhança, equivalentemente medidas de verossimilhança podem ser combinadas de forma multiplicativa.

[3] Segundo a notação desenvolvida na seção 2.1

Onde o parâmetro x é estimado, a partir da realização de um processo Poisson, como tendo o valor $\hat{x}$. Para estudar a incerteza do resultado do modelo y, vamos utilizar a função de verossimilhança de x, normalizada de forma a representar uma probabilidade. Lembremos que a função de verossimilhança de uma distribuição Poisson é:

$$L\left(\lambda|\hat{x}\right) = C\lambda^{\hat{x}}\mathrm{e}^{-\lambda} \tag{3.5}$$

Onde C é uma constante multiplicativa que deve ser ajustada de forma que a integral de $L(\lambda|\hat{x})$ seja igual a 1:

$$\begin{aligned}
\int_0^\infty C\lambda^{\hat{x}}\mathrm{e}^{-\lambda} &= 1 \\
C &= \left(\int_0^\infty \lambda^{\hat{x}}\mathrm{e}^{-\lambda}\right)^{-1} \\
C &= \Gamma(\hat{x}+1)^{-1}
\end{aligned}$$

O máximo da função $L(\lambda|\hat{x})$, descrita acima, ocorre em $\lambda = \hat{x}$. Logo, o valor de $\hat{x}$ deve ser usado como estimador pontual do valor mais provável do parâmetro x. Da mesma forma, intervalos de confiança para o parâmetro x devem ser construídos ao redor de $\hat{x}$.

Após tomar amostras desta distribuição, construímos a distribuição de resultados:

$$D(y) = Cy^{\hat{x}}\mathrm{e}^{-y} \tag{3.6}$$

Cuja média é dada por

$$\begin{aligned}
< D(y) > &= \int_0^\infty yCy^{\hat{x}}\mathrm{e}^{-y}dy \\
&= \int_0^\infty Cy^{\hat{x}+1}\mathrm{e}^{-y}dy \\
&= \Gamma(\hat{x}+2)C \\
&= \frac{\Gamma(\hat{x}+2)}{\Gamma(\hat{x}+1)} \\
&= \hat{x}+1
\end{aligned}$$

Note então que, se escolhemos representar a distribuição dos resultados pela sua média, vamos estar usando $\hat{x}+1$, enquanto que o valor mais verossímil para y é $\hat{x}$, o mesmo que para o parâmetro x. Da mesma forma, ao construir intervalos de confiança para o resultado y, estes devem ser feitos considerando valores de maior verossimilhança, e não valores ao redor da média $\hat{x}+1$.

3.4 Hipóteses compostas

Como exposto na seção 1.14, a lei da verossimilhança (LL, do inglês Law of Likelihood), enunciada por Ian Hacking, se refere a hipóteses simples:

LL "If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence." [Hacking, 1965]

Como, então, trabalhar com hipóteses compostas? A resposta tradicional de verossimilhantistas para essa questão é uma simples negativa: Edwards descarta hipóteses compostas como "desinteressantes para a ciência" [Edwards, 1972], enquanto Royall enxerga na requisição de que as hipóteses sejam simples um benefício, e não uma falha, da abordagem por verossimilhança [Royall, 1997], exemplificado no problema a seguir:

Suponha que um grupo de pesquisadores da área médica realizam um experimento para determinar a probabilidade de sucesso de um certo tratamento. Eles estão particularmente interessados em descobrir se o novo tratamento tem mais probabilidade de sucesso do que 0.2, que representa a probabilidade de sucesso de um tratamento concorrente. Após aplicar o novo tratamento em 17 pacientes, eles encontram sucesso em 9. O que esse resultado pode dizer sobre as hipóteses?

Uma análise frequentista vai confrontar a hipótese nula $H_0 : \theta = 0.2$, e concluir que a probabilidade de encontrar 9 ou mais sucessos em 17 realizações de um processo Bernoulli com $p = 0.2$ é de aproximadamente 0.04%, rejeitando a hipótese.

Uma análise Bayesiana vai considerar as hipóteses $H_1 : \theta \leq 0.2$ versus $H_2 : \theta > 0.2$. O estatístico Bayesiano deve escolher uma forma para representar seu conhecimento prévio, e pode, por exemplo, usar a *priori* uniforme, dada por $B(1, 1)$, como feito por Bayes, encontrando uma *posteriori* igual a $B(10, 9)$, ou uma *priori* de Jeffreys, dada por $B(\frac{1}{2}, \frac{1}{2})$, encontrando uma *posteriori* de $B(9.5, 8.5)$. Ambas as análises levam a probabilidades muito baixas para H_1: 0.09% e 0.11%, respectivamente.

Ainda, a razão de verossimilhança entre o estimador de máxima verossimilhança para $\theta = 0.52$ e qualquer hipótese simples $H_p : \theta = p$, $p \leq 0.2$ é de no mínimo 91.5. Todos esses cálculos sugerem que os dados sejam coerentes com a hipótese de maior suporte ser aquela que considera que o novo tratamento tem probabilidade de sucesso maior do que 0.2.

No entanto, a lei da verossimilhança de Hacking não permite essa afirmação, pois H_1 e H_2 não são hipóteses que atribuam um único valor de probabilidade às observações da variável aleatória sob consideração. Royall aponta que, embora algumas hipóteses simples componentes de H_1 sejam melhor suportadas do que as componentes de H_2, isso não é válido em geral: a hipótese $\theta = 0.2$ é mais suportada do que $\theta = 0.9$ por um fator de 22.

É importante notar que isso não é uma consequência da formulação das hipóteses envolvendo desigualdades, de fato as hipóteses $H_1^\dagger : \gamma = 0$ e $H_2^\dagger : \gamma = 1$ são equivalentes, e tão intratáveis quanto, as hipóteses H_1 e H_2, com γ sendo dado por

$$\gamma = \begin{cases} 0, & \theta \leq 0.2 \\ 1, & \theta > 0.2 \end{cases} \tag{3.7}$$

Embora o uso de hipóteses compostas seja necessário para tratar hipóteses formuladas a respeito do resultado de modelos, pouco progresso se fez na elaboração

de uma lei da verossimilhança que se aplique a hipóteses compostas. Os caminhos que podem ser trilhados aqui são:

1. Tratar a questão através da modelagem de *nuisance parameters*;
2. Formular uma lei da verossimilhança que seja aplicável a hipóteses compostas;
3. Apresentar uma metodologia baseada em uma extensão lógica da lei da verossimilhança.

A proposta 1 se baseia em métodos como a verossimilhança perfilhada e a verossimilhança condicional, que são propostas *ad hoc* usadas para reduzir o estudo de modelos estatísticos multiparamétricos a um parâmetro por vez. Um uso típico é dado ao ajustar uma distribuição normal a uma série de dados: embora, estritamente, as hipóteses que possam ser comparadas sejam dadas por pares $(\mu = \mu_0, \sigma = \sigma_0)$ representando um valor fixo para a média e desvio padrão dessa normal, é possível comparar uma aproximação para suporte relativo para diferentes valores para a média, com o desvio padrão livre - portanto, caracterizando a hipótese composta $(\mu = \mu_0, \sigma \geq 0)$. Vamos retomar esta proposta na seção 3.7.

A proposta 2 se baseia na formulação axiomática de uma lei geral, GLL (do inglês Generalized Law of Likelihood), que seja compatível com a lei da verossimilhança LL para hipóteses simples, mas que extenda seu domínio para hipóteses compostas. Em contraste, a proposta 3 propõe utilizar uma definição mais fraca de evidência, baseada em uma lei fraca da verossimilhança (WLL, do inglês Weak Law of Likelihood), tal que aceitar LL implique em aceitar WLL, mas a conversa não seja verdadeira.

3.5 Uma lei geral de verossimilhança

Talvez a generalização mais óbvia para a lei da verossimilhança para hipóteses compostas seja tomar o máximo (ou supremo, no casos em que o máximo não existe) da verossimilhança de suas hipóteses simples. Esse é o caminho perseguido por Zhang [2009]; Zhang and Zhang [2013] e Bickel [2010], por exemplo. Zhang considera duas hipóteses, $H_1 : \theta \in \Theta_1 \subset \Theta$ versus $H_2 : \theta \in \Theta_2 \subset \Theta$, e postula os seguintes axiomas (levemente modificados para coerência com a notação):

Axioma 3.5.1 *Se* $\inf \mathcal{L}(\Theta_1|\mathbf{x}) > \sup \mathcal{L}(\Theta_2|\mathbf{x})$*, então a observação* $\mathbf{x}$ *é uma evidência a favor de* Θ_1*.*

Axioma 3.5.2 *Se* $\mathbf{x}$ *é evidência a favor de* H_1^* *em relação a* H_2 *e* H_1^* *implica em* H_1*, então* $\mathbf{x}$ *é evidência de* H_1 *sobre* H_2*.*

O primeiro axioma estabelece que se a imagem de $\mathcal{L}(\Theta_1|\mathbf{x})$ e $\mathcal{L}(\Theta_2|\mathbf{x})$ são intervalos disjuntos, portanto, se toda hipótese simples que implica em H_1 é suportada *versus* toda hipótese simples que implica em H_2, então H_1 é suportada *versus* H_2. Não parece haver motivo para rejeitar esse axioma, além de uma indisposição prévia a tratar hipóteses compostas. Já o segundo axioma é uma forma de estabelecer

coerência lógica. É importante clarificar que esse requisito de coerência *não é* equivalente a supor que a estrutura lógica das hipóteses deve ser usada como base para justificar um grau de crença sobre H_1^* ou H_1: se H_1^* implica em H_1 e a recíproca não é verdadeira, o axioma 3.5.2 *não* justifica que H_1^* seja melhor suportada que H_1 por qualquer evidência. Destes axiomas, deriva-se uma lei geral da verossimilhança:

Teorema 3.5.1 ***(GLL)*** *Se* $\sup \mathcal{L}(\Theta_1|\mathbf{x}) > \sup \mathcal{L}(\Theta_2|\mathbf{x})$*, então existe evidência a favor de* H_1 *sobre* H_2*.*

Demonstração Seja $\sup \mathcal{L}(\Theta_1|\mathbf{x}) > \sup \mathcal{L}(\Theta_2|\mathbf{x})$. Então existe $\theta_1 \in \Theta_1$ tal que $\mathcal{L}(\theta_1|\mathbf{x}) > \sup \mathcal{L}(\Theta_2|\mathbf{x})$. Do axioma 3.5.1, a hipótese $H_1^* : \theta = \theta_1$ é suportada em relação a H_2. Mas H_1^* implica H_1, e a conclusão segue do axioma 3.5.2.

Embora não decorra dos axiomas, o uso de uma razão de verossimilhança generalizada $\sup \mathcal{L}(\Theta_1|\mathbf{x})/\sup \mathcal{L}(\Theta_2|\mathbf{x})$ parece natural para quantificar a força da evidência. É trivial que GLL é compatível com LL no caso de hipóteses simples, e com alguns casos particulares expostos por [Royall, 1997]. Essa formulação da GLL pode ser usada para apresentar questões de análise de incerteza e para formalizar o uso de verossimilhanças perfilhadas em alguns casos de modelos com *nuisance parameters*.

No entanto, a implicação mais surpreendente da GLL é que ela permite uma medida de confirmação *absoluta* de hipóteses. Uma hipótese $H_1 : \theta \in \Theta_1$ possui evidência favorável se $\sup \mathcal{L}(\Theta_1|\mathbf{x}) > \sup \mathcal{L}(\Theta_1^c|\mathbf{x}$, onde ^{c} representa o conjunto complementar, ou $L(\Theta_1, \Theta_1^c) > 1$. Embora esse suporte seja nulo no caso de hipóteses simples sobre parâmetros contínuos, esse não é caso quando Θ é finito.

Considere um exemplo, dado por Royall [1997]:

Existem três urnas com bolas brancas em diferentes proporções: um quarto (θ_1), metade (θ_2) e três quartos (θ_3); identificamos como H_i as três urnas possíveis. Se nenhuma bola branca for observada após 5 sorteios com reposição, essa observação, $\mathbf{x}$, implica que $\mathcal{L}(\theta_1|\mathbf{x}) \propto (\frac{3}{4})^5$, $\mathcal{L}(\theta_2|\mathbf{x}) \propto (\frac{1}{2})^5$ e $\mathcal{L}(\theta_3|\mathbf{x}) \propto (\frac{1}{4})^5$.

A LL nos permite afirmar que $L(\theta_1, \theta_2) = 7.6$ é evidência mediana a favor de H_1 sobre H_2, e $L(\theta_2, \theta_3) = 32$ é evidência forte de H_2 sobre H_3. No entanto, a LL não permite inferências sobre a hipótese composta $H_c : \theta = \theta_1 \vee \theta = \theta_3$ em comparação com H_2; equivalentemente, LL não permite inferências sobre a evidência absoluta H_2 sobre $\sim H_2$.

Por outro lado, a GLL permite a afirmação de que $L(\theta_1 \vee \theta_3, \theta_2) = 7.6$ representa suporte para H_c em relação a H_2, e ainda que o suporte absoluto para as três hipóteses é de $L(\theta_1, \sim\theta_1) = 7.6$, $L(\theta_2, \sim\theta_2) = 0.13$, $L(\theta_3, \sim\theta_3) = 0.004$, confirmando que apenas a hipótese 1 tem força de evidência superior a 1.

Uma desvantagem em aceitar a GLL é que o suporte a hipóteses compostas não se comporta da mesma forma que a razão de verossimilhanças para hipóteses simples. Suponha que o pesquisador, no mesmo problema das urnas, retira agora duas bolas brancas. Essa nova observação leva aos resultados $L(\theta_1, \sim\theta_1) = 0.11$, $L(\theta_2, \sim\theta_2) = 0.44$, $L(\theta_3, \sim\theta_3) = 2.25$, oferecendo suporte apenas para H_3. No entanto, não há como combinar essa evidência com a apresentada acima: as quantidades derivadas do supremo da verossimilhança não são multiplicativas para conjuntos de dados distintos, ao contrário da verossimilhança. A força de evidência

absoluta para θ_1 tomando o conjunto completo de dados é de aproximadamente $1.8 \neq 7.6 * 0.11$. Essa constatação, no entanto, não invalida o teorema da GLL, embora lance dúvidas quanto à aplicabilidade da razão de verossimilhança generalizada como força de evidência para hipóteses compostas.

Em um desenvolvimento independente, o estatístico indiano Debabrata Basu propôs uma generalização alternativa para a lei da verossimilhança em 1975:

> "*The strong law of likelihood*: For any two subsets A and B of Θ, the data supports the hypothesis $\omega \in A$ better than the hypothesis $\omega in B$ if
>
> $$\sum_{\omega \in A} \mathcal{L}(\omega) > \sum_{\omega \in B} \mathcal{L}(\omega) \tag{3.8}$$
>
> Let us recall the [assertion] that all our sets (the sample space, the parameter space, etc.) are finite."

[Basu, 2011]

Novamente, o grande problema ao aceitar esta generalização é encontrar uma medida de força de evidência compatível.

3.6 Uma lei fraca de verossimilhança

Confrontados com os insucessos de trabalhar com a verossimilhança de hipóteses compostas e os problemas derivados de uma medida absoluta de confirmação, os principais defensores da lei da verossimilhança viram por bem abandonar a consideração de hipóteses compostas. O outro curso de ação que podemos tomar é abandonar a lei da verossimilhança, e é o caminho que vamos seguir nesta seção.

Um exemplo devido a Fitelson [2007] mostra uma questão importante sobre a lei da verossimilhança: tome um baralho de 52 cartas bem embaralhado, e considere as hipóteses (simples) H_1: a primeira carta é o ás de espadas *versus* H_2: a primeira carta é preta. A observação de uma carta de espadas leva às seguintes verossimilhanças: $\mathcal{L}(H_1|\spadesuit) = 1k > \mathcal{L}(H_2|\spadesuit) = \frac{1}{2}k$, sendo k uma constante de proporcionalidade, e a lei da verossimilhança nos leva a declarar o suporte para H_1 sobre H_2. Mas a evidência apresentada sobre H_1 é inconclusiva, mas garante H_2. Fitelson traça essa discordância na relação entre os chamados "*catch-alls*", $P(E| \sim H_1)$ e $P(E| \sim H_2)$[4]. Uma formulação mais geral sobre a lei da verossimilhança, apelidada por J. Joyce de lei fraca da verossimilhança (WLL) é dada por:

WLL "Evidence E favors hypothesis H_1 over hypothesis H_2 *if* $P(E|H_1) > P(E|H_2)$ *and* $P(E| \sim H_1) \leq P(E| \sim H_2)$." [Fitelson, 2007]

É evidente que LL $\implies$ WLL. No entanto, as principais correntes de Bayesianismo moderno *também* aceitam WLL: dada uma medida de confirmação $c(H, E)$, a expressão Bayesiana equivalente a LL é:

[4]Em casos de cartas de baralhos, os *catch-alls* estão bem definidos; objetivistas podem ter problemas em aceitar a formulação de *catch-alls* em problemas mais gerais, como "essa observação provê suporte à teoria da evolução"

$\dagger_c$ "Evidence E favors hypothesis H_1 over hypothesis H_2 according to measure c *iff* $c(H_1, E) > c(H_2, E)$." [Fitelson, 2007]

Três possíveis formas para a medida de confirmação são:

Diferença $d(H, E) = P(H|E) - P(H)$

Razão $r(H, E) = \frac{P(H|E)}{P(H)}$

Razão de verossimilhanças $l(H, E) = \frac{P(E|H)}{P(E|\sim H)}$

É importante relembrar que a confirmação Bayesiana dada por $c(H, E)$ tem um caráter não-relacional, à partir do qual se *deriva* uma medida de confirmação relacional; verossimilhantistas enxergam na lei da verossimilhança uma relação primitiva. E enquanto a construção da medida de confirmação não-relacional depende da especificação de *prioris*, a WLL não faz uso direto deles, mas apenas da especificação de *catch-alls*. Uma teoria que utilizasse a WLL sem invocar *prioris* poderia embasar uma escola de inferência intermediária, sem o uso de *prioris* ao qual objetivistas objetam, nem a restrição arbitrária ao tratamento de hipóteses compostas.

Concluímos esta seção ponderando que o paradigma da verossimilhança parece estar encravado dentro do pensamento Bayesiano: se generalizamos a Lei da Verossimilhança, nos encontramos em um paradigma Bayesiano; se a enfraquecemos, encontramos todas as correntes do Bayesianismo moderno. Se o problema da inferência estatística é respondido pela Lei da Verossimilhança, essa resposta deve passar por explicar essa singularidade.

3.7 PLUE: uma proposta de perfilhamento de verossimilhança

Nesta seção, vamos descrever uma metodologia tentativa para a realização de análises de incerteza baseada no perfilhamento da verossimilhança dos parâmetros. Argumentamos que essa metodologia é intuitivamente atraente dentro de um paradigma de verossimilhança. Vamos nos referir ao nosso procedimento como PLUE - Profiled Likelihood Uncertainty Estimation.

3.7.1 Intuição

Suponha que temos em mãos um modelo de crescimento populacional estruturado como aqueles discutidos nas seções 2.4 e 2.5, e um conjunto de dados a partir dos quais podemos calcular as taxas de sobrevivência, crescimento e fertilidade para a espécie.

De posse desses dados, gostaríamos de fazer as seguintes perguntas:

"Qual o suporte que os dados dão para a afirmação de que a população está estável versus *em declínio? Qual o suporte que os dados dão para a afirmação de que a população vai se extinguir em menos de 10 anos* versus *em mais de dez anos?"*

Essas perguntas não podem ser respondidas de um ponto de vista verossimilhantista, por corresponderem a hipóteses compostas sobre os parâmetros. No entanto, podemos tomar o ponto de máxima verossimilhança como um ponto privilegiado, e restringir nossa pergunta à forma:

"Qual o suporte que os dados dão para o ponto de máxima verossimilhança versus *qualquer ponto cuja taxa de crescimento populacional seja negativa?"*

Esquematicamente, podemos ver na figura 3.1 que estamos comparando a verossimilhança do ponto de máximo global (x_{max}) com o máximo obtido em uma região distinta (x_{lim}). Esta pergunta compara uma hipótese simples, a que corresponde à máxima verossimilhança, com uma hipótese composta. Esse tipo de comparação evita os problemas e peculiaridades que surgem em propostas de leis gerais de verossimilhança devido à sobreposição entre a verossimilhança de hipóteses diferentes. Em termos dos axiomas de Zhang[5], estamos aceitando uma forma fraca do primeiro, mas não o segundo.

Esse raciocínio pode ser expandido para perguntarmos:

"Quais são os pontos do espaço de parâmetros para os quais o suporte é maior do que uma certa distância δ em relação ao ponto de máxima verossimilhança?"

Ao fazer essa pergunta para uma série de valores de δ, estamos efetivamente perfilhando a verossimilhança de cada ponto do espaço de parâmetros. Neste ponto, precisamos apontar que procedimentos de perfilhamento vêm sendo utilizados na inferência por verossimilhança há décadas para reduzir a dimensionalidade do espaço de parâmetros - seja na forma de uma simples eliminação de "*nuisance parameters*", seja de forma elaborada através de análises de componentes principais. Considere, por exemplo, o caso típico em que desejamos ajustar uma distribuição normal a uma série de dados, e ao invés de compararmos pares da forma $(\mu = \mu_0, \sigma = \sigma_0)$ representando um valor fixo para a média e desvio padrão dessa normal, desejamos fazer afirmações apenas sobre a média μ desta normal. Já apontamos que este raciocínio corresponde a encarar hipóteses compostas da forma $(\mu = \mu_0, \sigma \geq 0)$. Agora, vamos notar que isto é equivalente a considerar a função $g(\mu, \sigma) = \mu$ - ou seja, estamos projetando o espaço bidimensional formado por μ e σ em um espaço unidimensional. Nossa proposta pode ser considerada uma generalização deste procedimento: ao invés de considerar apenas funções que removam um dos parâmetros, estamos considerando qualquer função não-inversível.

Há que se considerar, no entanto, que técnicas de remoção de "*nuisance parameters*" dentro do paradigma de verossimilhança no caso geral se assentam sobre questões que não estão totalmente resolvidas - e que se situam na fronteira entre o pensamento verossimilhantista e a inferência fiducial [Edwards, 1972; Kalbfleisch and Sprott, 1970]. Se estes métodos não podem ser, em geral, encarados como forma de inferência tão rigorosa quanto a inferência pela função de verossimilhança, é bem aceito que esta é uma análise exploratória válida. Nossa generalização se encontra no mesmo estado: embora ela não possa ser vista como uma forma rigorosa de inferência, ela pode ser usada de forma exploratória sem incorrer em problemas mais profundos que a perfilhagem tradicional.

Lembramos que, se do ponto de vista da análise de verossimilhança, nossa proposta não apresenta grande inovação, na literatura de análise de sensibilidade ela é

[5] vide seção 3.5

única em propor uma metodologia que utilize toda a informação contida nas amostras coletadas.

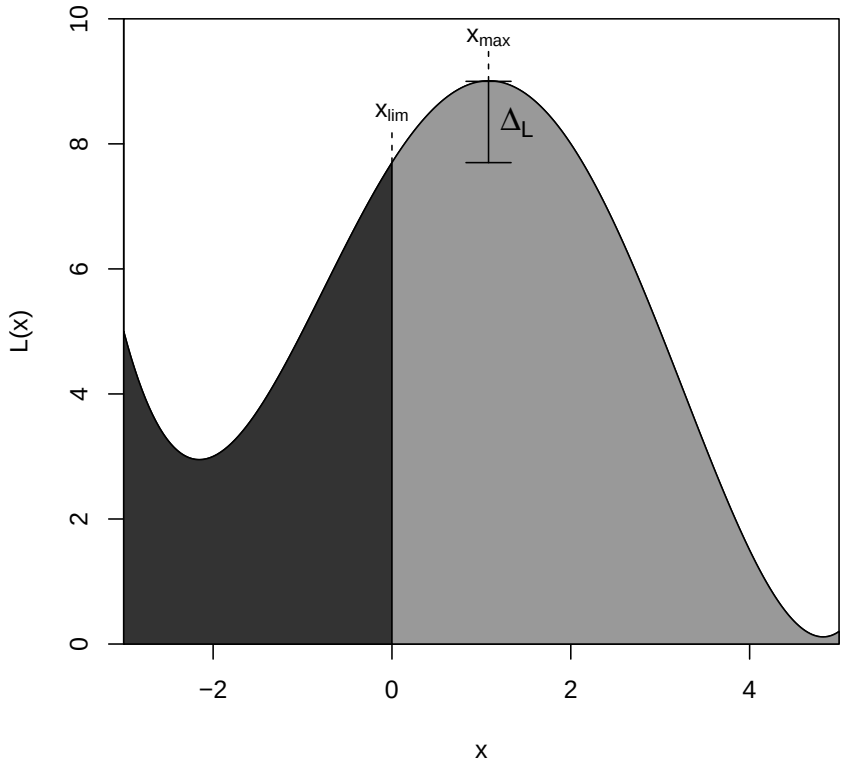


Figura 3.1: Representação esquemática de uma função de verossimilhança, mostrando dois pontos de interesse: o máximo global x_{max} e o máximo de uma região distinta x_{lim}. Veja detalhes no texto.

3.7.2 Método

Definições Considere um modelo de interesse biológico qualquer - vamos chamá-lo de modelo biológico, para diferenciar do modelo estatístico que será apresentado abaixo. Vamos considerar o caso simples de um resultado escalar y de um modelo com um vetor de entradas θ: $y = F(\theta)$. Vamos representar por χ o conjunto de dados coletados. Se temos n parâmetros e m observações, χ é uma tabela de n colunas por m linhas.

Formule diferentes modelos para explicar seus dados (ex, fertilidade constante ou agrupada, taxa de crescimento constante ou decrescente com a classe de tamanho, modelo com 4 ou com 5 classes de tamanho, etc) - este vai ser designado o modelo estatístico. Escreva a função de verossimilhança $\mathcal{L}(\theta|\chi)$ para cada modelo. Encontre o conjunto de parâmetros que melhor ajusta

seus dados para cada modelo e determine o valor de AIC para cada modelo estatístico.

Amostragem De posse do modelo estatístico de melhor AIC[6], utilize um método de Monte Carlo para gerar um grande número de amostras com densidade proporcional a $\mathcal{L}(\theta|\chi)$. Vamos chamar esta amostra discreta de $\mathbf{A}$. À cada amostra $A_{i\cdot}$ da função $\mathcal{L}(\theta|\chi)$, associamos $L_i = \mathcal{L}(A_{i\cdot}|\chi)$, o valor de verossimilhança desta amostra, e $Y_i = F(A_{i\cdot})$, o resultado do modelo biológico quando executado sobre esta amostra. Normalize L_i de forma que o mínimo de verossimilhança seja no 0.

Agregação A partir dos valores resultantes do modelo Y_i e seus valores de verossimilhança L_i associados, construa o perfil superior para a verossimilhança de y da seguinte forma: para cada incremento z, encontre o maior valor $\bar{y}$ em Y_i tal que $L_i \leq z$. Anote este valor como $P_{sup}(z) = \bar{y}$ e repita para um valor maior de z.

Proceda de forma análoga para construir o perfil inferior de verossimilhança. Os dois perfis, em conjunto, podem ser utilizados para investigar as regiões de plausibilidade para y.

3.8 Estudo de caso 3: um modelo mínimo

Um modelo de população estruturada que pode ser considerado mínimo é o modelo com juvenis não-reprodutivos e adultos reprodutivos, modelo discutido por Caswell [2008] como um exemplo simples da aplicação da análise de sensibilidade analítica. O modelo é associado à seguinte matriz:

$$A = \begin{bmatrix} \sigma_1(1-\gamma) & f \\ \sigma_1\gamma & \sigma_2 \end{bmatrix} \tag{3.9}$$

Aqui, σ_1 é a probabilidade de sobrevivência de juvenis, σ_2 é a probabilidade de sobrevivência de adultos, γ é a probabilidade de maturação, e f é a fertilidade dos adultos. Vamos representar por λ o maior autovalor dessa matriz.

Vamos primeiramente presumir que a sobrevivência independe do estágio ($\sigma_1 = \sigma_2 = \sigma$). Também vamos fazer a suposição de que é possível marcar inequivocamente quais dos juvenis nasceram no último ciclo, e quais adultos passaram pelo processo de maturação no último ciclo, para chegar ao modelo:

$$A = \begin{bmatrix} \sigma(1-\gamma) & f \\ \sigma\gamma & \sigma \end{bmatrix} \tag{3.10}$$

Para animais de tamanho grande, com um filhote por estação reprodutiva, f pode ser aproximado pela proporção de adultos que gera prole, σ é dado pela proporção de indivíduos que sobrevivem por um ciclo e γ pela proporção de juvenis

[6]É possível que uma abordagem de inferência multi-modelo possa ser utilizada em casos de empate de AIC [Burnham and Anderson, 2002]

que se tornam adultos por ciclo, de forma que os três parâmetros podem ser modelados por distribuições binomiais, com probabilidades θ_i desconhecidas e número de tentativas dados, respectivamente, por n_1, o número original de juvenis, n_2, o número original de adultos, e n_t, o tamanho total da população:

$$\gamma \sim \text{binom}(\theta_1, n_1) \tag{3.11}$$
$$f \sim \text{binom}(\theta_2, n_2) \tag{3.12}$$
$$\sigma \sim \text{binom}(\theta_3, n_t = n_1 + n_2) \tag{3.13}$$

Vamos utilizar ainda a suposição de que os parâmetros são independentes neste exemplo, para chegar às funções de verossimilhança retratadas na fig. 3.2 (detalhes sobre a construção dessa função podem ser vistos na seção 3.8.1).

Vamos examinar um exemplo numérico com a população inicial contendo 10 juvenis e 15 adultos. O tamanho populacional pequeno é importante para acentuar as diferenças entre as abordagens. Após um ciclo, observamos 3 adultos recém maduros, 2 nascidos e 23 sobreviventes. É fácil ver na figura 3.2 que a melhor estimativa para cada parâmetro é dada por $\sigma = 0.92$, $f = 0.13$ e $\gamma = 0.3$. Neste caso, o valor de λ é 1.02.

As funções de verossimilhança de cada parâmetro foram utilizadas para gerar 10000 amostras pelo método de Metropolis, a partir das quais geramos uma distribuição empírica para λ, de forma proporcional à verossimilhança dos parâmetros. Esta distribuição de valores de λ, conjuntamente com os valores de verossimilhança associados a cada input, foi usada para gerar um perfil de verossimilhança para o resultado do modelo. O mínimo de verossimilhança para λ é atingido em $\lambda = 1.02$.

As figuras 3.4 e 3.5 mostram resultados preliminares da aplicação de técnicas de análise de sensibilidade sobre as amostras geradas, análogas às discutidas no capítulo 2. É importante ressaltar que estas análises foram realizadas sobre uma vizinhança não infinitesimal (como seria o caso analítico) nem arbitrária (como seriam as análises descritas no cap. 2), centrada no ponto de máxima verossimilhança.

A análise realizada indica que o valor de λ estimado é pouco confiável, tendo um perfil muito aberto. É importante notar que esses resultados são para uma única amostra, e um tamanho amostral decididamente pequeno. Considerando uma amostra 3 vezes maior, na qual todas as proporções se mantenham as mesmas (ou seja, o número de indivíduos maturados, nascidos e sobreviventes é multiplicado por 3), a análise resulta em um perfil muito mais fechado (veja figura 3.6).

3.8.1 Detalhes matemáticos

Nesta seção, vamos desenvolver alguns detalhes matemáticos sobre o exemplo acima.

Para um dado número observado x_A de juvenis que passaram pelo processo de maturação, se tornando adultos, após um ciclo, a função de log-verossimilhança para γ é dada por

$$\mathcal{L}(\theta_1|x_A) = \log\left(\binom{n_1}{x_A}\theta_1^{x_A}(1-\theta_1)^{n_1-x_A}\right) \tag{3.14}$$

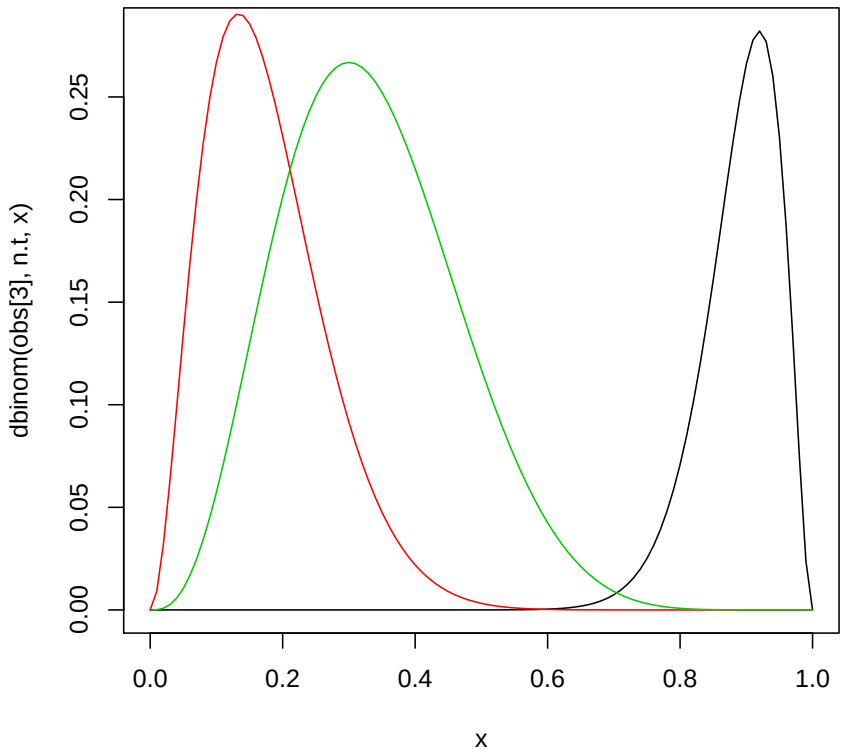


Figura 3.2: Função de verossimilhança para cada parâmetro do modelo. Preto = σ, vermelho = f e verde = γ.

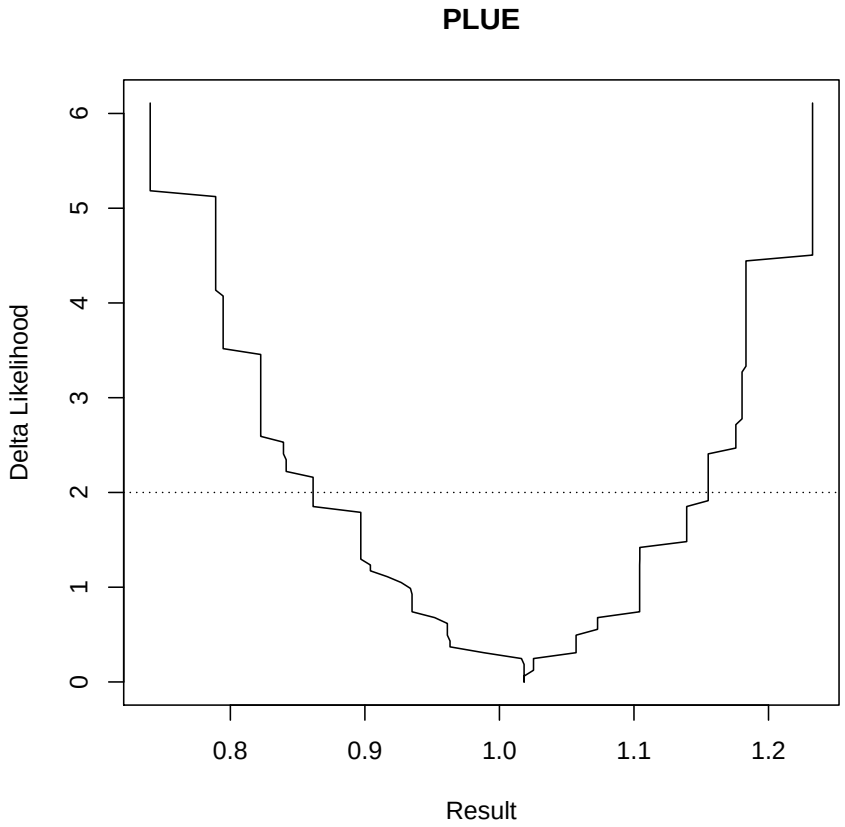


Figura 3.3: Análise de verossimilhança perfilhada sobre os resultados do modelo mínimo de população estruturada.

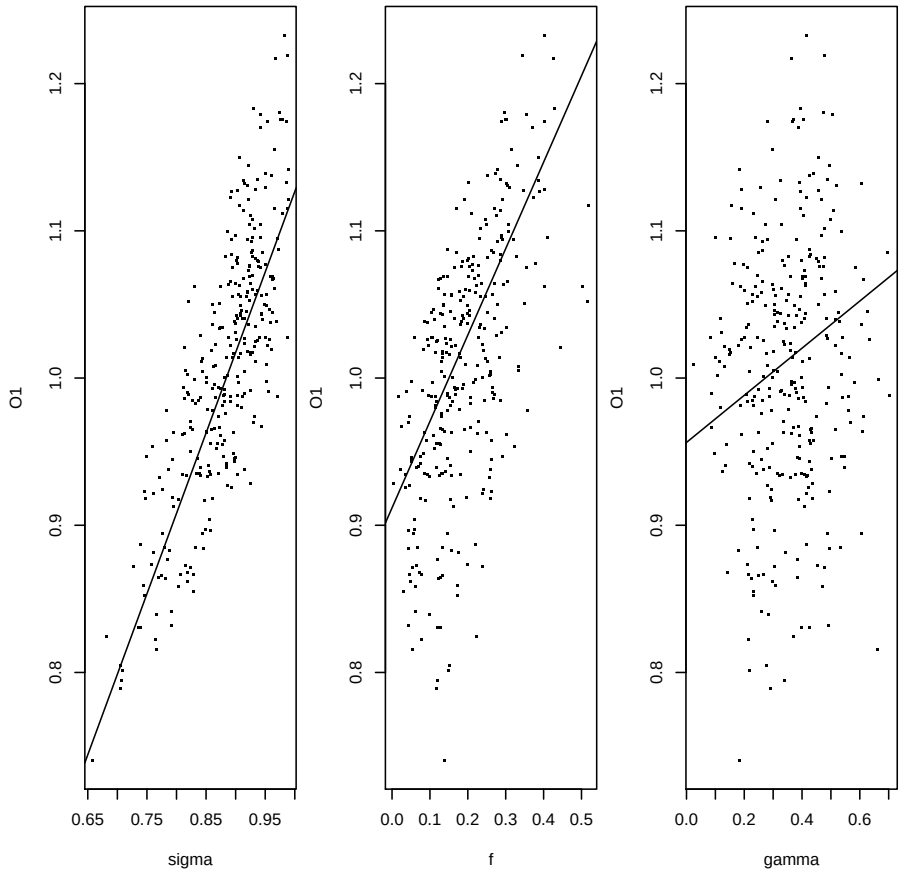


Figura 3.4: Gráfico de dispersão dos valores de parâmetros (no eixo x) e resultados do modelo de crescimento estruturado mínimo, gerados a partir de uma abordagem de verossimilhança.

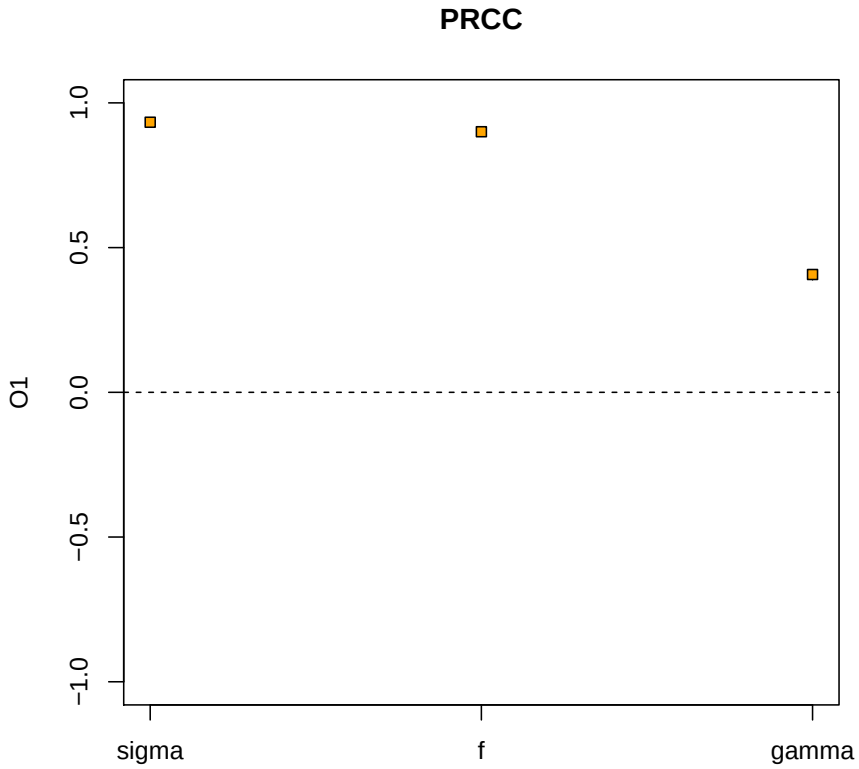


Figura 3.5: Análise de Partial Rank Correlation Coefficient entre as entradas do modelo e o resultado em um modelo estruturado mínimo de crescimento populacional

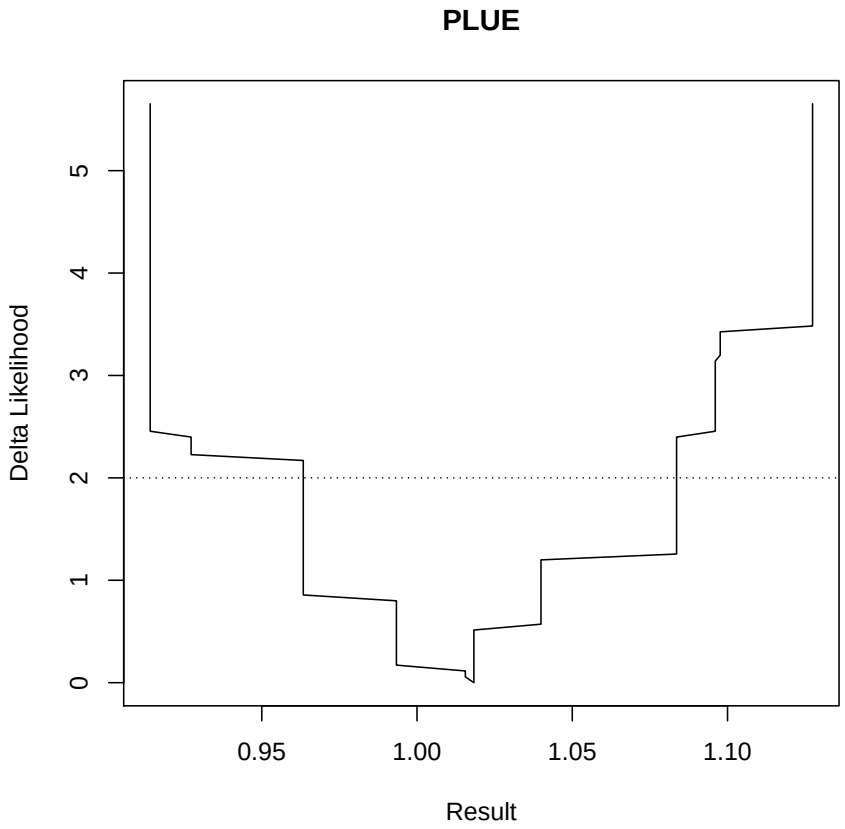


Figura 3.6: Análise de verossimilhança perfilhada sobre os resultados do modelo mínimo de população estruturada, mas com tamanho amostral maior. Em comparação com a figura 3.3, o perfil é muito mais fechado.

Da mesma forma, a função de log-verossimilhança para f é dada em função do número de juvenis nascidos no último ciclo, x_J:

$$\mathcal{L}\left(\theta_2|x_J\right) = \log\left(\binom{n_2}{x_J}\theta_2^{x_J}(1-\theta_2)^{n_2-x_J}\right) \tag{3.15}$$

Por fim, a função de log-verossimilhança referente a σ é dada em função do número de indivíduos sobreviventes, x_S, calculado como o número de indivíduos observado menos x_J:

$$\mathcal{L}\left(\theta_3|x_S\right) = \log\left(\binom{n_t}{x_S}\theta_3^{x_S}(1-\theta_3)^{n_t-x_S}\right) \tag{3.16}$$

Com o pressuposto forte de que as três variáveis são independentes, e escrevendo $n_t = n_3$ e $\{x_A, x_J, x_S\} = \{x_1, x_2, x_3\}$ para facilitar a notação[7] a função de verossimilhança para o vetor de parâmetros $\boldsymbol{\theta}$ é:

$$\mathcal{L}\left(\boldsymbol{\theta}|\mathbf{x}\right) = \sum_i \log\left(\binom{n_i}{x_i}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i}\right) \tag{3.17}$$

O resultado do modelo é λ, o maior autovalor de A obedecendo

$$\det\begin{bmatrix} \lambda-\sigma(1-\gamma) & -f \\ -\sigma\gamma & \lambda-d \end{bmatrix} = \lambda^2 - \lambda\,\mathrm{tr}(A) + \det(A) = 0 \tag{3.18}$$

$$\lambda = \frac{1}{2}\left(\mathrm{tr}(A) + \sqrt{\mathrm{tr}^2(A) - 4\det(A)}\right) \tag{3.19}$$

[7]Importante frisar: x_1 não corresponde aqui ao número de juvenis observados após um ciclo, etc.

Appendix A

Sensitivity analyses: a brief tutorial with R package pse

Chalom, A. [1] [2], Mandai, C.Y. [1] and Prado, P.I. [1]

This document presents a brief practical tutorial about the use of sensitivity analyses tools in the study of ecological models. To read about the underlying theory, please refer to our work in [Chalom and Prado, 2012].

We presume for this tutorial that you are already familiar with the R programing environment and with your chosen model. We will illustrate the techniques using a simple model of population growth.

You should have installed **R**[3] along with an interface and text editor of your liking, and the package "pse" (available on http://cran.r-project.org/web/packages/pse). This package is based on the "sensitivity" package, and is designed to resemble its uses, so researchers who already use it will be able to write code with the pse package easily. Major differences will be noted on the help pages and in this tutorial.

This tutorial focuses on the parameter space exploration of deterministic models. For a discussion of stochastic models, see the 'multiple' vignette on the same package.

A.1 Input parameters

The first thing that must be done is to determine exactly what are the input parameters to your model. You should list which parameters should be investigated, what are the probability density functions (PDFs) from which the parameter values will be calculated, and what are the arguments to these PDFs.

In the examples, we will use a simple model of logistical population growth, in which a population has an intrinsic rate of growth r, a carrying capacity of K and a starting population of X_0. In each time step, the population may grow or diminish after the following expression:

[1]Theoretical Ecology Lab, LAGE at Dep. Ecologia, Instituto de Biociências, Universidade de São Paulo, Rua do Matão travessa 14 nº 321, São Paulo, SP, CEP 05508-900, Brazil.

[2]email: andrechalom@gmail.com

[3]This tutorial was written and tested with **R** version 3.0.1, but it should work with newer versions

$$X_{t+1} = X_t + rX_t\,(1 - X_t/K) \tag{A.1}$$

We are interested in studying the effects of the parameters r, K and X_0 on the final population. After researching on our databases, we have decided that, for our species of interest, r and K follow a normal distribution with known parameters. However, we could not reliably determine what the initial population should be, so we have used an uniform distribution covering all the reasonable values. The following table summarizes this:

Parameter	Distribution	Arguments
r	normal	$\mu = 1.7$, $\sigma = 0.3$
K	normal	$\mu = 40$, $\sigma = 1$
X_0	uniform	min = 1, max = 50

We next translate this table to three **R** objects that will be used in the sensitivity analyses, containing (1) the names of the parameters, (2) the probability density functions, and (3) *a list containing the lists* with all the parameters to the density functions:

```
> factors <- c("r", "K", "X0")
> q <- c("qnorm", "qnorm", "qunif")
> q.arg <- list( list(mean=1.7, sd=0.3), list(mean=40, sd=1),
+         list(min=1, max=50) )
```

A fundamental question in this stage is to determine whether, inside the ascribed parameter ranges, *every parameter combination* is meaningful. See the next examples on this:

Example 1:

We would like to run a model for a species abundance distribution (SAD), and we decided to examine the effect of N, the total number of individuals in the community, and S, the total number of species. We can run the model with $N = 100$ and $S = 50$ or with $N = 15$ and $S = 3$, so there is nothing wrong with these values. However, the *combination* $N = 15$, $S = 50$ is meaningless, as it would imply that there are more species than individuals. One solution to this problem is to run the models with the parameters modified as following: N is the total number of individuals, and $\hat{s}$ is the average number of individuals for each species. So, $\hat{s} * N = S$, and now every combination of N and $\hat{s}$ is meaningful.

Example 2:

In a model of structured population growth, we have estimated independently two parameters for each class: S, the probability that a given individual survives and does not move into the next size class, and G, the probability that a given individual survives and grows into the next class. We can run the model with $S = 0.2$ and $G = 0.7$, or $S = 0.8$ and $G = 0.1$. However, if we try to run the model with $S = 0.8$ and $G = 0.7$, we arrive at the conclusion that, for every individual in the original size class, in the next time step we will have 0.8 individuals in the same

class and more 0.7 in the next, giving a total of 1.5 individuals! The problem is that the sum of S and G must be smaller than 1. One way to solve this is to define new parameters $\hat{s}$ and $\hat{g}$ such that $\hat{s}$ is the survival probability, independently of the individual growing, and $\hat{g}$ is the growth probability for each surviving individual. We can relate these parameters to the former ones, as $G = \hat{s} * \hat{g}$ and $S = \hat{s} * (1 - \hat{g})$.

Note:

When transforming parameters like done on the above examples, it is important to remember that the new parameters may not have the same probability density functions as the original ones.

A.1.1 Optional: More details about the quantiles

The quantile functions used can be any of the built-in quantile functions as **qnorm** for normal, **qbinom** for binomial, **qpois** for poison, **qunif** for uniform, etc; less common distributions can be found on other packages, like the truncated normal distribution on package "msm". You can even define other quantile functions, given that their first argument is the probability, and that they are able to work on a vector of probabilities. For example:

The quantiles of an empirical data set can be used by creating a wrapper function for the **quantile** function:

```
> qdata <- function(p, data) quantile(x=data, probs=p)
```

A discrete uniform density function, useful for parameters that must be integer numbers, can be given by

```
> qdunif<-function(p, min, max) floor(qunif(p, min, max))
```

A.2 Your model

The model that you wish to analyze must be formulated as an **R** function that receives a *data.frame*, in which every column represent a different parameter, and every line represents a different combination of values for those parameters. The function must return an array with the same number of elements as there were lines in the original data frame, and each entry in the array should correspond to the result of running the model with the corresponding parameter combination. We will cover the case in which a model outputs more than a single number in section A.4.

If your model is already written in R, and accepts a single combination of values, it is easy to write a "wrapper" using the function **mapply** to your model. In the example below, the function **oneRun** receives three numbers, corresponding to r, K and X_0, and returns a single value corresponding to the final population. The function **modelRun** encapsulates this function, in a manner to receive a data.frame containing all parameter combinations and returning the results in one array.

Make **SURE** that the order in which the parameters are defined above is the same in which they are being passed to the function.

```
> oneRun <- function (r, K, Xo) {
+     X <- Xo
+     for (i in 0:20) {
+        X <- X+r*X*(1-X/K)
+     }
+     return (X)
+ }
> modelRun <- function (my.data) {
+         return(mapply(oneRun, my.data[,1], my.data[,2], my.data[,3]))
+ }
```

If your model is written in a different language, as C or Fortran, it is possible to write an interface with **R** by compiling your model as a shared library, and dynamically loading this library [Geyer, 2012]. Also, you should consider uncoupling the simulation and the analyses (see section A.5).

A.3 Uncertainty and sensibility analyses

We first use the **LHS** function to generate a hypercube for your model. The mandatory arguments for this function are: *model*, the function that represents your model; *factors*, an array with the parameter names; *N*, the number of parameter combinations to be generated; *q*, the names of the PDF functions to generate the parameter values; and *q.arg*, a list with the arguments of each pdf. We have already constructed suitable objects to pass to this function above, so now we simply call the **LHS** function:

```
> library(pse)
> myLHS <- LHS(modelRun, factors, 200, q, q.arg, nboot=50)
```

The extra parameter *nboot* is used to bootstrap the correlation coefficients (see below).

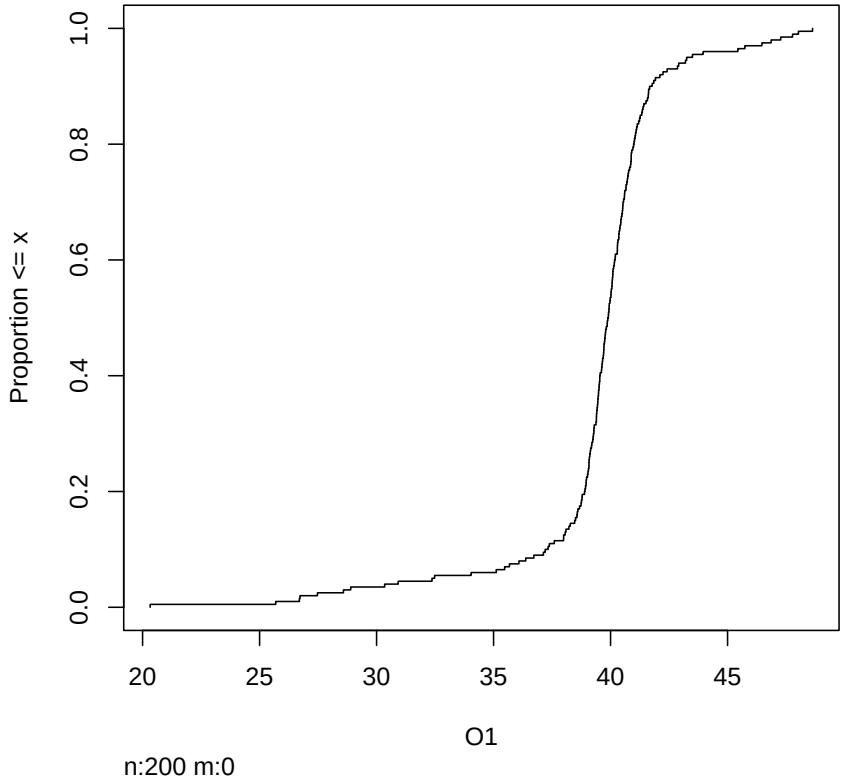
To access the values of the parameters used in the model, use the function **get.data(myLHS)**. To access the results, use **get.results(myLHS)**.

With the object returned by the function **LHS**, we will exemplify in this section four techniques that can be used: the uncertainty analysis using the **ecdf**, scatterplots of the correlation between each parameter and the result using the function **plotscatter**, partial rank correlation using the function **plotprcc** and agreement between different hypercube sizes with the function **sbma**.

A.3.1 ECDF

The ecdf, short for empirical cumulative distribution function, may be used to illustrate the distribution of the model results, in our case the final population. With this graph, we can see that the final population for our species is, with high probability, between 35 and 45.

```
> plotecdf(myLHS)
```

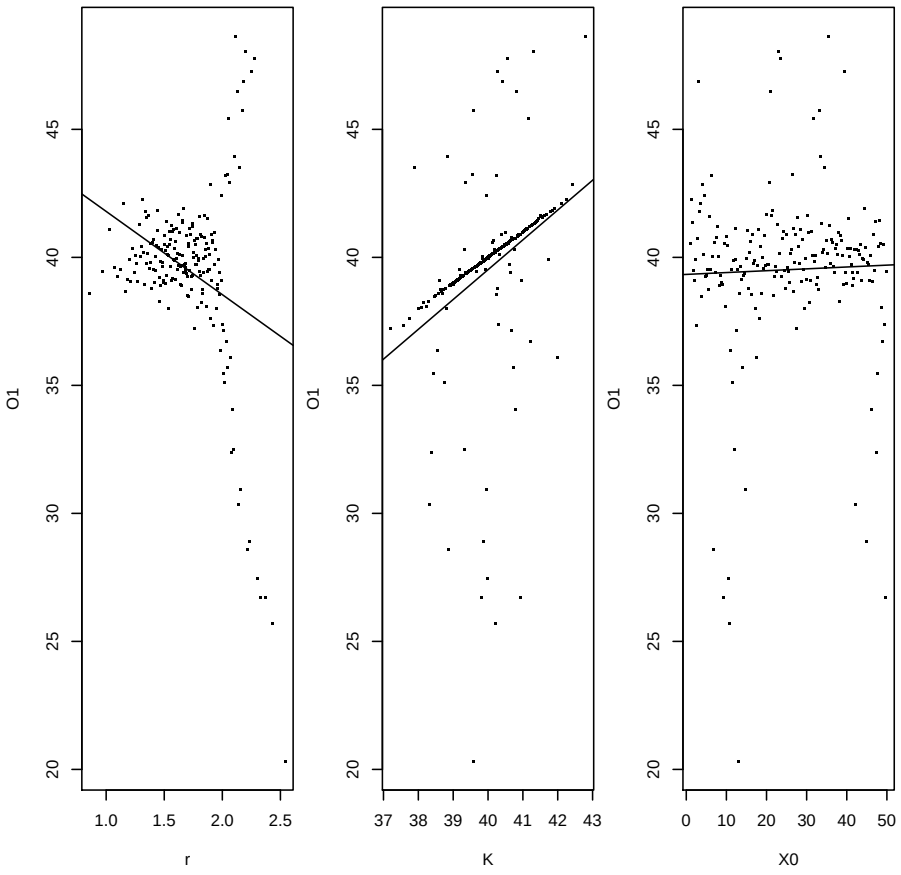


A.3.2 Scatterplots

Here, we can see a scatterplot of the result as a function of each parameter. As all the parameters are being changed for each run of the model, the scatterplots look like they were randomly generated, even if the underlying model is deterministic. Actually, what scatterplots show is the distribution of values returned by the model in the parameter space sampled by the hypercube and how sensible are these model responses to the variation of each parameter.

Note that population sizes bifurcate above a given value of parameter r. This is a well known behavior of many population models and will ultimately lead to chaotic solutions [May, 1976; Murray, 2002].

```
> plotscatter(myLHS)
```

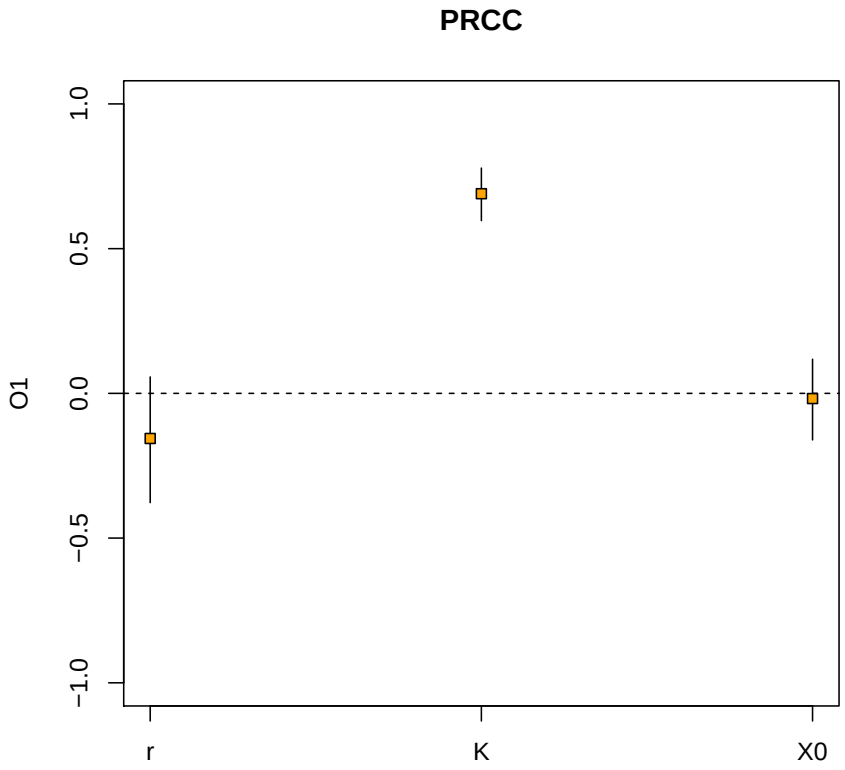


A.3.3 Partial correlation

The partial (rank) correlation coefficient (pcc or prcc) measures how strong are the linear associations between the result and each input parameter, after removing the linear effect of the other parameters.

The confidence intervals shown in this plot are generated by bootstrapping.

```
> plotprcc(myLHS)
```



In the ecological literature, it is usual to refer to the partial derivatives of the model response in respect to each parameter as "the sensitivity" of the model response in respect to each parameter. One analog measure in stochastic models is the Partial Inclination Coefficient (pic) of the model response in respect to each parameter.

```
> pic(myLHS, nboot=40)

[[1]]

Call:
pic.default(X = L, y = r, nboot = nboot, conf = conf)

Partial Inclination Coefficients (PIC):
       original         bias std. error  min. c.i.   max. c.i.
```

```
r  -3.250037198  0.061491430 1.36721722 -5.9616927 -1.16739143
K   1.164546035  0.039996005 0.21752022  0.6613275  1.54543939
X0  0.007266329 -0.000577073 0.01577244 -0.0247911  0.03757147
```

A.3.4 Agreement between runs

In order to decide whether our sample size was adequate or insufficient, we calculate the Symmetric Blest Measure of Agreement (SBMA) between the PRCC coefficients of two runs with different sample sizes.

```
> newLHS <- LHS(modelRun, factors, 250, q, q.arg)
> (mySbma <- sbma(myLHS, newLHS))

[1] 1
```

A value of -1 indicates complete disagreement between the runs, and a value of 1 indicates total agreement. As the SBMA seldom reaches 1 for realistic models, some criterion must be used to indicate when the agreement should be considered good enough. More details about how the SBMA is calculated can be found on [Chalom and Prado, 2012].

It should be stressed that there is no "magical" number for deciding how close to unity the SBMA should be. It is reasonable to expect agreements around 0.7 to 0.9 in well-behaved models, but two cases require attention. If the total number of factors is very low, the SBMA may converge slowly. Also, if none of the model parameters happen to be monotonically correlated with the output, the agreement between runs may stay as low as 0.2 even for very large hypercubes.

A.4 Multiple response variables

In the previous section, we have examined a model that returned a single number, namely, the final population. However, we might be interested in examining the effects of the parameters in several distinct responses from the model. The responses may be (1) different variables, like "total population" and "species richness", (2) the same variable in different time points, or (3) the same variable calculated by different methods.

In our example, we are interested in determining the effect of the parameters to the population in each of the first 6 time steps. The theory and tools for this analysis remain mostly the same. We will write our model to return an array now, as:

```
> factors <- c("r", "K", "X0")
> q <- c("qnorm", "qnorm", "qunif")
> q.arg <- list( list(mean=1.7, sd=0.3), list(mean=40, sd=1),
+         list(min=1, max=50) )
> Time <- 6
> oneRun <- function (r, K, Xo) {
+         X <- array();
+         X[1] <- Xo; # Caution, X1 gets overwritten
+         for (i in 1:Time) {
+                 Xl <- X[length(X)]
+                 X[i] <- Xl + r*Xl*(1-Xl/K)
+         }
+         return (X)
+ }
> modelRun <- function (dados) {
+         mapply(oneRun, dados[,1], dados[,2], dados[,3])
+ }
```

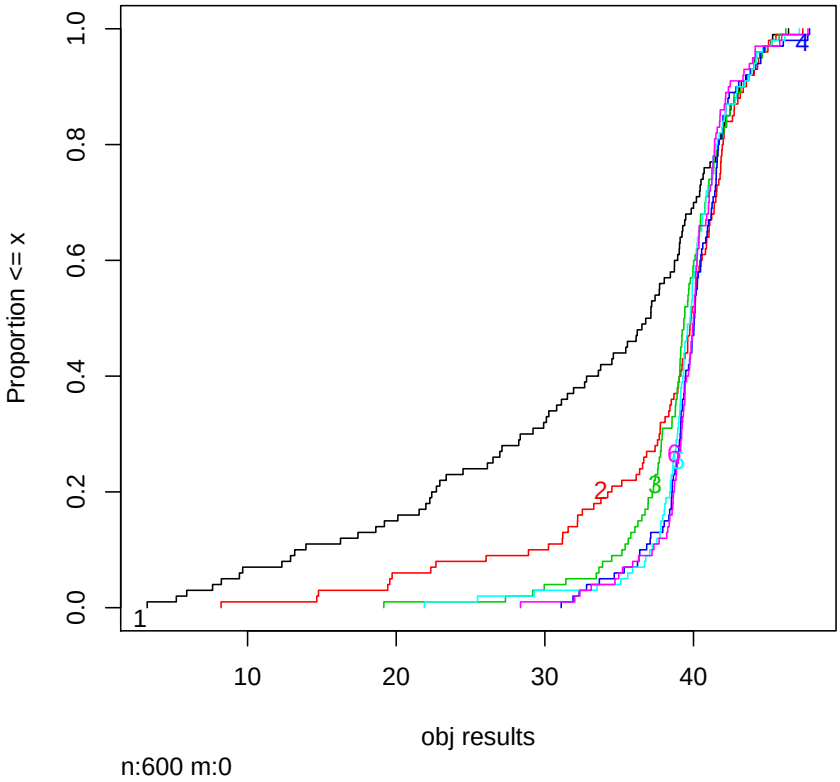
The hypercube is generated exactly in the same way. However, now we have the option to give names (which will be used in the plots below) to each response variable.

```
> res.names <- paste("Time",1:Time)
> myLHS <- LHS(modelRun, factors, 100, q, q.arg, res.names, nboot=50)
```

A.4.1 ECDF

The first plot we will produce will, again, be the ECDF. We may produce several plots using the parameter "stack=FALSE", or stack all the plots in the same graph, using "stack=TRUE":

```
> plotecdf(myLHS, stack=TRUE)
```

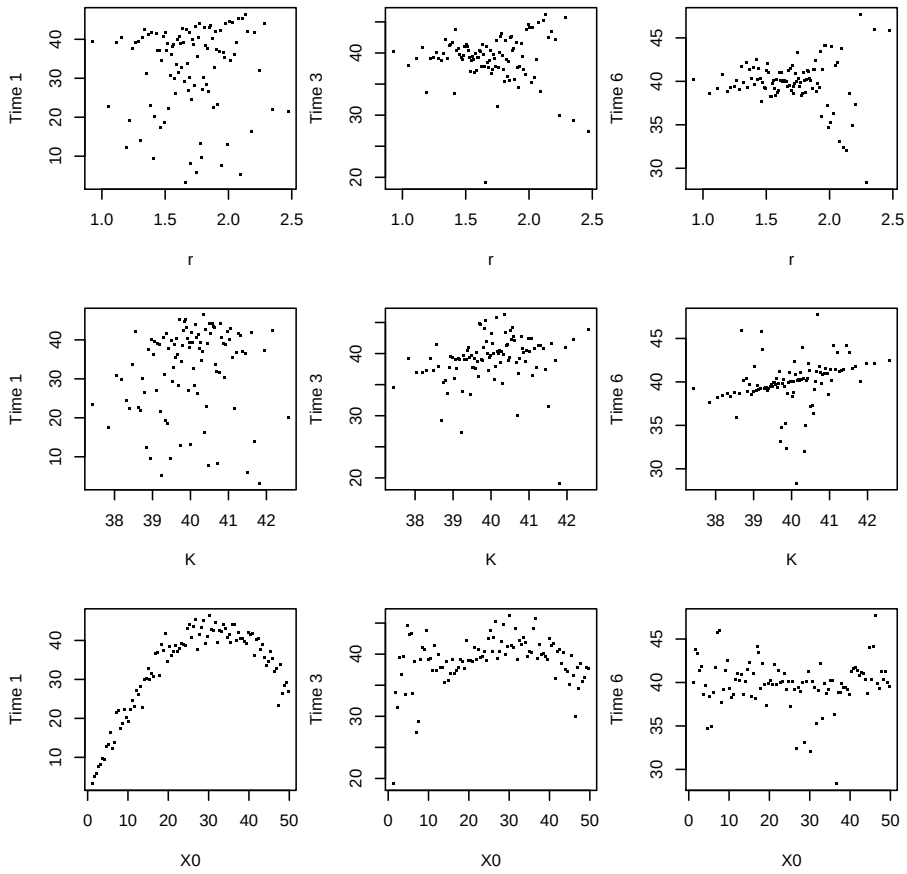


We may notice that the population values are spread over a wider range in the first time steps, but converge to a narrow distribution on time 6.

A.4.2 Scatterplots

Next, we investigate the correlation plots for the variables with each input parameter. To reduce the number of plots, we will present results just for the time steps 1, 3 and 6, using the "index.res" parameter, and suppress the linear model from being plotted with the parameter "add.lm":

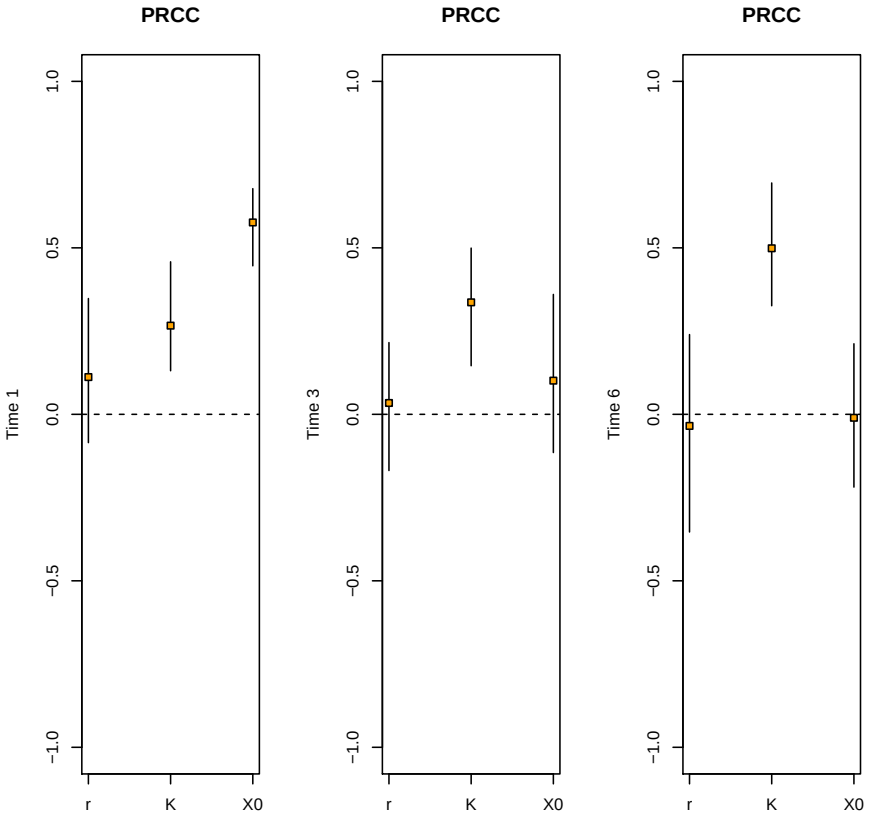
```
> plotscatter(myLHS, index.res=c(1,3,6), add.lm=FALSE)
```



A.4.3 Partial correlation

The partial correlation plots also accept the "index.res" argument:

```
> plotprcc(myLHS, index.res=c(1,3,6))
```



A.4.4 Agreement between runs

We have seen how the function **sbma** measures the agreement between two runs in the previous section. Now, we will use the function **target.sbma** to run several hypercubes until a pair of runs provides us with an agreement equal to or better than a limit specified in its first argument.

```
> targetLHS <- target.sbma (target=0.3, modelRun, factors,
+         q, q.arg, res.names, FUN=min)

[1] "INFO: initial run..."
[1] "INFO: LHS with N = 105"
[1] "sbma of -1 (target 0.3)"
[1] "INFO: LHS with N = 205"
[1] "sbma of 0.375 (target 0.3)"
```

As the SBMA is calculated for each response variable independently, we must decide how to combine these values. The argument "FUN=**min**" is telling the function to consider only the minimum value, and may be ignored for models that return a single response variable.

A.5 Uncoupling simulation and analysis

In many scenarios, it is necessary to run the simulation and the analyses at different times, and even in different computers. It may be the case, for example, that your lab computer is not fast enough to run the simulations, but that the high-performance cluster in which the simulations are to be run does not have **R** installed. In order to do this, however, you must generate the Latin Hypercube in the lab computer, transfer this information to the cluster, run the simulations there, transfer the results back to the lab computer, and then run the analyses.

In order to generate the samples without running a model, use the function **LHS** with the parameter *model=NULL* and save the samples in the desired format:

```
> uncoupledLHS <- LHS(model=NULL, factors, 50, q, q.arg)
> write.csv(get.data(uncoupledLHS), file="mydata.csv")
```

Then run the model using the data. To incorporate the results into the LHS object, use the function **tell**: [4]

```
> coupledLHS <- tell(uncoupledLHS, myresults)
```

Then you may proceed with the analyses using prcc, ecdf, etc.

[4]Please note that the **tell** method implemented in the sensitivity package alters its argument. This is **not** the case with the LHS **tell** method.

Appendix B

Multiple runs of the same parameter combination with R package pse

This document presents an extension to the practical tutorial found in the vignette "pse_tutorial" of this package. If you're not familiar with the basic concepts of parameter space exploration, please refer to the tutorial. To read about the underlying theory, please refer to our work in [Chalom and Prado, 2012]. You should have installed **R**[1] along with an interface and text editor of your liking, and the package "pse" (available on http://cran.r-project.org/web/packages/pse).

The problem we will address here is how to deal with the parameter space exploration of stochastic models. When considering deterministic models, the usual approach to parameter space exploration is to generate a hypercube containing several parameter combinations of interest, running the model with each paramter combination, and summarizing the results with empirical cummulative density functions (ECDFs) and partial rank correlation coefficients (PRCCs) between each model input and output variables. However, if the model isn't deterministic, a single run of the model may not be able to provide enough information about a particular point in the parameter space. Thus, it is necessary to run the model several times at the same combinations, and summarize the information for each point before proceeding to the uncertainty and sensibility analyses. This approach is discussed in [Marino et al., 2008], and our terminology is based on this paper. It should be noted that these problems are part of a fresh and open area of study, both in the biology and the engineering communities. The simplest solution is to evaluate average responses from many runs with each combination of parameters. This tutorial present some simple tools to do this and to evaluate how much of total variation is captured by averaging the responses.

It is usual to refer to the *aleatory* and *epistemic* components of the uncertainty as, respectively, the uncertainty due to random variation of the model behaviour for a fixed combination of parameters and the uncertainty due to our knowledge about the values of the parameters. For comparison, deterministic models only present

[1] This tutorial was written and tested with **R** version 3.0.1, but it should work with newer versions

epistemic uncertainty.

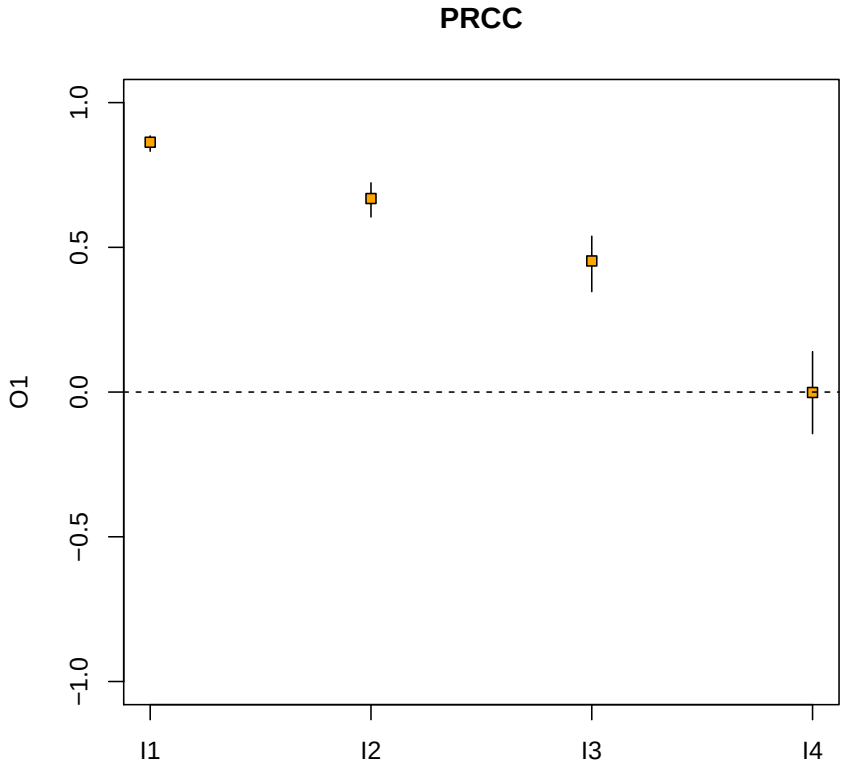
B.1 A simple example

We will show the use of multiple runs in the "pse" package with a very simple model described by:

```
> oneRun <- function (x1, x2, x3, x4)
+     10 * x1 + 5 * x2 + 3 * rnorm(1, x3, x4)
> modelRun <- function (my.data){
+     mapply(oneRun,
+            my.data[,1], my.data[,2], my.data[,3], my.data[,4])
+ }
```

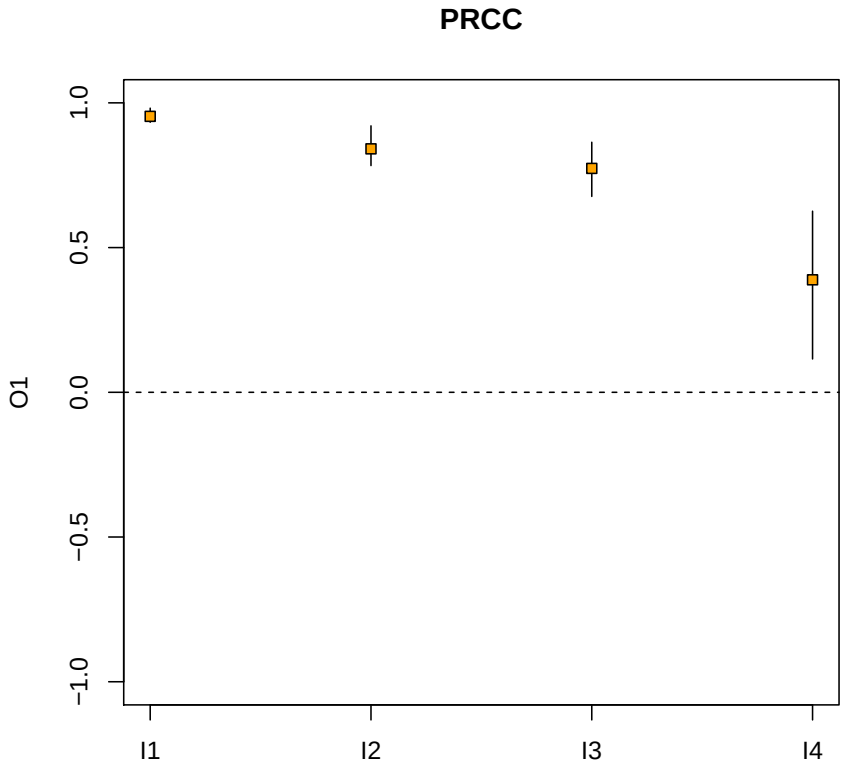
We then generate a Latin Hypercube where parameters x_i will be varied uniformly between 0 and 1. The code for a hypercube with a single run at each point and resulting partial correlation plot is:

```
> library(pse)
> LHS1 <- LHS(modelRun, N=300, factors=4, nboot=50)
> plotprcc(LHS1)
```



Now, we take a look at the same model, but *repeating* the simulation several times for each data point and averaging the results. The *repetitions* argument sets the number of evaluations for each combination of parameters. Note that the total number of model evaluations ($N \cdot repetitions$) is the same, in our case, as the LHS1 defined above:

```
> LHS2 <- LHS(modelRun, N=60, factors=4, repetitions=5, nboot=50)
> plotprcc(LHS2)
```



It should be clear from the graphs that the results we get from both schemes (which we will call single-run and repetition schemes) are different, even for a simple model like this. The repetition scheme usually results in larger values of PRCC for variables which are actually correlated to the model output, by mitigating the aleatory uncertainty in each data point. However, this comes at a cost of having less samples to use in statistical inference (resulting in loss of the power of significance tests, or in our case, an increase in the bootstrapped confidence intervals).

The repetition scheme also has the advantage that it permits a crude estimate of the aleatory uncertainty by means of the *coefficients of variation* (cv). The **cv** and **plotcv** functions can be used to identify whether the aleatory variability is comparable to the total model variability. This plot presents an empirical cumulative distribution function (ecdf) of the variation of the model responses obtained for each parameter combination (point wise cv). The dotted vertical line corresponds to the variation of the average model response through all combinations (global

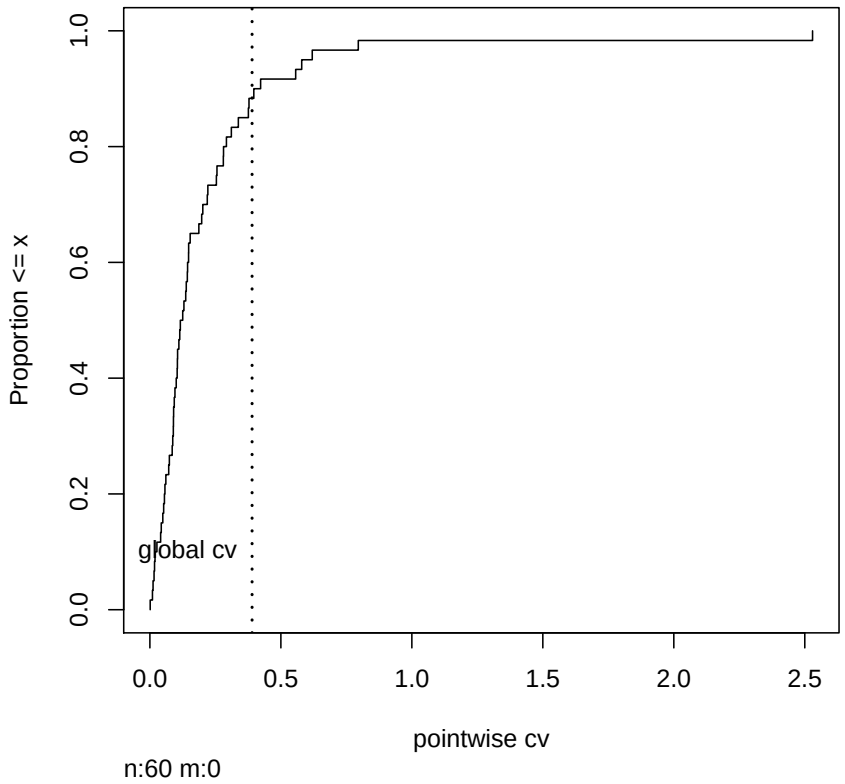
cv). If the global cv is far greater than all of the pointwise cvs, this means that the epistemic variability is far greater than the aleatory variation for any point. In contrast, if the global cv appears to the left of the graph, thus being smaller than most pointwise cvs, this is probably a sign that the aleatory variation may be masking the effect of the parameter variation, and so the sensitivity analyses will probably be compromised.

In our example, the cv of the whole result set is comparatively larger than most of the pointwise cvs. It is then reasonable to assume that the uncertainty and sensitivity analyses done with average responses are robust.

```
> cv(get.results(LHS2)) # global CV

[1] 0.3900826

> plotcv(LHS2)
```



One caveat of using the **cv** is that it is only meaningful if the distribution of results for a given point in the parameter space is unimodal. It is strongly recommended that you check the model behavior for multimodality before applying any of the uncertainty and sensitivity analyses discussed here. Strong multimodality is often evident from the scatterplots generated for single-run hypercubes.

B.2 Uncoupling analyses

The **tell** method of the LHS package can be used several times to add repetitions to an object. This can be done to uncouple the generation of the LHS and the running of the model, or even to provide more data points to a hypercube. For example, to add a repetition to the LHS2 object, simply execute:

```
> newdata <- modelRun(get.data(LHS2))
> LHS3 <- tell(LHS2, newdata)
```

This can be used to iteratively add repetitions until the PRCC is stable. In this example, the PRCC scores show very little change after 6 to 7 repetitions. This procedure can be coupled with SMBA evaluations between different hypercube sizes to find an acceptable bound for the number of total model evaluations.

Appendix C

PLUE: a suggested methodology for likehood profiling of model results

This document presents a brief introduction to a proposed methodology for the likelihood profiling of of the results from a computational model. This methodology is nicknamed PLUE, for Profiled Likehood Uncertainty Estimation, and is implemented in the *pse* package by the PLUE function. A paper describing the theoretical background for this proposal is under preparation for publication. The present document presumes you are familiar with general concepts from parameter space exploration. If you are not, please refer to our work in [Chalom and Prado, 2012]. The PLUE methodology is useful if you are interested in analysing a computational model and if you have already gathered some data from which you can estimate likelihood distributions for your input parameters. If you are interested in conducting an exploratory analysis and you don't have any data collected, you should use the tools described in the "pse_tutorial" vignette in this package. You should have installed **R**[1] along with an interface and text editor of your liking, and the package "pse" (available on http://cran.r-project.org/web/packages/pse).

The general question we are attempting to answer here is: *how much support does the data give to alternative hypothesis concerning the result of a (non-invertible) model?* It should be noted that while this question may not be always answered under a likehoodist approach to statistical inference, it does have an answer when we restrict one of the alternative hypothesis to being the maximum likehood estimator for the parameters. This answer is given by profiling the likelihood of the model parameters, and while this procedure leads to a function that is not a true likelihood function (thus not possessing many desirable properties), it is generally accepted as a valid exploratory analysis.

[1]This tutorial was written and tested with **R** version 3.0.1, but it should work with newer versions

C.1 Biological and statistical models

First, we should define our interest model. We will refer to this model as the biological model[2] to distinguish this from the statistical model we will be using to estimate likelihoods. This model must be formulated as an **R** function that receives a *data.frame*, in which every column represent a different parameter, and every line represents a different combination of values for those parameters. The function must return an array with the same number of elements as there were lines in the original data frame, and each entry in the array should correspond to the result of running the model with the corresponding parameter combination. For example, it can be this:

```
> oneRun <- function (r, K, Xo) {
+     X <- Xo
+     for (i in 0:20) {
+        X <- X+r*X*(1-X/K)
+     }
+     return (X)
+ }
> modelVec <- Vectorize(oneRun)
> model <- function(x) modelVec(x[,1], x[,2], x[,3])
```

Following the definition of the model, we should define the likelihood function for our parameters. To do this, we can formulate and test several statistical models. In order to fit competing models to the data and select the best of them, we recommend using the **R** package **bbmle**. Then, the best model should be written as a function receiving a numeric vector representing one realization of the parameter vector and returning the *positive* log-likelihood of that vector.

For example, the best model may be that the parameters r, K and Xo are all independent from each other, coming from two exponentials and from the *size* parameter of one binomial distribution, respectively, fitting the data in the *observations* data.frame below:

```
> r <- c(1.4, 1.2, 1.8)
> K <- c(70, 85, 98)
> Xo <- c(50, 60, 45)
> obs = data.frame(r=r, K=K, Xo=Xo)
```

The likelihood function, in this case, should be:

```
> LL <- function (x)
+ {
+         t <- sum(dexp(1/obs$r, as.numeric(x[1]), log=TRUE)) +
+                  sum(dexp(1/obs$K, as.numeric(x[2]), log=TRUE)) +
+                  sum(dbinom(obs$Xo, as.integer(x[3]), p=0.5, log=TRUE))
```

[2]Because the models of interest in my research are biological. It can also be a physical model, geochemical model, etc.

```
+         if (is.nan(t)) return (-Inf);
+         return(t);
+ }
```

Please note that this function uses the global variable *obs*, and that it return minus infinity instead of not-a-number in cases where the likelihood is not properly defined. This can happen, for instance, if any of the values of x is negative. Also, notice that this function converts the third element of x to integer, as the *dbinom* function does not accept a fractional value for the *size* parameter.

C.2 Profiling: sampling and aggregating the results

After carefully constructing the model of interest and the likelihood function, as described in the previous section, performing the PLUE analysis is simply a matter of calling the *PLUE* function. This function performs three steps. First, it performs a Monte Carlo sampling of the likelihood function in order to generate a large sample from the likelihood distribution. Then, the biological model is applied to this sample, and finally the model results are combined by means of profiling the likelihood function associated with each data point.

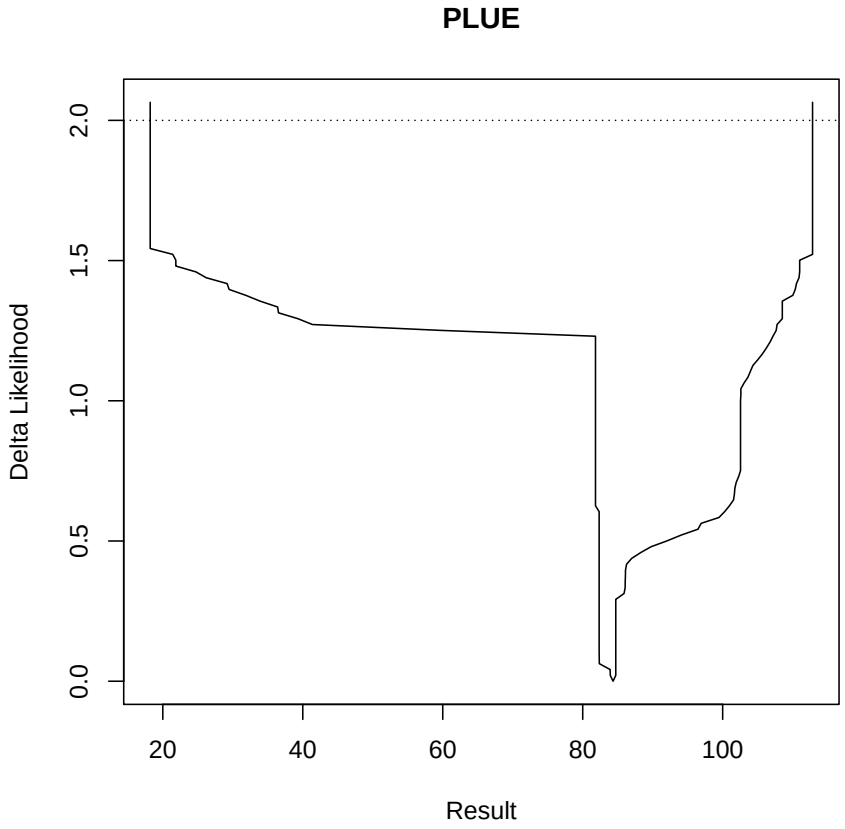
The *pse* package implements a simple Metropolis sampling function that can be used by setting *method='internal'* in the *PLUE* function call. For more elaborate sampling schemes, and more control over the process, we recommend using *method='mcmc'*, which uses the *mcmc* **R** package.

```
> library(pse)
> factors = c("r", "K", "X0")
> set.seed(42)
> N = 10000
> # The starting point for the Monte Carlo sampling
> start = c(mean(obs$r), mean(obs$K), 2*max(obs$Xo))
> plue <- PLUE(model, factors, N, LL, start)
```

Important note: the example above uses a N of 10.000, which is very low. For practical applications, always use larger samples and the 'mcmc' method.

In order to see the profiled likelihood of the model result, simply run:

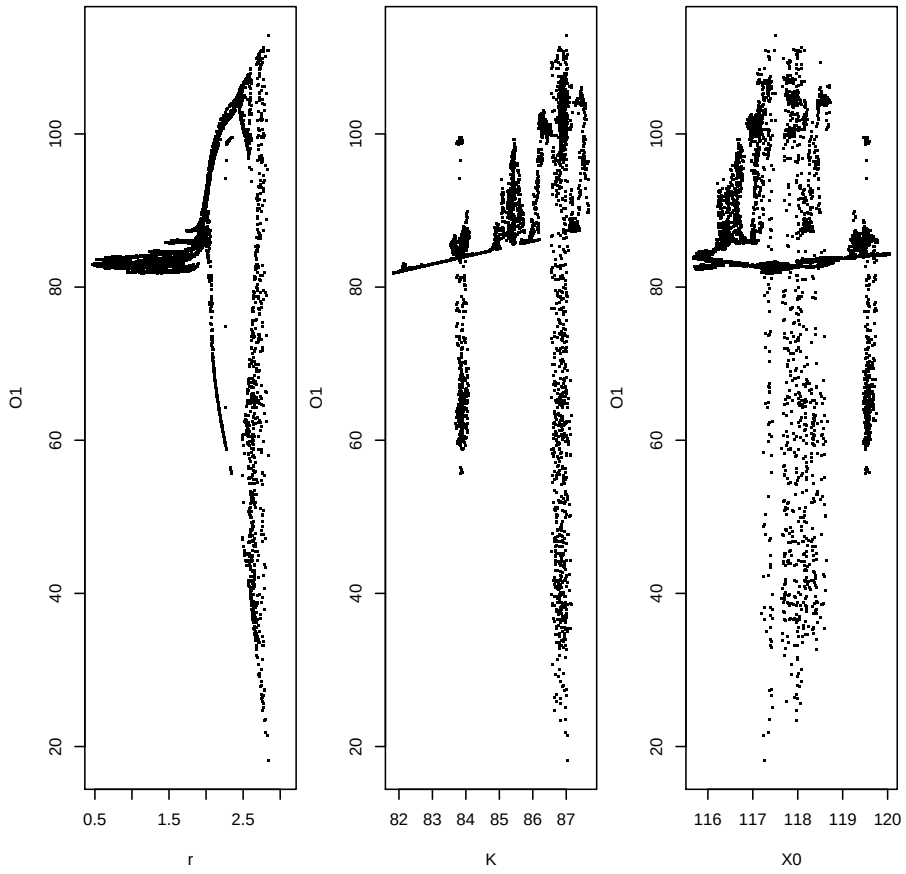
```
> plot(plue)
```



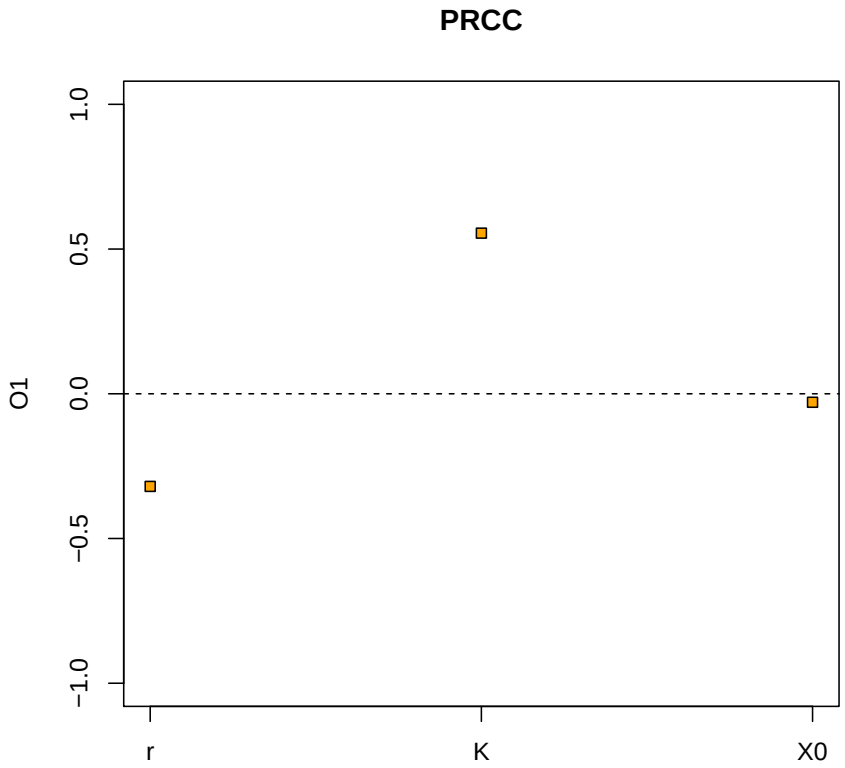
The profile seen in this figure shows that the model result is very unreliable. It can be considered very plausible, from the data collected, that the result of the model lie anywhere in the $20 - 120$ interval.

Additional plots may be seen by using the *plotscatter* and *plotprcc* functions:

```
> plotscatter(plue, add.lm=F)
```




```
> plotprcc(plue)
```



The interpretation of these graphs is analogous to the graphs generated by the Latin Hypercube Sampling, and described in the "pse_tutorial" vignette. However, it is important to notice that, instead of the arbitrary region of the parameter space that is sampled in the LHS scheme, the plots presented in this vignette are representing a discretization of the likelihood surfaces of the parameters, thus incorporating all the information about the data collected.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Aldrich, J. (2008). R. A. Fisher on Bayes and Bayes' Theorem. *Bayesian Analysis*, 3(1):161–170.

Anscombe, F. and Aumann, R. (1963). A definition of subjetive probability. *Annals of Mathematical Statistics*, 34(1):199–205.

Archer, G., Saltelli, A., and Sobol, I. (1997). Sensitivity measures, ANOVA-like techniques and the use of bootstrap. *J. Statist. Comput. Simul.*, 58:99–120.

Barber, J. and Ogle, K. (2014). To p or not to p? *Ecology*, 95(3):621–626.

Bart, J. (1995). Acceptance criteria for using indivual-based models to make management decisions. *Ecological applications*, 5(2):411–420.

Basu, D. (2011). Statistical information and likelihood. In DasGupta, A., editor, *Selected Works of Debabrata Basu*, pages 207–278. Springer.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York, USA.

Berthaume, M., Dechow, P., Iriarte-Diaz, J., Ross, C., Strait, D., Wang, Q., and Grosse, I. (2012). Probabilistic finite element analysis of a craniofacial finite element model. *Journal of Theoretical Biology*, 300(0):242 – 253.

Beven, K. and Binley, A. (1992). The future of distributed models: model calibration and predictive uncertainty. *Hydrol. Processes*, 6:279–298.

Bickel, D. (2010). The strength of statistical evidence for composite hypotheses: Inference to the best explanation. COBRA Preprint Series.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–326.

Bladt, M. and Nielsen, B. F. (2010). On the construction of bivariate exponential distributions with an arbitrary correlation coefficient. *Stochastic Models*, 26(2):295–308.

Bolker, B. (2008). *Ecological Models and Data in R*. Princeton University Press.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Burnham, K., Anderson, D., and Huyvaert, K. (2011). Aic model selection and multimodal inference in behavioral ecology: some background, observations and comparisons. *Behavioral Ecology and Sociobiology*, 61(1):23–25.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Carnap, R. (1962). *Logical Foundations of Probability*. University of Chicago Press.

Caswell, H. (1989). *Matrix population models*. John Wiley & Sons.

Caswell, H. (2008). Perturbation analysis of nonlinear matrix population models. *Demographic Research*, 18:59–116.

Caswell, H. (2009). Sensitivity and elasticity of density-dependent population models. *Journal of Difference Equations and Applications*, 15(4):349–369.

Caswell, H. (2010). Reproductive value, the stable stage distribution, and the sensitivity of the population growth rate to changes in vital rates. *Demographic Research*, 23:531–548.

Chalom, A. and Prado, P. (2012). Parameter space exploration of ecological models. arXiv:1210.6278 [q-bio.QM].

Cole, L. (1954). The population consequences of life history phenomena. *The Quarterly Review of Biology*, 29(2):103–137.

Confalonieri, R., Bellocchi, G., Bregaglio, S., Donatelli, M., and Acutis, M. (2010). Comparison of sensitivity analysis techniques: A case study with the rice model WARM. *Ecological Modelling*, 221(16):1897 – 1906.

Daston, L. and Galison, P. (2007). *Objectivity*. Zone Books, NY.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In Kyburg, H. E. and Smokler, H. E., editors, *Studies in subjective probability*, page 0. Robert E. Kreiger Publishing Co.

de Finetti, B. (1951). Recent suggestions for the reconciliation of theories of probability. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 217–225. Berkeley: University of California Press.

de Finetti, B. (2010). *Philosophical Lectures on Probability*. Springer-Verlag, New York.

de Morgan, A. (1847). *Formal logic, or, The calculus of inference, necessary and probable*. Taylor and Walton, London.

Dennis, B., Desharnais, R. A., Cushing, J., and Costantino, R. (1995). Nonlinear demographic dynamics: mathematical models, statistical methods, and biological experiments. *Ecological Monographs*, 65(3):261–282.

Duchesne, S., Boyer, P., and Beaugelin-Seiller, K. (2003). Sensitivity and uncertainty analysis of a model computing radionuclides transfers in fluvial ecosystems (CASTEAUR): application to ^{137}Cs accumulation in chubs. *Ecological Modelling*, 166(3):257 – 276.

Dukic, V. M. and Marić, N. (2013). Minimum correlation in construction of multivariate distributions. *Physical Review E*, 87(3):032114.

Edwards, A. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Estill, J., Aubriere, C., Egger, M., Johnson, L., Wood, R., Garone, D., Gsponer, T., Wandeler, G., Boulle, A., Davies, M., et al. (2012). Viral load monitoring of antiretroviral therapy, cohort viral load and hiv transmission in southern africa: A mathematical modelling analysis. *AIDS*, 26:000–000.

Fetzer, J. (1993). Peirce and propensities. In Moore, E., editor, *Charles S. Peirce and the philoposhy of science*, pages 60–71. University of Alabama Press.

Fisher, R. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222:309–368.

Fisher, R. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Fisher, R. (1955). Statistical methods and scientific induction. *Philosophical Transactions of the Royal Society of London. Series B*, 17:69–78.

Fisher, R., McDowell, N., Purves, D., Moorcroft, P., Sitch, S., Cox, P., Huntingford, C., P., M., and F.I., W. (2010). Assessing uncertainties in a second-generation dynamic vegetation model caused by ecological scale limitations. *New Phytologist*, 187:666–681.

Fitelson, B. (2007). Likelihoodism, Bayesianism, and relational confirmation. *Synthese*, 156(3):473–489.

Florian, A. (1992). An efficient sampling scheme: Updated latin hypercube sampling. *Probabilistic Engineering Mechanics*, 7:123–130.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. Number p. 2 in International student edition. W.W. Norton and Company.

Friedman, M. and Savage, L. J. (1948). Utility analysis of choices involving risk. *Journal of Political Economy*, 56(4):279–304.

Galavotti, M. (1989). Anti-realism in the philosophy of probability: Bruno de Finetti's subjectivism. *Erkenntnis*, 31:239–261.

Gandenberger, G. (2012). A new proof of the likelihood principle. *British Journal for the Philosophy of Science*.

Gardner, M. and Altman, D. (1986). Confidence intervals rather than p values: estimation rather than hypothesis testing. *British medical journal*, 292(6522):746.

Genest, C. and Plante, J. (2003). On blest's measure of rank correlation. *The Canadian Journal of Statistics*, 31(1):35–52.

Geyer, C. (2012). Calling C and Fortran from R. http://users.stat.umn.edu/ geyer/rc/.

Gillies, D. (2000). Varieties of propensity. *Brit. J. Phil. Sci.*, 51:807–835.

Good, I. (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. In Harper, W. and Hooker, C., editors, *Foundations of probability theory, statistical inference, and statistical theories of science*, page 0. Reidel.

Good, I. (1992). The Bayes/non-Bayes compromise: a brief review. *Journal of the American Statistical Association*, 87(419):597–606.

Goodman, N. (1983). *Fact, fiction, and forecast*. Harvard University Press, fourth edition edition.

Gudder, S. (1988). Quantum probability. In Kurt Engesser, Dov M. Gabbay, D. L., editor, *Handbook of quantum logic and quantum structures*, page 0. Elsevier.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge university press, New York.

Hamilton, A., Basset, Y., Benke, K., Grimbacher, P., Miller, S., Novotny, V., Samuelson, G., Stork, N., Weiblen, G., and Yen, J. (2010). Quantifying Uncertainty in Estimation of Tropical Arthropod Species Richness. *American Naturalist*, 176(1):90–95.

Helton, J., Davis, F., and Johnson, J. (2005). A comparison of uncertainty and sensivity analysis results obtained with random and latin hypercube sampling. *Reliability Engineering and System Safety*, 89:304–330.

Helton, J. and Davis, J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*, 81:23–69.

Hoeffding, W. (1940). Scale-invariant correlation theory. *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universit at Berlin*, 5(3):181–233.

Huber, M. and Marić, N. (2014). Minimum correlation for any bivariate geometric distribution. arXiv:1406.1779 [math.PR].

Huntington, D. and Lyrintzis, C. (1998). Improvements to and limitations of latin hypercube sampling. *Prob. Engng. Mech.*, 13(4):245–253.

Hájek, A. (2012). Interpretations of probability. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition.

Iman, R. and Conover, W. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in statistics*, B11(3):311–334.

Iman, R. and Conover, W. (1987). A measure of top-down correlation. *Technometrics*, 29(3):351–357.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med*, 2(8):e124.

Jaynes, E. (1968). Prior probabilites. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241.

Jennions, M. and Moeller, A. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav Ecol*, 14:438–445.

Johnson, J. B. and Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2):101–108.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 175–208.

Kleijnen, J. and Helton, J. (1999). Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: Review and comparison of techniques. *Reliability Engineering and System Safety*, 65:147–185.

Koopman, B. O. (1940). The axioms and algebra of intuitive probability. *Annals of Mathematics*, pages 269–292.

Kuczera, G. and Parent, E. (1998). Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology*, 211:69–85.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kyburg, H. (1974). *The Logical foundations of statistical inference*. Reidel, Dordrecht.

Laplace, P. (1814). *Essai philosophique sur les probabilités*. Bachelier, imprimeur-libraire, Paris.

LeBlanc, D. (2004). *Statistics: Concepts and Applications for Science*. Jones and Bartlett.

Lenhard, J. (2006). Models and statistical inference: the controversy between Fisher and Neyman-Pearson. *Brit J Phil Sci*, 57:69–91.

Letcher, B., Rice, J., Crowder, L., and Rose, K. (1996). Variability in survival of larval fish: disentangling components with a generalized individual-based model. *Can. J. Fish. Aquat. Sci.*, 53:787–801.

Lindley, D. (1982). Scoring rules and the inevitability of probability. *International Statistical Review*, 50(1):1–11.

Loève, M. (1977). *Probability Theory*. Springer Verlag.

Lovvorn, J. and Gillingham, M. (1996). Food dispersion and foraging energetics: A mechanistic synthesis for field studies of avian benthivores. *Ecology*, 77(2):435–451.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mandel, L. and Wolf, E. (1995). *Optical Coherence and Quantum Optics*. Cambridge University Press, Cambridge.

Marino, S., Hogue, I., Ray, C., and Kirschner, D. (2008). A methodology for performing global uncertainty and sensivity analysis in systems biology. *Journal of Theoretical Biology*, 254:178–196.

Maturi, T. and Elsayigh, A. (2010). A comparison of correlation coefficients via a three-step bootstrap approach. *Journal of Mathematics Research*, 2(2):3–10.

May, R. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261:459–466.

Mayo, D. (2010). An error in the argument from conditionality and sufficiency to the likelihood principle. In Mayo, D. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, pages 305–314. Cambridge University Press.

McKay, M. and Beckman, R. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–244.

Meisner, R. (2010). Advanced Simulation and Computing Program Plan FY11. Technical report, Office of Advanced Simulation and Computing, NNSA Defense Programs. NA-ASC-122R-10.

Meyer, K., Wiegand, K., Ward, D., and Moustakas, A. (2007). SATCHMO: A spatial simulation model of growth, competition, and mortality in cycling savanna patches. *Ecological Modelling*, 209(2–4):377 – 391.

Miller, R. (1975). Propensity: Popper or peirce? *British Journal for the Philosophy of Science*, 26(2):123–132.

Milne, P. (1993). The foundations of probability and quantum mechanics. *Journal of Philosophical Logic*, 22(2):129–168.

Moore, H. and Li, N. (2004). A mathematical model for chronic myelogenous leukemia (CML) and T cell interaction. *Journal of Theoretical Biology*, 227(4):513 – 523.

Morettin, P. and Bussab, W. (2009). *Estatística básica*. Editora Saraiva, São Paulo, 6 edition.

Morris, M. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Murray, J. D. (2002). *Mathematical Biology I: An Introduction, vol. 17 of Interdisciplinary Applied Mathematics*. Springer, New York, NY, USA,.

Nathan, R., Safriel, U., and Noy-Meir, I. (2001). Field validation and sensitivity analysis of a mechanistic model for tree seed dispersal by wind. *Ecology*, 82(2):374–388.

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337.

Peirce, C. (1883). A theory of probable inference. In Peirce, C. S., editor, *Studies in Logic*. Little, Brown, and Company, Boston.

Peirce, C. (1892). The law of mind. *The Monist*, 2(4):533–559.

Pfanzagl, J. (1967). Subjective probability derived from the Morgenstern-von Neumann utility theory. In Shubik, M., editor, *Essays in Mathematical Economics In Honor of Oskar Morgenstern*, pages 237–251. Princeton University Press.

Popham, W. (1986). Curriculum, instruction and assessment: amiable allies or phony friends? *The Teachers College Record*, 106(3):417–428.

Popper, K. (1963). *Conjectures and refutations*. Routledge, London.

Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science*, pages 25–42.

Potvin, C., Lechowicz, M., and Tardif, S. (1990). The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. *Ecology*, 71:1389–1400.

Quine, W. V. O. (1970). Natural kinds. In et al., N. R., editor, *Essays in Honor of Carl G. Hempel.* Dordrecht.

Reed, K., Rose, K., and Whitmore, R. (1984). Latin hypercube analysis of parameter sensitivity in a large model of outdoor recreation demand. *Ecological Modelling*, 24(3-4):159–169.

Ross, G. (1990). *Nonlinear estimation*. Springer series in statistics. Springer-Verlag.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, London.

Saltelli, A. (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Wiley, Hoboken, NJ.

Saltelli, A., Tarantola, S., and Chan, K. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56.

Savage, L. (1961). The foundations of statistics reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

Savage, L. (1967). Implications of personal probability for induction. *The Journal of Philosophy*, 64(19):593–607.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Segovia-Juarez, J., Ganguli, S., and Kirschner, D. (2004). Identifying control mechanisms of granuloma formation during *m. tuberculosis* infection using an agent-based model. *J Theor Biol*, 231(3):357–376.

Shafer, G. and Vovk, V. (2003). The origins and legacy of Kolmogorov's Grundbegriffe. http://www.probabilityandfinance.com/articles/04.pdf.

Shirley, M., Rushton, S., Smith, G., South, A., and Lurz, P. (2003). Investigating the spatial dynamics of bovine tuberculosis in badger populations: evaluating an individual-based simulation model. *Ecological Modelling*, 167(1–2):139–157.

Silva Matos, D., Freckleton, R., and Watkinson, A. (1999). The role of density dependence in the population dynamics of a tropical palm. *Ecology*, 80(8):2635–2650.

Smith, E. (2002). Uncertainty analysis. In *Encyclopedia of Environmetrics*, volume 4, pages 2283–2297. John Wiley and Sons, Chichester, UK.

Sparks, D. and Sparks, D. (2003). *Environmental Soil Chemistry*. Academic Press.

Steinberg, D. and Lin, D. (2006). A construction method for orthogonal latin hypercube designs. *Biometrika*, 93(2):279–288.

Stigler, S. (1978). Mathematical statistics in the early States. *The Annals of Statistics*, 6(2):239–265.

Stoica, P. and Selen, Y. (2004). Model-order selection: a review of information criterion rules. *Signal Processing Magazine, IEEE*, 21(4):36–47.

Tang, B. (1998). Selecting latin hypercubes using correlation criteria. *Statistica Sinica*, 8:965–977.

Thébault, E. and Fontaine, C. (2010). Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329:853.

Tiemeyer, B., Moussa, R., Lennartz, B., and Voltz, M. (2007). Mhydas-drain: A spatially distributed model for small, artificially drained lowland catchments. *Ecological Modelling*, 209(1):2 – 20.

Tung, T. Q. and Lee, D. (2010). Psexplorer: whole parameter space exploration for molecular signaling pathway dynamics. *Bioinformatics*, 26(19):2477–2479.

Turchin, P. and Hanski, I. (1997). An empirically based model for latitudinal gradient in vole population dynamics. *The American Naturalist*, 149(5):842–874.

Venn, J. (1866). *The Logic of Chance*. Macmillan and co., London and Cambridge.

Von Mises, R. (1941). On the foundations of probability and statistics. *The Annals of Mathematical Statistics*, 12(2):191–205.

Vořechovskỳ, M. and Novák, D. (2009). Correlation control in small-sample monte carlo type simulations i: A simulated annealing approach. *Probabilistic Engineering Mechanics*, 24(3):452–462.

Vrugt, J., ter Braak, C. J., Gupta, H., and Robinson, B. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrological modeling. *Stoch Environ Res Risk Assess*, 23(7):1011–1026.

Weakliem, D. L. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):359–397.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

Xu, C., He, H., Hu, Y., Chang, Y., Li, X., and Bu, R. (2005). Latin hypercube sampling and geostatistical modeling of spatial uncertainty in a spatially explicit forest landscape model simulation. *Ecological Modelling*, 185(2–4):255 – 269.

Yang, Y. and Atkinson, P. (2008). Parameter exploration of the raster space activity bundle simulation. *J. Geograph. Syst.*, 10:263–289.

Ye, K. (1998). Orthogonal column latin hypercubes and their application in computer experiments. *Journal of the American Statistical Association*, 93(444):1430–1439.

Zabell, S. (2009). Carnap and the logic of inductive inference. In *Handbook of the History of Logic*, pages 265–309. Elsevier BV.

Zhang, Z. (2009). A law of likelihood for composite hypotheses. arXiv:0901.0463 [math.ST].

Zhang, Z. and Zhang, B. (2013). A likelihood paradigm for clinical trials. *Journal of Statistical Theory and Practice*, 7(2):157–177.