Rijksuniversiteit Groningen

Distributional Inference: The Limits of Reason

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
vrijdag 7 maart 2003
om 14.15 uur
door

Casper Johannes Albers

geboren op 30 juni 1975
te Hengelo (Overijssel)

Promotor:
Prof.dr. W. Schaafsma

Beoordelingscommissie:
Prof.dr. H.G. Dehling
Prof.dr. R.C. Jansen
Prof.dr. A.W. van der Vaart

Preface

Stripped to its essentials, statistics is about inferences and decisions on the basis of numerical evidence which is statistical in the sense that *things could have been different*. Which inferences and decisions are 'most appropriate', given some initial ideas and a set of data? That is the question to which the mathematical statistician restricts his attention. The applied statistician has a more comprehensive task. He will have to participate in the discussion about the design of the experiments, the choice of the data to be evaluated, the interpretation of the inferences, etcetera.

It is true, and completely natural, that the 'procedures' proposed by mathematical statisticians are based on principles which can be modified. In practice, the relevant data should be sufficiently abundant to accept the corresponding inferences or decisions as 'reasonable'. If factual information is scarce or irrelevant or not trustworthy then one should not rely on the precise results prescribed by such procedures. Anybody should be aware of what Kant called *the limits of reason* ('die Grenzen der Vernunft').

Procedures for making inferences on the basis of data are usually based on a mathematical model which comprises a specification of the probabilistic context within which the data arise, and a specification of the inferential or decision-making context.

Probability statements and distributional inferences are of particular interest, e.g. as an intermediary between data and decision. 'The making of statistical inferences in distributional form is conceptionally complicated because the epistemic 'probabilities' assigned are mixtures of fact and fiction. In this respect they are essentially different from 'physical' or 'frequency-theoretic' probabilities. The distributional form is so attractive and useful, however, that it should be pursued'[1].

In Part I of this thesis it will be made very clear, by elaborating on examples, that the precise probability statements and distributional inferences prescribed by some 'rational' mathematical-statistical procedure are sometimes not relevant. If, however, the factual information is sufficiently abundant, then the inferences based on it deserve to play a part in the discussion. 'As workers in Science we aim, in fact, at methods of inference which shall be equally convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utilizing the knowledge inferred'[2]. Mathematical statisticians try to be as 'objective' as possible, even if they declare themselves to be subjectivists.

[1] A.H. Kroese et al., Distributional Inference, *Statistica Neerlandica*, **49**:1, 63–82, 1995

[2] R.A. Fisher, *Statistical Methods and Scientific Inference*, third edition, Macmillan, New York, 1973, p. 107

In Part II two types of (distributional) inference will be examined. The first one is that of a density (e.g. the density of a future observation) which has to be estimated on basis of a combination of a sample and a priori 'knowledge'. The second one is that of a statement which has to be made about the truth or falsity of the hypothesis that the density is exactly the one specified. In the latter case an accept-reject statement may be more appropriate than the assignment of a probability.

In Part III applications are made of the theory developed in Part II, but also with a concrete problem in mind. This two-way traffic is essential. Statistics needs applications and many applications need statistics. 'As regards mathematics, you cannot separate it from its applications to the external world, and you cannot separate statistics from mathematics, or mathematical statistics from applied statistics.'[3]

The reader might wonder whether something 'new' can be found in this thesis: many of the arguments to be used were already availably half a century ago. It is the combination of such arguments which is pursued. A specific feature of Chapter 3 is that it goes one step into the direction of a Bayesian approach by claiming that an a priori guess is available. The second step, the specification of an entire a priori distribution, is not made because it does not seem appropriate. The density estimates provided seem to be 'very good', though, slight modifications can still lead to further improvements.

Multi-modal 'compromises' can, of course, differ. The rational man likes logical validity and unicity of solution. Unfortunately, such niceties are not attainable in statistics. One will study a variety of approaches to the same problem, each one resulting in an expert opinion. If these opinions are sufficiently alike then any one of them will do, a summary can be presented. If the opinions are too much different, then one should *not* present such concluding summary.

Directions for the reader: this thesis contains a report of some of the problems I encountered for the past four years. Although these problems were of interest for my own development, not all of them are of direct interest to the 'average reader'. Such cases are mentioned where they occur, and the sections concerned could be skipped from reading.

Casper Albers
Groningen, December 2002

[3] M. KAC ET AL., *Discrete thoughts — Essays on Mathematics, Science, and Philosophy*, BIRKHÄUSER, BOSTON, 1986

Acknowledgements
I am indebted to my promotor Willem Schaafsma, who was always available with his support and enthusiasm. The theory of Chapter 2 was developed jointly with Barteld Kooi. The estimator of Chapter 3 was studied before by Rob de Bruin and Diemer Salomé. Other issues were brought under our attention by or were discussed with (in order of appearance in this thesis):

Sandy Zabell (Northwestern University of Illinois) commented on the Kullback-Leibler number (Section 1.3), Rieks op den Akker (University of Twente) made some valuable remarks on the prisoner's dilemma (Section 1.9), Lammert ter Veld (Groningen) referred us to the Protagoras paradox (Section 2.2), Roland Auer and Wim Oudshoorn (Groningen) suggested a decision function in Section 2.4, Wilbert Kallenberg (Twente) and Subha Chakraborti (University of Alabama) commented on Chapter 4, Richard Davis (University of Colorado) and Thomas Mikosch (University of Copenhagen) brought the unit root problem (Chapter 6) under our attention, Jan Hulscher (Groningen) and Gerlof de Roos advised on Chapter 7, Alwin Stegeman (Groningen) gave helpful advice on printing theses, and, finally, Willem Albers (Twente) advised for the Summary and Samenvatting.

The coworkers at the *Centre for Quantitative Methods* (CQM) in Eindhoven, especially Mynt Zijlstra, Jaap Praagman, Marijke Swaving, Bert de Vries, and Ruud van Lieshout, assisted me in constructing the first drafts of Chapter 5.

R.N.M. Duin of the *Rijksinstituut voor Kust en Zee* kindly granted us permission to use the data set *Monitoring Watertoestand des Lands* (Chapter 5). G.Th. de Roos granted permission to use his ornithological data (Chapter 7).

To my former Probability & Statistics colleagues in Groningen: Alwin, Bojan, Daniel, Diemer, Evgeni, Herold, Hui, Mook, Thomas, Willem, Wout: thank you for the nice working atmosphere in the periods you were in Groningen. Finally, thanks to Marijn.

Contents

Preface **iii**

I The limits of reason 1

1 How to assign probabilities if you must **3**

1.1 Predictive distributional inference, an example 3
1.2 Background information 4
1.3 A Fisher-Neyman-Pearson-Wald approach 6
1.4 A detailed discussion of some specific procedures 10
1.5 What if Player I uses a randomized strategy 12
1.6 Adapting the theory to alternative loss functions 14
1.7 Extension 20
1.8 An classroom example with Bernoulli trials 22
1.9 Related well-known examples 26

2 Trying to resolve the two-envelope problem **29**

2.1 The two-envelope problem 29
2.2 Explorations 30
2.3 Probability theory 33
2.4 Mathematical statistics 35
2.5 Game theory 38
2.6 Discussion: the limits of reason 43

II Statistical inference, distributional inference in particular 45

3 Estimating a density by adapting an initial guess **47**

3.1 Introduction 48
3.2 The original density estimation method 50

3.3 Asymptotic properties of the quantile function estimate 53
3.4 Asymptotic properties of the density function 57
3.5 Comparison with other methods 60
3.6 Improvements through U-statistic symmetrization 62
3.7 Simulation studies 64
3.8 Specifying the initial guess 68
3.9 Extending the theory to the bivariate case 72
3.10 Discussion 73

4 A goodness of fit test, smoother than smooth 75
4.1 Goodness of fit testing 75
4.2 Specifications 78
4.3 The extreme case $m = 1$ 80
4.4 The case $m = 2$ 82
4.5 The general case $m \geq 2$ 85
4.6 Relation with Neyman's smooth tests 88
4.7 Relations with other goodness of fit tests 92

III Applications 93

5 Analyzing water-quality data 95
5.1 Description of the data 95
5.2 All measurements are above the threshold 96
5.3 Some measurements below threshold 98
5.4 Substances involving multiple thresholds 101
5.5 Complications 101

6 The interface between time series and multivariate analysis 103
6.1 Introduction to the unit root problem 103
6.2 The case $n = 2$ 106
6.3 Exact inferences in the case $n > 2$ 110
6.4 Focusing on exact tests for the null hypothesis 111
6.5 Survey of the literature for the case $n = \infty$ 113
6.6 Discussion 114

7 Making statistical inferences about a frequency unseen 115

7.1 Introduction 115
7.2 Primary approaches 118
7.3 Obtaining a solution by ignoring day effects 119
7.4 Obtaining a solution by taking day effects into account 121
7.5 Discussion 125

Appendices 127

A Proof of Theorem 4.1 129

B Critical values for the $||f_n^{(m)} - \psi||_1$-test statistic 133

Bibliography 137

Author index 145

Summary 147

Samenvatting 149

Part I

The limits of reason

Chapter 1

How to assign probabilities if you must

> *'Without the information the model was clear: a random drawing from 1,...,6. After the information one is completely muddled. Giving out information should be accompanied by the knowledge of the information-policy.'*
>
> J. Hemelrijk[1]

Empirical evidence can sometimes be incorporated in a probabilistic analysis by conditioning with respect to the observations. Usually, the underlying probability distribution and also the conditional distribution are not completely known. The assignment of probabilities will then require a compromise. The making of such compromise goes beyond mathematical theory: a statistical discussion is needed. It depends on the context whether the result of such discussion is almost compelling, reasonable, or not really agreeable. This is illustrated by means of a simple example from the area of predictive distributional inference.

This chapter is an extended version of the article with the same title in Statistica Neerlandica (Albers and Schaafsma, 2001b) (2001b), and the technical report with the same title (Albers, 2000) (2000).

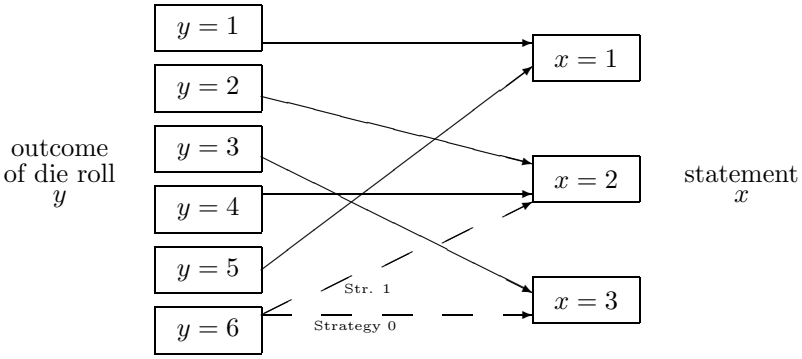
1.1 Predictive distributional inference, an example

Most theories of probability, Bayesian statistics included, prescribe to incorporate empirical evidence by computing conditional distributions. The availability of such prescriptions suggests that the approach is compelling. This may be the case if a fair die is rolled once and somebody has been instructed to tell us, without lying, whether the number of eyes y is even or odd, but the compellingness disappears if the instructions are less specific. If somebody provides us with the information that y is even, it can be very misleading to infer that the probabilities of the possibilities $y = 2,\ 4$ and 6 are equal to $\frac{1}{3}$.

The example

A fair die has been rolled once and the true number of eyes y has been made available to some person (or Nature), henceforth referred to as Player I. Player I has to provide Player II (the statistician) with true information x about y. He has to choose one of the statements made in Figure 1.1. Note that there is a difficulty if $y = 6$. In that case, Player I has to choose between $x = 2$ and $x = 3$.

[1] Rules for building statistical models, *Statistica Neerlandica*, **32**:3, 1978.



$x = 1$: y is neither even, nor a triple
$x = 2$: y is even
$x = 3$: y is a triple

Figure 1.1: Visualization of the outcome space

Question
Suppose Player I provides us with the information x (either 1, 2, or 3). Which probabilities $q_x(\eta)$ should be assigned to the theoretical possibilities η for y?

Preliminary exploration
If $x = 1$, then the theoretical possibilities are 1 and 5 and nobody will criticize the specification $q_1(1) = q_1(5) = \frac{1}{2}$ because the die is fair. If $x = 2$ we are in trouble because if we assume that Player I chooses $x = 2$ in case $y = 6$ (Strategy 1) then $q_2(2) = q_2(4) = q_2(6) = \frac{1}{3}$ is logical, but if we assume that he would choose $x = 3$ (Strategy 0) then $x = 2$ implies that $y = 6$ is impossible and that $q_2(2) = q_2(4) = \frac{1}{2}$ is appropriate . If $x = 3$, a similar discussion can be made. HEMELRIJK (1978) states that 'Information about the result of an experiment is only reliable if the receiver knows the information-policy used, i.e. if he knows which information would be given for every possible result of the experiment'. The nonexistence of a unique solution, due to an unknown information-policy, in the cases $x = 2$ and $x = 3$ implies that answering the question is only possible in the sense of making a compromise.

1.2 Background information

The situation now obtained is characteristic for almost everything in statistics: the solution depends on an unknown true value, here strategy number $i \in \{0, 1\}$. We shall discuss a logical approach, a Bayesian approach and we give a preview of the Fisher-Neyman-Pearson-Wald approach we recommend. This section will be concluded with some additional comments.

The situation in the preliminary exploration above is such that it is logical to specify

that, e.g., $q_2(6) \in \{0, \frac{1}{3}\}$ or, more precisely, that $q_2(6)$ is equal to 0 if $i = 0$ and equal to $\frac{1}{3}$ if $i = 1$. The latter statement is logically valid, but useless because we do not know whether $i = 0$ or $i = 1$. Simply stating that $q_2(6)$ is either 0 or $\frac{1}{3}$ is a possibility which is in line with some theories in logic. The interesting book by CLEAVE (1991) starts with Carnap's statement that 'in logic, the science of valid inference, there are no morals'. Indeed, if one refuses to choose between 0 and $\frac{1}{3}$ then no moral principles will be needed. The statistician, however, will accept the task of specifying a number $q_2(6)$, at least if a 'reasonable' compromise is possible.

The Bayesians adhere to the perspective that Player I chooses Strategy 1 with probability ρ. As a consequence $q_2(6)$ will be chosen equal to $\mathrm{P}(Y = 6|X = 2) = \rho/(2+\rho)$. The question now is *'which ρ?'*. In this problem, the choice $\rho = \frac{1}{2}$ cannot be defended on the basis of symmetry arguments. The Bayesian might continue by simply using $\rho = \frac{1}{2}$ because this 'is in the middle' of the interval $[0, 1]$ of possibilities, or he might continue by simply using ρ itself as the outcome of a uniformly distributed random variable. Such constructions are not necessarily reasonable. That is why we abandon the Bayesian approach and start from scratch in Section 1.3 with a 'classical' statistical approach where i is regarded as an unknown number. In Section 1.5 we accept the idea that i is the outcome of a random variable such that the possibility 1 has probability ρ. The discussion leads to the conclusion that this does not help much if one knows nothing about the true value of ρ.

The latter situation is not very realistic in the sense that if one knows that i is the outcome of a random variable then one will usually also have some information about ρ. Ignoring such prior information, computations to be discussed in the sequel lead to $\rho = .495$ (and, hence $q_2(6) = .198$) as the (Bayesian) solution, which corresponds to the minimax regret procedure. The minimax risk procedure is characterized by $\rho = .667$ (and $q_2(6) = .250$). A referee of ALBERS AND SCHAAFSMA (2001b) remarked that, without moral information the choice of principle cannot be discussed. We agree to a certain extent: many statisticians have experienced that Wald's minimax risk principle is often too conservative and that it may even lead to degenerate results such that, e.g., the observations are completely ignored (see, e.g., SCHAAFSMA, 1969). Such experiences suggest that the choice of ρ 'should be' closer to .495 than to .667.

The reason we pay much attention to this elementary and impractical example is that we are interested in the foundations of predictive inference, predictive distributional inference in particular. Our example is a very simple example of a problem where a predictive inference is required. That our problem is, in some sense, one step too simple, will be shown and dealt with in Section 1.7.

Note that the whole perspective of the example may change drastically if additional information is provided. If, e.g., Player II observes that Player I flips a coin before issuing the statement $x = 2$ then he should *not* conclude that $\rho = \frac{1}{2}$ and, hence $q_2(6) = \frac{1}{2}/(2 + \frac{1}{2}) = .20$ is appropriate. It is 'logical' to conclude that Player I has seen a six, and hence $q_2(6) = 1$, because in any other case it makes no sense to flip a coin.

With these preliminarities in mind, the reader will, hopefully, appreciate the following

discussions illustrating

1. Fisher's desire to create an inductive logic;
2. Popper's statement that *induction is a myth*;
3. The fact from life that *induction is a must*.

Epistẽmẽ (infallible knowledge about the universe) is beyond reach but we can do our best in providing 'approximations'. We are aware of the fact that, in some sense, we have done our best too much. The reader can restrict himself to Sections 1.3, 1.4, 1.5, and 1.9.

1.3 A Fisher-Neyman-Pearson-Wald approach

Fisher proposed methods of inference while Neyman and Pearson, and especially Wald, tried to make comparative analyses of such methods in the hope that some method comes out best (from certain perspectives). The first and very essential step in this approach is to shift the attention from the concrete situation with a given statement x (either 1,2, or 3) to that where probabilities $q_\xi(\eta)$ have to be assigned for each value ξ a priori possible for x. It seems reasonable to restrict the attention to the class

$$\mathcal{D} = \left\{Q_{a,b} \mid \tfrac{1}{3} \le a \le \tfrac{1}{2},\ \ \tfrac{1}{2} \le b \le 1\right\}$$

of procedures $Q_{a,b}$ defined in Table 1.1. One of the arguments behind this restriction is that in a Bayesian context (Strategy 1 is chosen with probability ρ) the posterior probabilities $\mathrm{P}(Y = \eta | X = \xi)$ are of this kind, with $a = (2+\rho)^{-1}$ and $b = (2-\rho)^{-1}$. Note that in this section we do *not* make the assumption that Strategy i is the outcome of a random variable.

Simply choosing $q_2(2) = q_2(4) = q_2(6) = \frac{1}{3}$ and $q_3(3) = q_3(6) = \frac{1}{2}$ is a possibility, but not a clever one. It corresponds to $Q_{1/3,1/2}$ and is represented by the left-lower point A of the rectangle in the left graph of Figure 1.2. If Player I is known to act according to Strategy 1 then $Q_1 = Q_{1/3,1}$ is the procedure to choose. It is represented as the left-upper point of the rectangle. If he would choose Strategy 0 then $Q_0 = Q_{1/2,1/2}$ (the right-lower point) is 'logically valid' from the probabilistic viewpoint. As we are unaware of Player I's strategy and yet are forced to assign probabilities, a *compromise* will be needed. Averaging the parameters of Q_0 and Q_1

Table 1.1: Procedures $Q_{a,b}$

$q_\xi(\eta)$	$\eta=1$	$\eta=2$	$\eta=3$	$\eta=4$	$\eta=5$	$\eta=6$
$\xi=1$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0
$\xi=2$	0	a	0	a	0	$1-2a$
$\xi=3$	0	0	b	0	0	$1-b$

we obtain $Q_{5/12,3/4}$ (point E). This solution is not satisfactory from an intellectual viewpoint: it is like cutting a Gordian knot without examining it. To improve this situation, we adopt the perspective of the theory of statistical decision functions. In its usual form this theory tries to prescribe how statements should be made about true values of unknown parameters (here strategy number i). Our situation is different in the sense that a predictive statement is required, namely about the true value of some random variable Y. The distributional form of inference is more natural if such predictive inference is required than in the classical situation. On the other hand, it is more complicated to arrive at satisfactory results (unless a Bayesian approach is adopted). We, for example, do not know whether the fundamental Wald-Lehmann minimal-complete class theorem is valid in predictive inference at large. In our special example it will follow from the concrete analysis that the minimal complete class corresponds to the class of all Bayes procedures. We, however, doubt whether this holds in general, an extension in Section 1.7 did not provide an answer. In this respect it is interesting to quote Eaton (1999, p. 851), who claims

> 'The more stringent evaluation of predictive distributions using decision theoretic notions (minimaxity, admissibility, etcetera) has received very little attention in the literature. A few results can be found in Eaton (1982, 1992) but a body of work providing hard evidence - that is, i.e. theorems - that specific predictive distributions will perform well in particular situations is, in the main, lacking.'

Anyway, given the observation x, we have to choose a probability distribution $Q = Q(x)$ on $\{1, \ldots, 6\}$ with probabilities $q_x(1), \ldots, q_x(6)$ and think in terms of the loss $L(y, Q)$ to be incurred if the true value y is revealed. An important requirement is that the loss is *proper*: if y is the outcome of a random variable Y with its probabilities $p(\eta) = \mathrm{P}(Y = \eta)$ known, then L is said to be proper if

$$\mathbf{E}\, L(Y, Q) = \sum_{\eta=1}^{6} L(\eta, Q) p(\eta)$$

is minimum, as a function of Q, if the corresponding probabilities satisfy $q(\eta) = p(\eta)$. At this moment the elaborations are restricted to the logarithmic loss function

$$L(y, P(x)) = -\log(q_x(y)).$$

The properness of this loss function is an immediate consequence of the positiveness of the Kullback-Leibler information number because, if P denotes the true distribution of Y with $\mathrm{P}(Y = \eta) = p(\eta)$, and Q is any other distribution with $Q(\{\eta\}) = q(\eta)$, then

$$\mathbf{E}\, L(Y, Q) - \mathbf{E}\, L(Y, P) = \sum_{\eta=1}^{6} \log\left(\frac{p(\eta)}{q(\eta)}\right) p(\eta)$$

is positive if $P \neq Q$ (Good, 1952). Using this loss function we can determine the risk (expected loss) of the procedure $Q_{a,b}$. As the distribution of (X, Y) depends on

$\mathrm{P}_i(X=\xi, Y=\eta)$	$\eta=1$	$\eta=2$	$\eta=3$	$\eta=4$	$\eta=5$	$\eta=6$
$\xi=1$	$1/6$	0	0	0	$1/6$	0
$\xi=2$	0	$1/6$	0	$1/6$	0	$i/6$
$\xi=3$	0	0	$1/6$	0	0	$(1-i)/6$

Table 1.2: Distributions P_i

whether Player I chooses Strategy 0 (saying $x=3$ if $y=6$) or Strategy 1 (saying $x=2$), there are two distributions involved. They are given in Table 1.2.

We shall have to consider the risks (expected losses)

$$\begin{aligned} R(0,Q_{a,b}) &= -\tfrac{1}{6}\log\left(\left(\tfrac{1}{2}\right)^2 a^2 b(1-b)\right) \\ R(1,Q_{a,b}) &= -\tfrac{1}{6}\log\left(\left(\tfrac{1}{2}\right)^2 a^2 b(1-2a)\right), \end{aligned}$$

obtained from Tables 1.1 and 1.2. $R(0,Q_{a,b})$ is minimum if both a^2 and $b(1-b)$ $((a,b)\in[\tfrac{1}{3},\tfrac{1}{2}]\times[\tfrac{1}{2},1])$ are maximum, i.e. if $Q_0=Q_{1/2,1/2}$ is used. Similarly $R(1,Q_{a,b})$ is minimum if $Q_1=Q_{1/3,1}$ is used. The fact that the minimization of these risks as a function of a can be separated from the minimization as a function of b indicates that our example is not representative for predictive inference at large. The minimum risk achieved by using the best procedure is called the envelope risk. It depends on i and is given by

$$\begin{aligned} R^*(0) &= R(0,Q_0)=\log 2 \\ R^*(1) &= R(1,Q_1)=\tfrac{1}{3}\log 2+\tfrac{1}{2}\log 3. \end{aligned}$$

The risk of any procedure $Q_{a,b}$ is larger than the envelope risk. The difference is referred to as the regret or shortcoming. Using $S(i,Q_{a,b})$ as notation, we have

$$\begin{aligned} S(0,Q_{a,b}) &= R(0,Q_{a,b})-R^*(0)=-\tfrac{1}{6}\log\left(16a^2b(1-b)\right) \\ S(1,Q_{a,b}) &= R(1,Q_{a,b})-R^*(1)=-\tfrac{1}{6}\log\left(27a^2b(1-2a)\right) \end{aligned}$$

Note that $S(0,Q_{a,b})$ is a decreasing function of a and an increasing function of b if $(a,b)\in[\tfrac{1}{3},\tfrac{1}{2}]\times[\tfrac{1}{2},1]$. For $S(1,Q_{a,b})$ the situation is reversed. In our example, the minimal complete class of procedures corresponds to the class of Bayes procedures. This class can be obtained by minimizing the convex combination

$$\begin{aligned} &(1-\rho)S(0,Q_{a,b})+\rho S(1,Q_{a,b})=-\tfrac{1}{6}\left(2\log a+\rho\log(1-2a)+\right. \\ &\left.\log b+(1-\rho)\log(1-b)+(1-\rho)\log 2^4+\rho\log 3^3\right) \end{aligned}$$

of both shortcomings. With elementary calculus it can be seen that this linear combination is minimum as a function of a and b if $a=(2+\rho)^{-1}$ and $b=(2-\rho)^{-1}$.

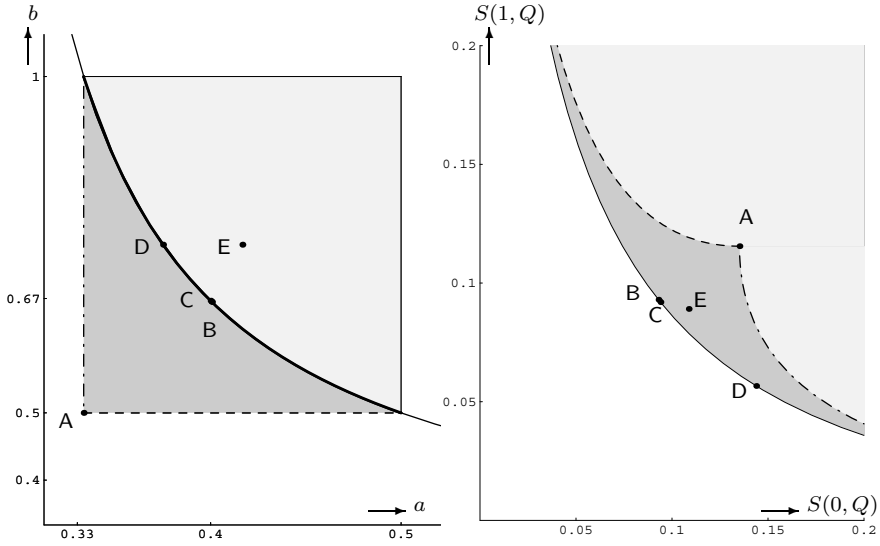


Figure 1.2: Visualization of the procedures in the (a, b)-plane (on the left) and the $(S(0, Q), S(1, Q))$-plane (right). The solid curves correspond to the procedures Q_ρ. Only the part inside the box in the figure to the left deserves consideration. The dashed lines correspond to procedures $Q_{a,1/2}$, the dot-dashed ones to $Q_{1/3,b}$. Explanation of the points A, B, C, D and E will be given in the text, especially at the end of Section 1.3.

This is expressed by $b = a/(4a - 1)$ and by the boldfaced curve in Figure 1.2 (left). Henceforth we use the notation

$$Q_\rho = Q_{\frac{1}{2+\rho}, \frac{1}{2-\rho}}$$

to denote the procedure of the form $Q_{a,b}$ that minimizes $(1-\rho)S(0, Q_{a,b})+\rho S(1, Q_{a,b})$. Note that $Q_0 = Q_{1/2,1/2}$ and $Q_1 = Q_{1/3,1}$ are as before. Looking for a compromise $Q_{a,b}$ between Q_0 and Q_1, one should not go (too far) beyond the boldfaced curve in Figure 1.2 (left) which characterizes the procedures which are 'admissible' in the sense of Wald's theory. For the general case we have

$$\begin{aligned} S(0, Q_\rho) &= -\tfrac{1}{6} \log \left(16 \frac{1-\rho}{(2+\rho)^2(2-\rho)^2} \right) \\ S(1, Q_\rho) &= -\tfrac{1}{6} \log \left(27 \frac{\rho}{(2+\rho)^3(2-\rho)} \right). \end{aligned}$$

The corresponding points constitute the left-lower bound of the regret set

$$\mathcal{S} = \left\{ (S(0, Q_{a,b}), S(1, Q_{a,b})) \mid \tfrac{1}{3} \le a \le \tfrac{1}{2}, \ \ \tfrac{1}{2} \le b \le 1 \right\}$$

which is, of course, the risk set when the origin is shifted to $(R^*(0), R^*(1))$. In this example the minimal complete class corresponds to the class of all admissible procedures as well as to the class of all Bayes procedures.

1.4 A detailed discussion of some specific procedures

Figure 1.2 provides visualizations of the parameters (a, b) of the procedures $Q_{a,b}$ (left) and of the corresponding points $S((0, Q_{a,b}), S(1, Q_{a,b}))$ (right). Table 1.3 gives details about the following specific points.

Point A

The naive procedure $Q_{1/3,1/2}$ has already been discussed at the beginning of the previous section. It corresponds to the idea that the probabilities $q_2(2) = q_2(4) = q_2(6) = \frac{1}{3}$ have to be assigned if $x = 2$ (because 2, 4 and 6 are equiprobable if Strategy 1 is chosen) and that the probabilities $q_3(3) = q_3(6) = \frac{1}{2}$ have to be assigned if $x = 3$ (because 3 and 6 are possible and equiprobable if Strategy 0 is used). The snake in the grass is that if $x = 2$, and Strategy 0 would have been chosen in case $y = 6$, then 2, 4 and 6 are not equiprobable at all, because $y = 6$ is impossible. A similar argument holds for $x = 3$, if Strategy 1 would have been chosen in case $y = 6$. It is very difficult for some probabilists to accept that the information *'the die is fair, the number of eyes is even'* does not imply that the outcomes 2, 4 and 6 are equiprobable. In the present context such a probabilist will be less unwilling to deviate from that what he regards as the foundations of Probability Theory than in, e.g., the quiz-master's problem. The reason is that we can present a precise analysis of the situation. The crux is, of course, that *the source of the information ('the number of eyes is even') has to be made part of the probabilistic model.* In the present Fisher-Neyman-Pearson-Wald approach this is done by referring to the true but unknown number i of the strategy which Player I would apply if he would have been confronted by $y = 6$. It follows from the discussion that the naive procedure $Q_{1/3,1/2}$ represented by A is not appropriate: a should be larger than $\frac{1}{3}$ and b larger than $\frac{1}{2}$. How much larger, that is the question.

Table 1.3: Specialization of some regret points

Procedure		Description	$S(0, Q)$	$S(1, Q)$
A	$Q_{1/3,1/2}$	naive conditional probabilities	.135	.116
B	$Q_{.495}$	minimax regret: $S(0, Q_\rho) = S(1, Q_\rho)$	.093	.093
C	$Q_{1/2}$	Bayes with uniform prior	.094	.092
D	$Q_{2/3}$	minimax risk: $R(0, Q_\rho) = R(1, Q_\rho)$	.144	.057
E	$Q_{5/12,3/4}$	naive compromise	.109	.089
	$Q_{2/5}$		.072	.119
	$Q_{.55}$		.107	.080

Point B
We are attracted by the idea to minimize the maximum regret (see Section 1.2). This can be achieved by looking for the Bayes procedure Q_ρ which satisfies $S(0, Q_\rho) = S(1, Q_\rho)$. The computations provide $\rho = .495$, etcetera and thus lead to procedure $Q_{.4007,.6648}$. See Table 1.3 and Figure 1.2 (point B) and notice that both regrets .093 are smaller than those of the naive procedure.

Point C
The Bayesian approach has a considerable appeal but it requires the choice of the prior probability ρ which, in the context of this section, is a fictitious construct (see Section 1.5 for a different possibility). We can ignore probabilistic terminology by simple stating that we want to minimize some weighted average of the risks, or, equivalently, of the regrets, $\frac{1}{2}S(0, Q_{a,b}) + \frac{1}{2}S(1, Q_{a,b})$. This is then achieved by the procedure $Q_{1/2}$ represented by the point C. Note that one coordinate of the corresponding regret point is larger, and one is smaller than that of B.

Point D
WALD (1947) was fascinated by the theory of games as presented in VON NEUMANN AND MORGENSTERN (1944) . This leads to minimizing the maximum risk procedure, which is obtained by equating the two risks. This provides $b = 2a$ and corresponds to point D.

Point E
At the beginning of this section we argued that a compromise will be needed and we suggested the algebraically natural candidate $Q_{5/12,3/4}$ is not satisfactory from an intellectual viewpoint. Indeed, $(\frac{5}{12}, \frac{3}{4})$ is beyond the boldfaced curve in Figure 1.2 (left) which corresponds to the admissible procedures. This generates the task to construct a procedure with both regrets decreased. Our first try was to solve $1/(2 + \rho) = \frac{5}{12}$ which provides $\rho = \frac{2}{5}$, the regret $S(1, Q_{2/5})$ being larger than that of the naive compromise. Solving $1/(2 - \rho) = \frac{3}{4}$ leads to the minimax risk procedure. Our final try was to take $\rho = .55$ (thus $Q_{.3922,.6897}$), which indeed has smaller regrets than the naive compromise.

A vexed issue at the end
The parameter ρ was used as a technical device to generate the minimal complete class $\{Q_\rho \mid 0 \leq \rho \leq 1\}$ of Bayes procedures. The procedure $Q_{1/3,1/2}$ was much too naive in the sense that it assumes that Strategy 1 has been chosen if $x = 2$ and Strategy 0 if $x = 3$. This does not necessarily correspond to the facts. It cannot be denied, however, that the outcome x contains some information with respect to Player I's choice of strategy. If we try to exploit this information by choosing ρ depending on x, then the procedure obtained may be reasonable but it will not be admissible in the sense of Wald, the admissibility requires that the procedure is in the class $\{Q_\rho \mid 0 \leq \rho \leq 1\}$.

1.5 What if Player I uses a randomized strategy?

As complete consensus can obviously not be achieved, additional knowledge would be welcome. In this section we restrict the attention to the idea that Player I chooses Strategy 1 with probability ρ, this 'physical' probability being either fully known (Situation 1) or fully unknown (Situation 2). In the end of this section, we will discuss the third possibility of partial knowledge.

Situation 1 may appear if we are involved in a two-person zero-sum game. Knowledge of the pay-off matrix will then not only affect the choice of ρ (if one accepts the minimax principle) but also the restriction to the class $\mathcal{D} = \{Q_{a,b} \mid \frac{1}{3} \leq a \leq \frac{1}{2}, \frac{1}{2} \leq b \leq 1\}$. This makes it clear that the 'solutions' presented in Section 1.4 are only reasonable if further information is absent.

Henceforth the attention is concentrated on Situation 2: the randomization probability ρ will then appear as the unknown true value of the parameter $\theta \in \Theta = [0,1]$, and the factual pair (x,y) is the outcome of a pair (X,Y) of random variables with distribution uniquely determined by ρ. As nothing is known about ρ, it is intuitively clear that the additional information is not worth much. We introduce random variables (X_θ, Y_θ) having the distribution P_θ which (X,Y) would have had, given $\rho = \theta$. If i is replaced by θ in Table 1.2 then one obtains a table of $\mathrm{P}(X_\theta = \xi, Y_\theta = \eta)$ values.

The risk $-\mathbf{E}\,\log(q_X(Y))$ depends on the true value ρ of θ. In general we have

$$\begin{aligned} R(\theta, Q_{a,b}) &= -\mathbf{E}\,\log(q_{X_\theta}(Y_\theta)) \\ &= -\tfrac{1}{6}(\log(2^{-2}a^2 b) - \theta\log(1-2a) - (1-\theta)\log(1-b)). \end{aligned}$$

Note that this risk is equal to $(1-\theta)R(0,Q_{a,b}) + \theta R(1,Q_{a,b})$ and, hence, is given by $R(0,Q_{a,b})$ and $R(1,Q_{a,b})$ in the end points.

For fixed θ, the risk $R(\theta, Q_{a,b})$ is minimum if $(a,b) = (1/(2+\theta), 1/(2-\theta))$. The procedure Q_θ thus obtained is Bayes with respect to all prior distributions τ on $[0,1]$ which have θ as their expectation. The envelope risk is

$$\begin{aligned} R^*(\theta) &= R(\theta, Q_\theta) \\ &= (1-\theta)R(0,Q_\theta) + \theta R(1,Q_\theta) \end{aligned}$$

where $R(0,Q_\theta)$ is obtained from $R(0,Q_{a,b})$ by substituting

$$(a,b) = \left(\frac{1}{2+\theta}, \frac{1}{2-\theta}\right)$$

and $R(1,Q_\theta)$ follows from $R(1,Q_{a,b})$. The envelope risk displayed in Figure 1.3, is such that $R^*(0)$ and $R^*(1)$ correspond to $R^*(0) = \log 2$ and $R^*(1) = \frac{1}{3}\log 2 + \frac{1}{2}\log 3$. The maximum is reached for $\theta = \frac{2}{3}$; the minimax risk procedure is $Q_{2/3}$. The shortcoming $S(\theta, Q_{a,b}) = R(\theta, Q_{a,b}) - R^*(\theta)$ is a convex function of θ because $R^*(\theta)$ is concave and $R(\theta, Q_{a,b})$ is linear in θ. As a consequence the minimax regret and the

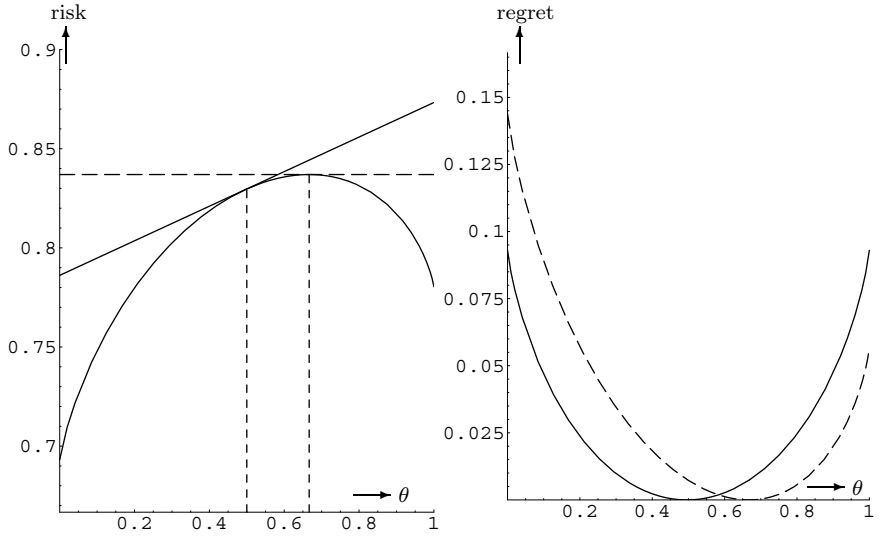


Figure 1.3: The randomized strategy game. On the left: envelope risk $R^*(\theta)$ (curve), minimax risk procedure (dashed line) and minimax regret procedure (solid line). On the right, the same procedures, but now with the regret along the vertical axis.

minimax procedure are exactly the same as in Section 1.4. The area under the regret function is minimized by using $Q_{1/2} = Q_{2/5,2/3}$ which, thus, is the minimal average regret procedure. The main difference with Figure 1.2 is that shortcoming functions are now visualized as functions of θ.

The elaboration about randomized strategies suggests that nothing is gained if we know that Player I chooses Strategy 1 with some completely unknown probability ρ. One might argue that this state of ignorance changes as soon as the outcome x is available. The underlying random variable X assumes the values $1, 2$, or 3 with probabilities $\frac{1}{3}$, $\frac{1}{3} + \frac{1}{6}\rho$, $\frac{1}{3} - \frac{1}{6}\rho$ respectively. Shouldn't this information be used to replace the a priori choice of $\rho = \frac{1}{2}$ and $a = \frac{2}{5}$, $b = \frac{2}{3}$ by an a posteriori choice of $\rho = \frac{1}{2}$, $\frac{1}{2} + \epsilon$, $\frac{1}{2} - \epsilon$ if $x = 1$, 2, 3? In practice, Bayesian statisticians adapt their prior if it is in conflict with actual data. In the present example it is easy to see that this approach leads to a procedure that is 'inadmissible' from a theoretical viewpoint: the x-dependent choice of ρ suggested corresponds to a $Q_{a,b}$ with

$$(a, b) = \left(\frac{1}{\frac{5}{2} + \epsilon}, \frac{1}{\frac{3}{2} + \epsilon} \right)$$

not in the arc of admissible procedures.

In practice, the additional information will not be of the form of the two extremes

$q_\xi(\eta)$	$\eta=1$	$\eta=2$	$\eta=3$	$\eta=4$	$\eta=5$	$\eta=6$
$\xi=1$	c	0	0	0	$(1-c)$	0
$\xi=2$	0	a	0	d	0	$(1-a-d)$
$\xi=3$	0	0	b	0	0	$(1-b)$

Table 1.4: Procedures $Q_{a,b,c,d}$

('ρ fully known' or 'ρ completely unknown') suggested in the beginning of Section 1.5. A confidence interval or distributional inference about ρ may be available. This may affect the conclusion that $Q_{a,b}$ with $(a,b) \approx (0.40, 0.67)$ is the rule to choose. That a compromise solution may be affected if additional information is provided, is completely natural though it illustrates the hazards involved in any discussion of the types presented.

1.6 Adapting the theory to alternative loss functions

Our theory started from the class $\mathcal{D}$ of procedures and concentrated the attention on risks and regrets based on logarithmic loss. It resulted in the opinion that (a,b) should be close to the arc

$$\left\{(a,b) \mid (a-\tfrac{1}{4})(b-\tfrac{1}{4}) = \tfrac{1}{16}\right\}$$

because these points generate the class $\{Q_\rho \mid 0 \le \rho \le 1\}$ of all Bayes procedures which corresponds to the minimal complete class. An extensive discussion suggested that the procedure to be chosen should be of the form Q_ρ where ρ is not too much different from $\frac{1}{2}$. The agreement among the probabilities actually assigned is such that none of these probabilities is completely compelling in the cases $x=2$ or $x=3$. They, however, are not unreasonable because the agreement between, e.g., $q_2(6) = .198$ (minimax average risk or regret) and $q_2(6) = .250$ (minimax risk) is quite satisfactory.

The theory was based on the restriction to the class $\mathcal{D}$ of procedures $Q_{a,b}$ (see the beginning of Section 1.2). Instead of $\mathcal{D}$ one might consider the more general class of procedures of the form $Q_{a,b,c,d}$ defined in Table 1.4. If one uses a *proper* loss function (see the end of this section) this extension is useless because the class $\mathcal{D}$ is essentially complete, in fact $\{Q_\rho : 0 \le \rho \le 1\} \subset \mathcal{D}$ is already (minimal) complete.

The reason for this is as follows. To derive the Bayes procedure w.r.t. some τ on $\Theta = [0,1]$, let us invoke a probabilistic model where (x,y) appears as the outcome of a pair (X_τ, Y_τ) of random variables. The distribution of this pair corresponds to the marginal distribution of $(\tilde{T}, \tilde{X}, \tilde{Y})$ where $\mathcal{L}(\tilde{T}) = \tau$ and $\mathcal{L}((\tilde{X}, \tilde{Y}) \mid \tilde{T} = \theta) = P_\theta =$

$\mathcal{L}\,(X_\theta, Y_\theta)$. The Bayes risk is

$$\begin{aligned} r(\tau, Q_{a,b,c,d}) &= \int R(\theta, Q_{a,b,c,d})\,\mathrm{d}\tau(\theta) \\ &= \int \mathbf{E}\,L(Y_\theta, Q_{a,b,c,d}(X_\theta))\,\mathrm{d}\tau(\theta) \\ &= \mathbf{E}\,L(\tilde{Y}, Q_{a,b,c,d}(\tilde{X})) \\ &= \mathbf{E}\,\left(\mathbf{E}\,L(\tilde{Y}, Q_{a,b,c,d}(\tilde{X}))|\tilde{X}\right) \end{aligned}$$

The properness of L implies that we have to use the procedure with parameters a, b, c, and d, such that

$$Q_{a,b,c,d}(x) = \mathcal{L}\,(\tilde{Y}|\tilde{X} = x)$$

But the marginal distribution of $(\tilde{X}, \tilde{Y})$ has already been described. It is that of $(X_\tau, Y_\tau) = (X_\theta, Y_\theta)$ where $\theta = \int_0^1 u\,\mathrm{d}\tau(u)$. Hence

$$\mathcal{L}\,(\tilde{Y}|\tilde{X} = x) = Q_{\frac{1}{2+\theta}, \frac{1}{2-\theta}}(x).$$

Another 'basis' of our theory was the usage of the logarithmic loss function. The question arises whether the agreement will be affected if the loss function is replaced by another (proper) one. Let us consider three proper loss functions: logarithmic, BRIER (1950) and EPSTEIN (1969) loss, respectively

$$\begin{aligned} L_{\log}(y, Q(x)) &= -\log q_x(y), \\ L_{\mathrm{B}}(y, Q(x)) &= (1 - q_x(y))^2 + \sum_{\eta \neq y} (q_x(\eta))^2, \\ L_{\mathrm{E}}(y, Q(x)) &= \sum_{\eta=1}^{6} \left(\mathbf{1}_{\{y,\ldots,6\}}(\eta) - \sum_{\nu=1}^{\eta} q_x(\nu)\right)^2. \end{aligned}$$

For the seven possible combinations of x and y, the losses incorporated are as in Table 1.5 From these losses, we can compute the risk functions $R.(0, Q_{a,b})$ and $R.(1, Q_{a,b})$ for Strategies 0 and 1 respectively, as well as the risk functions $R.(\theta, Q_{a,b})$ corresponding to the randomized strategy.

$$\begin{aligned} R_{\log}(\theta, Q_{a,b}) &= -\tfrac{1}{6}\left(\log(2^{-2}a^2b) - \theta\log(1-2a) - (1-\theta)\log(1-b)\right) \\ R_{\mathrm{B}}(\theta, Q_{a,b}) &= \tfrac{1}{6}\left(6(2+\theta)a^2 - 12a + 2(2-\theta)b^2 - 4b + 7\right) \\ R_{\mathrm{E}}(\theta, Q_{a,b}) &= \tfrac{1}{6}\left(10(2+\theta)a^2 - 20a + 3(2-\theta)b^2 - 6b + 11\right) \end{aligned}$$

Due to the linear relation $\mathrm{P}_\theta = (1-\theta)\mathrm{P}_0 + \theta\mathrm{P}_1$ and properness of the loss functions, the different envelope risks are obtained when the same procedures are used, as was

(x, y)	logarithmic	Brier	Epstein
$(1,1)$	$\log(2)$	0.5	1
$(1,5)$	$\log(2)$	0.5	1
$(2,2)$	$-\log(a)$	$6a^2-6a+2$	$10a^2-12a+4$
$(2,4)$	$-\log(a)$	$6a^2-6a+2$	$10a^2-8a+2$
$(2,6)$	$-\log(1-2a)$	$6a^2$	$10a^2$
$(3,3)$	$-\log(b)$	$2(1-b)^2$	$3(1-b)^2$
$(3,6)$	$-\log(1-b)$	$2b^2$	$3b^2$

Table 1.5: 7×3*-matrix of possible losses*

already noted before. These procedures are of the form $Q_\theta = Q_{(2+\theta)^{-1},(2-\theta)^{-1}}$ with endpoints $Q_{\frac{1}{2},\frac{1}{2}}$ and $Q_{\frac{1}{3},1}$ if $\theta = 0, 1$. The envelope risk functions are

$$\begin{aligned}
R^*_{\log}(\theta, Q_\theta) &= -\tfrac{1}{6}(-\log 2^2 - ((2+\theta)\log(2+\theta) + (2-\theta)\log(2-\theta)) \\
&\quad + (\theta\log\theta + (1-\theta)\log(1-\theta))) \\
R^*_{\mathrm{B}}(\theta, Q_\theta) &= \tfrac{1}{6}\left(7 + \tfrac{1}{2+\theta} + \tfrac{5}{2-\theta} - \tfrac{28}{(2+\theta)(2-\theta)}\right) \\
R^*_{\mathrm{E}}(\theta, Q_\theta) &= \tfrac{1}{6}\left(11 + \tfrac{1}{2+\theta} + \tfrac{8}{2-\theta} - \tfrac{44}{(2+\theta)(2-\theta)}\right)
\end{aligned}$$

The shortcomings, or regrets, are obtained by subtracting the envelope risks from the risks. Thus, $S(\theta, Q_\rho) = R(\theta, Q_\rho) - R(\theta, Q_\theta)$ for $\theta, \rho \in [0,1]$. The regret functions for our three different loss functions are

$$\begin{aligned}
S_{\log}(\theta, Q_\rho) &= -\tfrac{1}{6}(-\theta\log\tfrac{\rho}{(2+\theta)^3(2-\theta)} + (1-\theta)\log\tfrac{1-\rho}{(2+\theta)^2(2-\theta)^2} + \theta\log\theta + \\
&\quad + (1-\theta)\log(1-\theta) - (2-\theta)\log(2-\theta) - (2+\theta)\log(2+\theta)) \\
S_{\mathrm{B}}(\theta, Q_\rho) &= \tfrac{(\rho-\theta)^2/6}{(2+\rho)^2(2-\rho)^2(2+\theta)(2-\theta)}((8-2\theta)\rho^2 + (-16+2\theta)\rho + (16-4\theta)) \\
S_{\mathrm{E}}(\theta, Q_\rho) &= \tfrac{(\rho-\theta)^2/6}{(2-\rho)^2(2+\rho)^2(2-\theta)(2+\theta)}((26-7\theta)\rho^2 + \\
&\quad (-56+52\theta)\rho + (104-28\theta))
\end{aligned}$$

These formulas are sufficient material for a graphical analysis of the procedures following from the three different loss functions. We concentrate the attention on the motivation behind the five points $\mathsf{A}, \ldots, \mathsf{E}$ considered before. In Table 1.6, for each of these points the corresponding procedure Q and regret points $S(0, Q), S(1, Q)$ are given.

A graphical representation can be seen in Figure 1.4 where, from above to below, two graphs for logarithmic, Brier and Epstein loss are displayed. The five points, along with the curve $\{Q_\rho | \rho \in [0,1]\}$ are plotted. The three left graphs are plotted in the (a, b)-plane, where the rectangle bounds the procedures that deserve consideration.

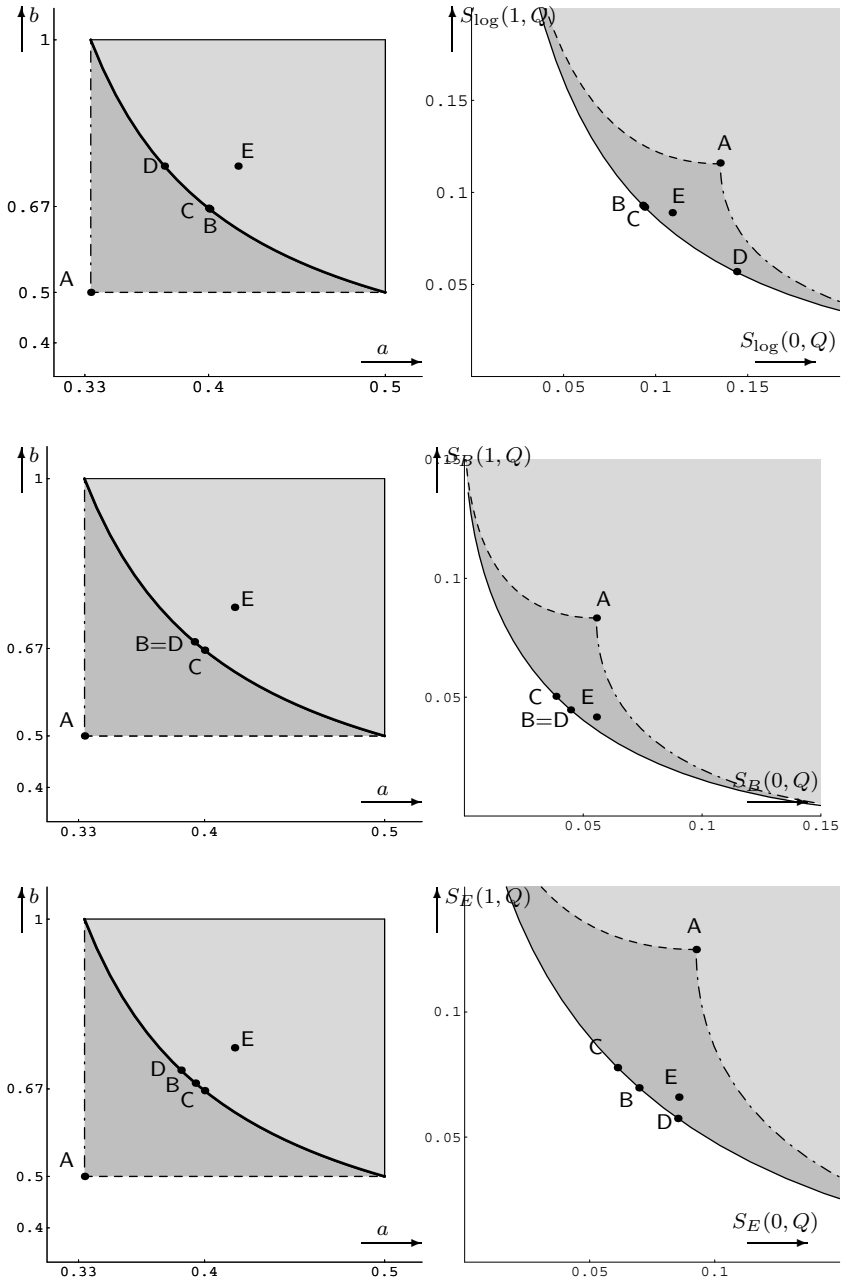


Figure 1.4: Visualization of logarithmic, Brier and Epstein scoring rules.

Table 1.6: Overview of regret points

	Logarithmic			Brier			Epstein		
	Q	$\theta=0$	$\theta=1$	Q	$\theta=0$	$\theta=1$	Q	$\theta=0$	$\theta=1$
A	$Q_{\frac{1}{3},\frac{1}{2}}$	.135	.116	$Q_{\frac{1}{3},\frac{1}{2}}$	.0556	.0833	$Q_{\frac{1}{3},\frac{1}{2}}$	.0926	.1250
B	$Q_{.496}$	.093	.093	$Q_{.536}$	.0447	.0447	$Q_{.532}$	.0697	.0697
C	$Q_{\frac{1}{2}}$	.094	.092	$Q_{\frac{1}{2}}$	.0385	.0504	$Q_{\frac{1}{2}}$	.0611	.0778
D	$Q_{\frac{2}{3}}$	.144	.057	$Q_{.536}$	.0447	.0447	$Q_{.584}$	.0852	.0575
E	$Q_{\frac{5}{12},\frac{3}{4}}$	.109	.089	$Q_{\frac{5}{12},\frac{3}{4}}$	.0556	.0417	$Q_{\frac{5}{12},\frac{3}{4}}$	.0856	.0660

The three right graphs present the $(S(0,Q_{a,b}),S(1,Q_{a,b}))$-points, part of the rectangle is mapped and folded.

In Figure 1.5 we have made a display of the risk for the case of logarithmic loss. Figure 1.6 provides a similar display, this time of the regret. For Brier and Epstein loss similar visualizations can be made. These figures display a similar behaviour. The left part of Figure 1.5 shows us the envelope risk $R^*_{\log}(\theta)$ (curve), the minimax risk procedure $R_{\log}(\theta,Q_{2/3})$ (dashed line) and the minimax regret procedure $R_{\log}(\theta,Q_{.4957})$ (solid line). In left of Figure 1.6 we see the same procedures. In the right part of Figure 1.5 the riskplane $R_{\log}(\theta,Q_\rho)$ $([\theta,\rho]\in(0,1)^2)$ is displayed. Notice that for fixed ρ, the risk is linear in θ. The three solid lines correspond to the envelope risk ($\theta=\rho$) and minimax risk ($\rho=\frac{2}{3}$) and minimax regret ($\rho=0.4957$) procedures. The right part of Figure 1.6 shows a similar display, now in the regretplane. The figures on the left of Figures 1.5 and 1.6 can thus be described as a two-dimensional display of the right figures ('ignoring' the value of ρ).

Conclusions

The procedures corresponding to A, C and E are independent of the choice of loss function (as long as this loss function is proper), as was to be expected. Of course, the corresponding risks and regrets do shift. Note that in the case of Brier-loss, the values of the minimax risk and minimax regret procedures coincide because, in this case, the envelope risks in 0 and 1 are equal. The position of the minimax risk and minimax regret procedures are different, but remain in the form Q_θ. The differences are also small, all procedures suggest a strategy Q_θ with θ close to or in $[\frac{1}{2},\frac{2}{3}]$. If the loss function is proper, then the class $\{Q_\rho|\rho:0\le\rho\le1\}$, of Bayes rules (they are also Somewhere Minimal Risk), will not be affected, $Q_{1/2}$ will minimize the average risk or regret (if a randomized strategy is considered, integration should be with respect to Lebesgue measure). The position of the ρ values corresponding to the minimax regret or the minimax risk procedures will become somewhat different.

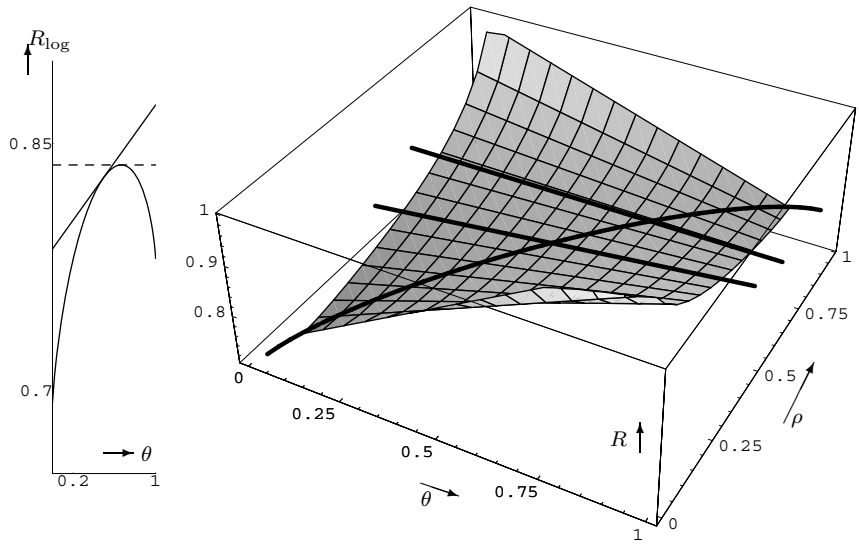


Figure 1.5: Risks for logarithmic loss.

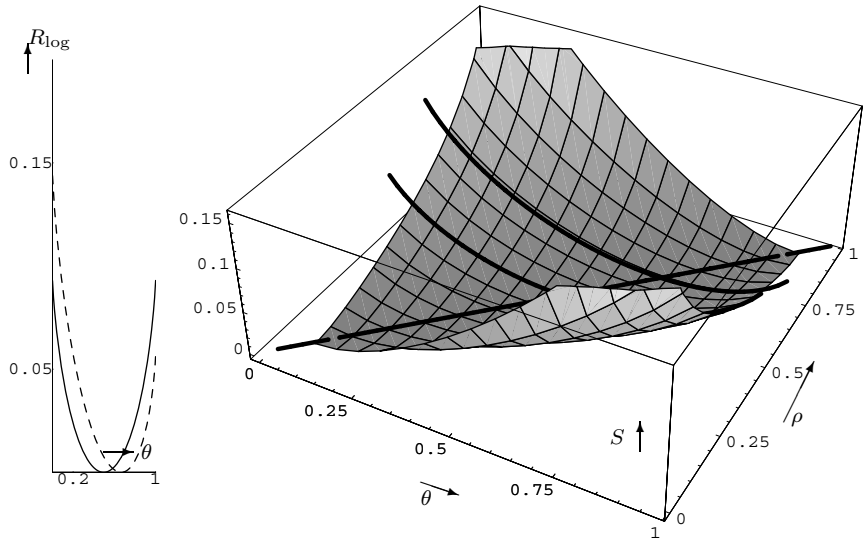


Figure 1.6: Regrets for logarithmic loss.

1.7 Extension

This and the next section are concerned with generalizations to and extensions of the example discussed so far. The reader can continue directly with Section 1.9 without loss of understanding of the main ideas of this chapter. The example discussed so far is in some sense paradigmic for statistics at large. In another sense it is not sufficiently general to be called paradigmic because the linearity

$$\mathrm{P}_\theta = (1-\theta)\mathrm{P}_0 + \theta\mathrm{P}_1$$

of the model has induced the very special property that all admissible procedures in Section 1.5 (Situation 2) are also 'Somewhere Minimum Risk'. Let us therefore now discuss a new example which is, at first sight, one step less regular.

A new example
Just as in the original example, a fair 'die' (now a regular simplex) has been rolled and Player I has to make a statement x about the true number of eyes y according to the following 'rules': when $y = 1$, Player I has to make statement $x = 1$, when $y = 2$ the statement must be $x = 2$, and if $y = 3$ or $y = 4$ Player I is allowed to choose between $x = 1$ and $x = 2$. Treatment of this example mimics that of the original one, though it looks a bit more complex. That, on the other hand, it is also more general than the original example will become clear during the analysis.

Now there are four possibilities for Player I to specify his choice of x, depending on the value y observed. These possibilities can be indicated as $\{(i,j);\ i,j \in \{0,1\}\}$ where $i = 1$ if Player I chooses $x = 1$ after seeing $y = 3$ (and $i = 2$ when $x = 2$ is chosen in this case) while $j = 1$ if Player I chooses $x = 1$ after seeing $y = 4$ (and $j = 0$ if he chooses $x = 2$). Our corresponding strategies will be indicated as $Q_{(i,j)}$ or Q_{2i+j}. Here Q_{2i+j} prescribes what we should infer if we assume that Player I chooses the possibility (i, j).

We can restrict, for trivial reasons, to procedures $Q_{a,b,c,d}$ as displayed in Table 1.7. Note that $Q_0 = Q_{0,0,1/3,1/3}$, $Q_1 = Q_{0,1/2,1/2,0}$, $Q_2 = Q_{1/2,0,0,1/2}$, and $Q_3 = Q_{1/3,1/3,0,0}$. Table 1.8, with probabilities $\mathrm{P}_{2i+j}(X = \xi, Y = \eta)$ is the analogue of Table 1.2.

The risks involved in the four possibilities (pure strategies (i, j) or $2i + j$ of Player I)

Table 1.7: Procedures $Q_{a,b,c,d}$

$q_\xi(\eta)$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$
$\xi = 1$	$1 - a - b$	0	a	b
$\xi = 2$	0	$1 - c - d$	c	d

$\mathrm{P}_{2i+j}(X=\xi, Y=\eta)$	$\eta=1$	$\eta=2$	$\eta=3$	$\eta=4$
$\xi=1$	$\frac{1}{4}$	0	$\frac{i}{4}$	$\frac{j}{4}$
$\xi=2$	0	$\frac{1}{4}$	$\frac{1-i}{4}$	$\frac{1-j}{4}$

Table 1.8: Procedures P_{2i+j}

are

$$\begin{aligned}
R(0, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)cd\right)\\
R(1, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)cb\right)\\
R(2, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)ad\right)\\
R(3, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)ab\right),
\end{aligned}$$

the envelope risks $R^*(k)$ are $R^*(0) = R^*(3) = \frac{3}{4}\log 3$ and $R^*(1) = R^*(2) = \log 2$ and the corresponding regrets (shortcomings)

$$\begin{aligned}
S(0; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(3^3(1-a-b)(1-c-d)cd\right)\\
S(1; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(2^4(1-a-b)(1-c-d)cb\right)\\
S(2; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(2^4(1-a-b)(1-c-d)ad\right)\\
S(3; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(3^3(1-a-b)(1-c-d)ab\right).
\end{aligned}$$

To construct the class $\mathcal{B} = \{Q_\rho \mid \rho \in S_3\}$ of all Bayes procedures, let ρ_1, ρ_2, and ρ_3 be a priori probabilities for possibilities 1, 2, and 3 (and hence $1-\rho_1-\rho_2-\rho_3$ for possibility 0). (It is trivial that all three $\rho_i \geq 0$ and that $\rho_1+\rho_2+\rho_3 \leq 1$.) The class of Bayes procedures consists (exactly) of the procedures that minimize the Bayes risks

$$(1-\rho_1-\rho_2-\rho_3)S(0) + \rho_1 S(1) + \rho_2 S(2) + \rho_3 S(3).$$

Elementary calculus provides the solutions $Q_{a,b,c,d}$ with

$$\begin{aligned}
a &= \tfrac{\rho_2+\rho_3}{1+\rho_1+\rho_2+2\rho_3}\\
b &= \tfrac{\rho_1+\rho_3}{1+\rho_1+\rho_2+2\rho_3}\\
c &= \tfrac{1-(\rho_2+\rho_3)}{3-(\rho_1+\rho_2+2\rho_3)}\\
d &= \tfrac{1-(\rho_1+\rho_3)}{3-(\rho_1+\rho_2+2\rho_3)}.
\end{aligned}$$

Here we used the average regret as Bayes risk. We could equally well have used the average risk, this would, obviously, have provided the same Bayes rule. The issues to be considered next are to define the *minimax risk* and *minimax regret* procedure.

The symmetry of $R(\cdot, Q_{a,b,c,d})$ suggests that the maximum risk is minimized if we can equalize an Bayes rule. This is obviously achieved if

$$\rho_1 = \rho_2 = \rho_3 = \tfrac{1}{4},$$

because the corresponding Bayes rule is such that

$$a = b = c = d = \tfrac{1}{4}.$$

Quite interesting, the minimax regret procedure is the same $Q_{1/4,1/4,1/4,1/4}$, but now with $\rho_1 = \rho_2 = \frac{1}{2}$ and,hence, $\rho_3 = 0$. If we plug in the above representations in ρ_1, ρ_2, ρ_3 for a, b, c, and d in the $S(\cdot, Q_{a,b,c,d})$, we obtain after observing symmetry in ρ_1 and ρ_2 (and hence $\rho_1 = \rho_2$) and $\rho_0 = 1 - \rho_1 - \rho_2 - \rho_3$ and ρ_3 (hence $\rho_0 = \rho_3$) that

$$(1 - \rho_1 - \rho_2 - \rho_3)S(0) + \rho_1 S(1) + \rho_2 S(2) + \rho_3 S(3)$$

$$= -\tfrac{1}{4}\left(2\rho_1 \log 2^4 + 2(\tfrac{1}{2} - \rho_1)\log 3^3 - 6\log 2\right)$$

and, again,

$$a = b = c = d = \tfrac{1}{4}.$$

This coincidence of the minimax risk and minimax regret rule and the symmetry in the solutions shows that this second example is also not sufficiently general to be called paradigmic. Let us therefore consider a completely new example, in a different context.

1.8 An classroom example with Bernoulli trials

This chapter, so far, illustrates that straightforward conditioning to incorporate empirical evidence can be misleading. This phenomenon is well-known from other problems, such as the quiz-master paradox and the prisoner's dilemma (see Section 1.9 for formulations). The source of the information should be formalized and made part of the probabilistic model, which will then become 'statistical' in the sense that the unknown true value of a parameter appears. HEMELRIJK (1978) states that 'the question of how to choose a statistical mathematical model has led to considerable confusion and controversy, and still does. Mathematical statisticians wisely save their skins by using the axiomatic approach, leaving the controversy to others and the confusion to the users of their theory. For axioms, however useful, say nothing about their application.'

Similar issues are involved elsewhere though they often go unnoticed. The following problem provides an example because neither information is provided about the player's set of alternatives nor about the rules he has to obey: 'A bridge player announces that his hand (of 13 cards) contains (i) an ace (that is, at least 1 ace), (ii) the

ace of hearts. What is the probability that it will contain another ace?' (PARZEN, 1960, p. 75).

From the statistician's viewpoint as well as from that of the epistemologist, this problem is 'ill-posed'. There, indeed, is a difference between the attitudes of the analyst (mathematician, probabilist) who proceeds on the basis of an elegant rationalization, and that of the statistician who has to start from the observation. In mathematical statistics we try to combine these rationalist and empiricist perspectives. In Sections 1.1 up to 1.6 we had considerable difficulty in obtaining consensus. In Section 1.7 the symmetry of the situation enforced some type of consensus. We shall now discuss a different elementary example, where, in the end, (approximate) consensus will be reached.

Problem
$X_1, X_2, \ldots$ are outcomes of independent Bernoulli trials, each with unknown success-rate p. The theoretical possibilities θ for p constitute the parameter space $\Theta = [0, 1]$. The outcome $x = (x_1, \ldots, x_n)$ observed has to be used to make an inference about the outcome y of $Y = X_n X_{n+1}$.

Exploration
Note that $y \in \mathcal{Y} = \{0, 1\}$. The 'inference' about y required can have the form of a distributional inference

$$Q(x) = (1 - \alpha(x))\epsilon_0 + \alpha(x)\epsilon_1.$$

Amongst other proper loss functions, the logarithmic loss function $L_{\log}(y, Q) = -\log q_x(y)$ and the Epstein scoring rule

$$L_{\mathrm{E}}(y, Q(x)) = \left(\mathbf{1}_{\{0\}}(y) - 1 + p\mathbf{1}_{\{1\}}(x)\right)^2 = \tfrac{1}{2} L_{\mathrm{B}}(y, Q(x))$$

can be used. Note that, in this case, the relationship between Epstein and Brier loss is such that $L_{\mathrm{E}}(y, Q(x)) = \frac{1}{2} L_{\mathrm{B}}(y, Q(x))$. The 'inference' $(x_n, y) = (0, 1)$ yields maximum loss (infinite loss if the logarithmic loss function is used) for all proper scoring rules, and $(x_n, y) = (0, 0)$ yields zero loss, because inferring $y = 0$ is obviously the only right choice after observing $x_n = 0$.

The inference about y required can also be identified as a point estimate (prediction) of y. Squared error loss is then attractive because $\mathbf{E}\,(Y - c)^2$ is minimum if

$$\begin{aligned} d(x_1, \ldots, x_n) &= \mathbf{E}\,(X_n X_{n+1} | X_1 = x_1, \ldots, X_n = x_n) \\ &= x_n \mathbf{E}\,(X_{n+1} | X_1 = x_1, \ldots, X_n = x_n) \\ &= x_n \mathbf{E}\,(X_{n+1}) \\ &= p x_n. \end{aligned}$$

Unfortunately, we do not know the true value p of the parameter θ. It is intuitively clear that the ideal 'solution' $d(x_1, \ldots, x_n) = p x_n$ should be replaced by a practical

one where p is replaced by some estimate. The theoretician, however, might argue that $\mathbf{E}\, Y = p^2$ and that, hence, it is reasonable to use the best unbiased estimator[2]

$$d_1(x) = \frac{s_n(s_n-1)}{n(n-1)}$$

of p^2. Here and elsewhere $s_m = \sum_{i=0}^{m} x_i$ is used for notation. After some computations it is clear that the prediction risk is given by $\mathbf{E}\,(d_1(X_1,\ldots,X_n)-Y)^2 =$

$$\begin{aligned} &= (1-p)\mathbf{E}\,((d_1-Y)^2|X_n=0) + p\mathbf{E}\,((d_1-Y)^2|X_n=1) \\ &= \frac{(1-p)p^2(4p-2n(5p-2)-n^3(-2-p+p^2)+n^2(p^2+3p-4))}{(n-1)n^2}. \end{aligned}$$

An obvious drawback of d_1 is that it can happen that $d_1(x) > 0$ whilst $x_n = 0$ and therefore $Y = 0$ is certain. It seems more reasonable to start from the procedure $d_p(x) = px_n$ which would be used if p were known, and to replace p by an estimate, e.g.,

$$\begin{array}{lcll} \hat{p} &=& s_n/n & \\ \hat{p} &=& s_{n-1}/(n-1) & \\ \hat{p} &=& (s_n+1)/(n+2) & \text{(Laplace)} \\ \hat{p} &=& (s_n+\sqrt{n}/2)/(n+\sqrt{n}) & \text{(Bernstein)} \end{array}$$

or, more generally,

$$\hat{p} = as_{n-1} + bx_n + c$$

where a, b and c have to be chosen such that

$$d_{a,b,c}(x) = x_n(as_{n-1} + bx_n + c)$$

displays good behaviour. Since $d_{a,b,c}(x) = ax_n s_{n-1} + (b+c)x_n$, it is obvious that we can generalize even more, by focussing on

$$d_{a,b}(x) = x_n(as_{n-1} + b).$$

Let us now for a moment look at the special case where $a = 1/(n-1)$ and $b = 0$. Here the actual risk is equal to $\mathbf{E}\, X_n^2((n-1)^{-1}S_{n-1} - X_{n+1})^2 = p^2(1-p)\frac{n}{n-1}$ and, hence, the risk function and regret function are

$$\begin{aligned} R(\theta, d_{1/(n-1),0}(x)) &= \frac{n}{n-1}\theta^2(1-\theta) \\ S(\theta, d_{1/(n-1),0}(x)) &= \frac{1}{n-1}\theta^2(1-\theta) \end{aligned}$$

[2]To construct the estimator d_1 start from X_1X_2. Because of independence, it is trivial that this is an unbiased estimator. Then use the fact that S_n is a sufficient statistic, and use Rao-Blackwell and Lehmann-Scheffé to find $d_1(x) = \mathbf{E}\,(X_1X_2|S_n)$.

and the latter one converges nicely to 0.

Bayesian methods
Another important special case is that where we try to minimize some integral

$$\int_0^1 R(\theta, d(x))\omega(\theta)\,\mathrm{d}\theta$$

by using Bayesian methods. If we use a Beta weight function, then the computations become feasible by invoking a Bayesian context with a probability distribution Q determined by $T \sim \text{Beta}(f,g)$ and $\tilde{X}_1, \ldots, \tilde{X}_{n+1}|(T=\theta)$ i.i.d. Bernoulli(θ). Interprete

$$\int_0^1 R(\theta, d)\frac{\theta^{f-1}(1-\theta)^{g-1}}{\beta(f,g)}\,\mathrm{d}\theta = \mathbf{E}\,\mathbf{E}\,(d(\tilde{X}_1,\ldots,\tilde{X}_n) - \tilde{X}_n\tilde{X}_{n+1})^2|T,$$

this can be treated as usual by conditioning with respect to the observation, i.e. by writing the total expectation as

$$\mathbf{E}\,\mathbf{E}\,((d(\tilde{X}_1,\ldots,\tilde{X}_n) - \tilde{X}_n\tilde{X}_{n+1})^2|\tilde{X}_1,\ldots,X_n)$$

and choosing d such that

$$\mathbf{E}\,((d(\tilde{x}_1,\ldots,\tilde{x}_n) - \tilde{x}_n\tilde{X}_{n+1})^2|\tilde{X}_1,\ldots,X_n = x_1,\ldots,x_n)$$

is minimum, i.e.

$$\begin{aligned}
d(x) &= x_n\mathbf{E}\,(\tilde{X}_{n+1}|\tilde{X}_1,\ldots,X_n = x_1,\ldots,x_n)\\
&= x_n\mathrm{P}_Q(\tilde{X}_{n+1}=1|\tilde{X}_1,\ldots,X_n = x_1,\ldots,x_n)\\
&= x_n\frac{\mathrm{P}_Q(\tilde{X}_1=\tilde{x}_1,\ldots,\tilde{X}_{n+1}=\tilde{x}_{n+1})}{\mathrm{P}_Q(\tilde{X}_1=\tilde{x}_1,\ldots,\tilde{X}_n=\tilde{x}_n)}\\
&= x_n\frac{\beta(s_n+f+1, n+g-s_n)}{\beta(s_n+f, n+g+1-s_n)+\beta(s_n+f+1, n+g-s_n)}\\
&= x_n\frac{s_n+f}{n+f+g}
\end{aligned}$$

This corresponds to $d_{a,b}(x)$ when (a,b) relate to (f,g) such that $a = 1/(n+f+g)$, $b = (f+1)/(n+f+g)$. Which f and g to choose remains unanswered here. Many other weight functions than Beta weight functions are possible, of course. Let us continue with the risk and regret function of $d_{a,b}$ We have

$$\begin{aligned}
\mathbf{E}\,(d_{a,b}(X) - X_nX_{n+1})^2 &= \mathbf{E}\,\mathbf{E}\,X_n^2(aS_{n-1}+b-X_{n+1})^2|X_n\\
&= p\mathbf{E}\,(aS_{n+1}+b-X_{n+1})^2\\
&= p\left[a^2\mathrm{Var}S_{n-1} + \mathrm{Var}X_{n+1} + ((an-a-1)p+b)^2\right]\\
&= (a^2n-a^2+1)p^2(1-p) + p((an-a-1)p+b)^2
\end{aligned}$$

and obtain the regret function

$$\begin{aligned} S(\theta, d_{a,b}) &= R(\theta, d_{a,b}) - \theta^2(1-\theta) \\ &= (n-1)a^2\theta^2(1-\theta) + \theta((an-a-1)\theta + b)^2 \end{aligned}$$

which we can study from various perspectives. One could try to find a class $\mathcal{D} \in \{d_{a,b}\}$ that minimizes the maximum regret. But also the classes of minimax risk rules, Bayes rules, etcetera, are of interest. In practice, the induced inferences will be approximately equal.

1.9 Related well-known examples

The following problems are related to the die-rolling games of this chapter.

The quiz-master problem
Also known as the three-doors problem and the Monty Hall dilemma, this is one of the most well-known paradoxical problems in popular statistics. The problem is as follows: 'Suppose you are a contestant in a TV game show and you made it to the grand final. This is your chance to win a brand new car. The game show host shows you three doors, behind one of them is the car, behind the other two goats. To take the car home, you have to pick the right door. After you have picked a door, say door A, the host (who knows what's behind which door) opens one of the other doors, say B, showing you a goat. He gives you the opportunity to switch to door C. Is this to your advantage?' KOOI (1999)

Many people's initial intuition will say 'No' to this question. You picked at random, all probabilities must be $\frac{1}{3}$. However, these of course 'change' when door B is opened. A simple probabilistic exploration of the problem shows that it is in your advantage to change doors, since the car is behind door C with probability 2/3. This only holds in the context where both game show host and contestant are 'machines' who will choose at random between the available options, with equal probabilities.

This problem is known for about 40 years, see e.g. MOSTELLER (1965, p. 4) and SELVIN (1975). Another reference is KOOI (1999), where the problem is dealt with using Probabilistic Epistemic Logic.

The prisoner's dilemma
This problem is probably even older then the quiz-master problem (according to RAPOPORT (1974) the first appearance of this problem was in 1952 by Flood). One version of this two-player nonconstant-sum game is as follows: 'two prisoners who escaped and participated in a robbery have been recaptured and are awaiting trial for their new crime. Although both are guilty, the district attorney is not sure he has enough evidence for conviction. To entice them to testify against each other, the D.A. tells each prisoner the following:

> ' "If only one of you confesses and testifies against your partner, the person who confesses will go free while the person who does not confess will surely

> be convicted and given a 20-year jail sentence. If both of you confess, you will both be convicted and sent to prison for 5 years. If neither of you confesses, I can convict you both of a misdemeanor and you will each get 1 year in prison." What should each prisoner do?'

(Winston, 1994, p. 850). On this dilemma, entire books are written. In a mathematical-statistical context, the ones by Rapoport et al. (1965) and Poundstone (1992) are very interesting.

Again, the solution depends heavily on the information of the game being the knowledge the two prisoners have about each other. An advantage of our problem is that it is less vexatious and more in line with what we have to do in mathematical statistics at large and in distributional inference in particular.

The three prisoners problem

> 'Three prisoners, A, B, and C are locked in their cells. It is common knowledge that one of them will be executed the next day and the others will be pardoned. Only the Governor knows which one will be executed. Prisoner A asks the guard a favor: "Please ask the governor who will be executed, and then take a message to one of my friends B and C to let him know that he will be pardoned in the morning." The guard agrees, and comes back later and tells A that he gave the pardon message to B. What are A's changes of being executed, given this information?'

(Russel and Norvig, 1995, ex. 14.12). See also Diaconis and Zabell (1986). This problem attracted our attention through Op den Akker (1998).

Similar to the quiz-master paradox, inferences about the probability (that A is executed) depend on the information-strategy, this time of the guard. Quite interestingly, this is not remarked in neither Russel and Norvig (1995) nor Op den Akker (1998).

Finally we want to mention yet another example, namely the two-envelope paradox. This paradox, and its accompanying problem, will constitute the core of the following chapter. There, the problem will be dealt with in a logical, probabilistic, statistical and game-theoretic way.

Chapter 2

Trying to resolve the two-envelope problem

> *'If a question be put to [the wise man] about duty or about a number of other matters in which practice has made him an expert, he would not reply in the same way as he would if questioned as to whether the number of the stars is even or odd, and say that he did not know; for in things uncertain there is nothing probable, but in things where there is probability the wise man will not be at a loss either what to do or what to answer'*
>
> CICERO[1]

The two-envelope *paradox* is easily explained, but the corresponding *problem* is more difficult. Our discussion contains contributions from economy, psychology, logic, probability, and mathematical statistics, respectively, as well as an in-depth contribution from game theory. We conclude that the two-envelope problem does not allow a satisfactory solution. An interpretation is made for statistical science at large. The major part of this chapter coincides with ALBERS ET AL. (2003).

2.1 The two-envelope problem

In 1943, KRAITCHIK discussed the paradox of the neckties:

> 'Each of two persons claims to have the finer necktie. They call in a third person who must make a decision. The winner must give his necktie to the loser as consolation. Each of the contestants reasons as follows: 'I know what my tie is worth. I may lose it, but I may also win a better one, so the game is to my advantage'. How can the game be to the advantage of both?'

The snake in the grass is that equal 'probabilities' or 'chances' are assigned to winning and losing. 'In reality, however, the probability is not an objectively given fact, but depends on one's knowledge of the circumstances. In the present case it is wise not to try to estimate the probability.' (KRAITCHIK, 1943)

A similar paradox is the two-envelope paradox. It is unclear who gave the problem its modern form; Kraitchik already formulated a problem where two people compare

[1] Academica 2.110; taken from ZABELL, Symmetry and its discontents, **Causation, chance and crecedence**, Kluwer, 1988

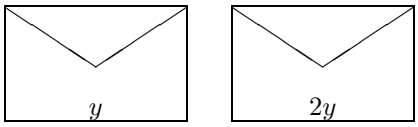


Figure 2.1: Two indistinguishable envelopes

the number of pennies in their purses. GARDNER (1982) called this the wallet game. ZABELL (1988a, 1988b) heard it from Budrys, see NALEBUFF (1989).

Two-envelope paradox.
Two indistinguishable envelopes, 1 and 2, see Figure 2.1, are submitted to 'you' (the decision maker; later the designation 'we' is used as the combination of 'you' and the statistician who tries to assist you). Envelope 2 contains a check worth twice the unknown value, say y, in Envelope 1. You choose one of the envelopes, say Envelope z, *at random* and, after opening it, you observe the value $x = zy$ contained by it. Finally you are allowed to decide between

a_1: keep the envelope you have;

a_2: return this envelope and take the other one.

Two variants will be discussed. In Variant 1, the discrete case, you are told that $y \in \mathbb{N}$; Variant 2, the continuous case, tells you $y \in \mathbb{R}^+$. A paradox appears if you would argue that choosing *at random* implies that, as the other envelope contains either $\frac{1}{2}x$ (namely if $z = 2$) or $2x$ (when $z = 1$), the expectation of the value contained by it is $\frac{1}{2} \cdot \frac{1}{2}x + \frac{1}{2} \cdot 2x = \frac{5}{4}x$, and that, hence, swapping is advantageous on the average. The snake in the grass is that you should not use the marginal or prior probability $\mathrm{P}(Z = 1)$, but the conditional or posterior probability $\mathrm{P}(Z = 1|X = x)$, given the knowledge you have, namely that $X = ZY$ has the outcome x observed. Unfortunately, the relevant posterior probabilities $\mathrm{P}(Z = z|X = x)$ $(z = 1, 2)$ are unknown to you. Moreover you might question whether y can really be regarded as the outcome of a random variable Y. The paradox has been explained, but you are still in need of a solution to the problem. Should you decide upon a_1 or upon a_2, given the outcome $X = x$? That is the question.

This problem has been discussed extensively in the literature, see, e.g., ZABELL (1988a, 1988b), NALEBUFF (1989), and the (probabilistic) references cited in Section 2.3.

2.2 Explorations

It is obvious that you are in need of some additional information, factual, contextual, or otherwise. To settle the issue 'scientifically' you might consult a variety of experts, such as an *economist*, *psychologist*, or *logician*. The economist will argue that the utility of an action is not necessarily proportional to its monetary value. The

law of diminishing returns prescribes this relation usually to be a concave function. Henceforth, it is assumed that there is complete correspondence between utility and monetary value. The psychologist will tell you that some people are risk-averse while others are risk-prone. To proceed 'in the most general way' we assume that the decision maker has a risk-neutral attitude and, hence, will try to maximize expected utilities (see Section 2.6 for a more elaborate treatment of the economist, psychologist, etcetera). This, however, requires probabilistic terminology. The logician SMULLYAN (1997) maintained that 'probability is really quite inessential to the heart of the two-envelope paradox'. He presents it as a *logical paradox*, i.e. 'two contrary, or even contradictory, propositions to which we are led by apparently sound arguments. The arguments are considered sound because, when used in other contexts, they do not seem to create any difficulty' (VAN HEIJENOORT, 1967). Here are the propositions derived by Smullyan:

Proposition 1: The amount you will gain by trading, if you do gain, is greater than the amount you will lose, if you do lose.

Proposition 2: The two amounts are really the same.

He proves both of them in the following way[2]:'To prove Proposition 1, let x be the amount you are now holding. Then the other envelope either contains $2x$ or $\frac{x}{2}$. If you gain by trading, you will gain x dollars (moving from x to $2x$), whereas if you lose by trading, you will lose only $\frac{x}{2}$. Since x is greater than $\frac{x}{2}$, then Proposition 1 is established.

To prove Proposition 2, let d be the difference between the two amounts in the envelopes (or what is the same thing, the lesser of the two amounts). Well, if you gain on the trade, you will gain d dollars. If you lose on the trade, you will lose d dollars. Since d is equal to d, then Proposition 2 is established.' (SMULLYAN, 1997, p. 174)

Both proofs seem to be sound, but it cannot be the case that both are correct. The problem turns out to be that some of the terms that are used in both proofs are ambiguous, they can be interpreted in two different ways. If you read the proofs you tend to interpret these terms in each proof in such a way that the proof is correct. But if you take the interpretation that makes one of the proofs correct and use that interpretation in the other proof then the other proof is not correct.

The ambiguous terms are 'the amount you will gain by trading, if you do gain' and 'the amount you will lose by trading if you do lose'. What do these terms refer to? This all depends on what you mean by 'if you do gain' for example. Under what circumstances do you gain? In the first proof this is the case if the amount in the envelope you did not pick is double the amount you are now actually holding. In the second proof this is the case if you picked the envelope which contains the highest amount actually available.

The difference between these interpretation becomes clear if we look at the case where

[2]In his book, Smullyan uses the letter n in his proof of proposition 1. We have replaced it with x to keep a uniform notation. In a recent publication CHASE (2002) elaborates on Smullyan.

you lose by trading. Now in the case of proof one and the first interpretation, 'the amount you gain by trading, if you do gain' refers to twice the amount that is actually in your envelope, which is four times what is in the other envelope. In case of proof two, on the other hand, 'the amount you gain by trading, if you do gain' refers to the amount you would win by trading if you had picked the other envelope, which is one times what is in the other envelope. A similar analysis is provided in (CHASE, 2002). Regardless of which interpretation you choose, this does not 'resolve' the problem.

To make the choice between a_1 and a_2 *additional knowledge* will be needed. In this respect it is natural to refer, as we shall do in Section 2.3, to the probabilistic knowledge that $\mathrm{P}(Z = 1) = \mathrm{P}(Z = 2) = \frac{1}{2}$. It will turn out that this 'factual' knowledge does not yet settle the issue. That is why we will look for more additional knowledge, perhaps of a less factual kind, in the hope that after all some 'reasonable' solution will appear. It is in this respect that we will refer to the following logical paradox. The point we shall make is that the context of this paradox enforces a solution (at least temporarily).

> *The Protagoras paradox* (GELLIUS, 1946)
> 'Euathlus, a wealthy young man, was desirous of instruction in oratory and the pleading of causes. He became a pupil of Protagoras, the keenest of all sophists, and promised to pay him a large sum of money, as much as Protagoras had demanded. He paid him half of the amount at once, before beginning his lessons, and agreed to pay the remaining half on the day when he first pleaded before jurors and won his case. Afterwards, when he had been for some little time a pupil and follower of Protagoras, and had in fact made considerable progress in the study of oratory, he nevertheless did not undertake any cases. And when the time was already getting long, and he seemed to be acting thus in order not to pay the rest of the fee, Protagoras formed what seemed to him at the time a wily scheme; he determined to demand his pay according to the contract, and brought suit against Euathlus.
>
> And when they had appeared before the jurors to bring forward and to contest the case, Protagoras began as follows: "Let me tell you, most foolish of youths, that in either event you will have to pay what I am demanding, whether judgment be pronounced for or against you. For if the case goes against you, the money will be due me in accordance with the verdict, because I have won; but if the decision be in your favour, the money will be due me according to our contract, since you will have won a case."
>
> To this Euathlus replied: "I might have met this sophism of yours, tricky as it is, by not pleading my own cause but employing another as my advocate. But I take greater satisfaction in a victory in which I defeat you, not only in the suit, but also in this argument of yours. So let me tell you in turn, wisest of masters, that in either event I shall not have to pay what you demand, whether judgment be pronounced for or against me. For if the jurors decide in my favour, according to their verdict nothing will be due you, because I

> have won; but if they give judgment against me, by the terms of our contract I shall owe you nothing, because I have not won a case." '

This is the paradox. But the context is such that Protagoras went to court. This had the following consequences according to Gellius: 'Then the jurors, thinking that the plea on both sides was uncertain and insoluble, for fear that their decision, for whichever side it was rendered, might annul itself, left the matter undecided and postponed the case to a distant day. Thus a celebrated master of oratory was refuted by his youthful pupil with his own argument, and his cleverly devised sophism failed.'

We conclude that, due to the context within which it appeared, this paradox is solved, at least temporarily, in favor of Euathlus. Later authors, e.g. STEWART (2000), missed this point. Dealing with the two-envelope paradox we shall look for similar sources of additional information, using probability theory (Section 2.3), mathematical statistics (Section 2.4), and game theory (Section 2.5), unfortunately with less success.

2.3 Probability theory

Apart from Smullyan's, most attempts to solve the two-envelope problem are in a Bayesian spirit (e.g. ZABELL, 1988a, 1988b; NALEBUFF, 1988, 1989; BROOME, 1995; CLARK AND SHACKEL, 2000; JACKSON ET AL., 1994; LINZER, 1994; and MCGREW ET AL., 1997). The formulation of the two-envelope problem chosen in Section 2.1 is such that the gain or utility

$$\begin{aligned} U(x,y,z;a_1) &= x = yz \\ U(x,y,z;a_2) &= 2^{3-2z}x = y(3-z) = \begin{cases} 2x & \text{if } z=1 \\ \frac{1}{2}x & \text{if } z=2 \end{cases} \end{aligned}$$

depends on the true values x, y, and z governing the actual experiment (one of these three can be 'deleted' because $x = yz$). To incorporate the information that $\mathrm{P}(Z=1) = \mathrm{P}(Z=2) = \frac{1}{2}$, we regard (x,z) as the outcome of a pair of random variables (X,Z) which, in principle, may assume any value $(\xi,\zeta) \in \mathcal{X} \times \{1,2\}$ where $\mathcal{X}$ is $\mathbb{N}$ or $\mathbb{R}^+$. Using the utility function in the form

$$U(\xi,\zeta;a) = \begin{cases} \xi & \text{if } a = a_1 \\ 2^{3-2\zeta}\xi & \text{if } a = a_2 \end{cases}$$

it is 'rational' to choose a such that the conditional (or posterior) expectation

$$\mathbf{E}\,(U(x,Z;a)|X=x) = \begin{cases} x & \text{if } a = a_1 \\ 2x\mathrm{P}(Z=1|x) + \frac{1}{2}x\mathrm{P}(Z=2|x) & \text{if } a = a_2 \end{cases}$$

is maximum or, equivalently, to use the 'procedure' $d^* : \mathcal{X} \to \{a_1, a_2\}$ defined by

$$d^*(\xi) = \begin{cases} a_1 & \text{if } \mathrm{P}(Z=1 \mid X=\xi) < \frac{1}{3} \\ a_2 & \text{if } \mathrm{P}(Z=1 \mid X=\xi) > \frac{1}{3}, \end{cases}$$

the choice in the case of equality being arbitrary. 'Procedure' d^* is such that the expected utility is maximum, both a priori (unconditionally) and a posteriori (conditionally, given X). Unfortunately d^* is not yet a workable procedure: the conditional probability $\mathrm{P}(Z = 1|X = x)$ has not yet been specified. Such specification requires additional information, e.g. about the way y has come into being. *Can y be regarded as the outcome of a random variable Y and, if so, can the distribution of Y be specified?* The personalist Bayesian will answer both questions affirmatively and argues as follows

In the discrete case, let $f(\eta) = \mathrm{P}(Y = \eta)$ be specified ($\eta \in \mathcal{Y}$). The joint distribution of (X, Y, Z) is then determined by (1) $X = YZ$, (2) Y and Z are stochastically independent, (3) Y has density f, (4) $\mathrm{P}(Z = 1) = \mathrm{P}(Z = 2) = \frac{1}{2}$. We obtain

$$\begin{aligned}\mathrm{P}(Z = 1|X = x) &= \frac{\mathrm{P}(Z = 1, X = x)}{\mathrm{P}(X = x)} \\ &= \frac{\mathrm{P}(Z = 1, Y = x)}{\mathrm{P}(Z = 1, Y = x) + \mathrm{P}(Z = 2, Y = \frac{1}{2}x)} \\ &= \frac{f(x)}{f(x) + f(\frac{1}{2}x)}.\end{aligned}$$

If x is odd, then $f(\frac{1}{2}x) = 0$ and $\{Z = 1\}$ is sure: swapping provides you with $2x$. If x is even, then it depends on f whether (or not) $f(\frac{1}{2}x)/f(x)$ is smaller (or larger) than 2 and swapping (or not swapping) is most profitable.

In the continuous case the difference between odd and even disappears and we have

$$\begin{aligned}\mathrm{P}(Z = 1|X = x) &= \lim_{\Delta \downarrow 0} \mathrm{P}(Z = 1|x - \Delta < X < x + \Delta) \\ &= \lim_{\Delta \downarrow 0} \frac{\mathrm{P}(|x - Y| < \Delta, Z = 1)}{\sum_{z=1}^{2} \mathrm{P}(|x - zY| < \Delta, Z = z)} \\ &= \frac{f(x)}{f(x) + \frac{1}{2} f(\frac{1}{2}x)}\end{aligned}$$

(see BROOME, 1995). Now the optimal procedure prescribes to decide upon a_2 if $f(\frac{1}{2}x)/f(x)$ is smaller than 4.

The Achilles' heel of this Bayesian solution is the assumption that y is the outcome of a random variable Y with *known* density f. In Section 2.4 we shall worry about this assumption (see also Section 2.6). In the remainder of this section, a matter of theoretical interest is discussed.

Suppose that f is such that, in the discrete case $f(\frac{1}{2}x)/f(x) \leq 2$ holds for all even values of x, or that, in the continuous case, $f(\frac{1}{2}x)/f(x) \leq 4$ holds for all x. Swapping is then always indicated according to the theory. This sounds paradoxical because if 'you' get Envelope z and 'I' get Envelope $3 - z$, we will both believe to gain by swapping. The explanation is that this situation can only appear if $\mathbf{E}\, Y = \infty$ and we both may expect an infinite amount of money, no matter what we do. The existence

of paradoxical f's is a matter of elementary analysis. In the discrete case,

$$f(\eta) = \begin{cases} (h-1)\, h^{(-s-1)}, & h > 0 \qquad \text{if } \eta = 2^s, (s = 0, 1, 2, \ldots) \\ 0 & \text{otherwise} \end{cases}$$

works. NALEBUFF (1989, p. 189) introduced this density (with $h = 3/2$), and BROOME (1995) linked this density to Daniel Bernoulli's St. Petersburg Paradox (for an explanatory and interesting reference, see DEHLING, 1997). In the continuous case

$$f(\eta) = \begin{cases} 0 & \text{if } \eta < 0 \\ (h-1)(\eta+1)^{-h}, \quad h \in (1, 2] & \text{if } \eta \geq 0 \end{cases}$$

(the case $h = 2$ was mentioned by BROOME, 1995) will do.

2.4 Mathematical statistics

The previous section is based on the assumption that y is the outcome of a random variable Y and that the distribution of Y is specified by a known function f. In practice, this assumption is questionable. It might happen, of course, that previous experiences provided a reliable estimate of f. It might also be the case that certain theories imply 'partial knowledge' of f. The thrill of the two-envelope problem, however, is in the situation that the decision-maker has no other information than that the outcome of X is x and, of course, that $\mathrm{P}(Z = 1) = \mathrm{P}(Z = 2) = \frac{1}{2}$. Theoreticians are fascinated by such problems. They will try to enforce a solution by using their mind. In the present section we start out with two rationalizations both providing a 'unique' solution (though these solutions should not be regarded as satisfactory). At the end we establish that one of these solutions (namely 'swap, no matter x') is *inadmissible* from the viewpoint of Wald's theory of statistical decision functions.

Approach 1

At the beginning of Section 2.3 it was established that the optimal 'procedure' prescribes to assign a_1 (a_2) if the posterior probability $\mathrm{P}(Z = 1|X = x)$ is smaller (larger) than $\frac{1}{3}$. The question is how to estimate this posterior probability. Note that, in Kolmogorov's theory, this conditional probability is interpreted as the conditional expectation $\mathbf{E}\,(\mathbf{1}_{\{Z=1\}}|X = x)$ and that, in general, the conditional expectation $\mathbf{E}\,(\mathbf{1}_{\{Z=1\}}|X)$ is the projection of $\mathbf{1}_{\{Z=1\}}$ on the space of all functions of X, a subspace of $L_2(\Omega, \mathcal{F}, \mathcal{P})$. This suggests to estimate $\mathrm{P}(Z = 1|X = x)$ by constructing a procedure $d : \mathcal{X} \to [0, 1]$, this construction being such that the mean squared error of prediction $\mathbf{E}\,\left(\mathbf{1}_{\{Z=1\}} - d(X)\right)^2$ is minimum. The value $d(x)$ is then the estimate required.

The joint distribution of X and Z is determined by that of (X, Y, Z) discussed in Section 2.3. The density f of Y appears as the unknown parameter governing the risk. In the continuous case, absence of information about f allows an interpretation in the sense of invariance under scale transformations. If the predictor d is required to

be scale invariant then it is constant and the solution $d \equiv \frac{1}{2}$ appears as *uniformly, best invariant predictor of* $\mathbf{1}_{\{Z=1\}}$. As $\frac{1}{2} > \frac{1}{3}$, its consequence is to swap, no matter the outcome x observed. The arguments involved are considerably manipulative because the real issue, namely that of deciding whether or not $\mathrm{P}(Z = 1|X = x)$ is smaller than $\frac{1}{3}$, is replaced by another one, namely the prediction of $\mathbf{1}_{Z=1}$. Moreover some information about f will exist: scale invariance is a mathematical idealization, not something factual to rely on.

Approach 2

All Bayesian and almost all non-Bayesian statisticians know that invariance considerations are elegant as well as dangerous. They would prefer to incorporate additional information, e.g. that $\mathbf{E}\, Y$ is approximately equal to some a priori value, say $\mathbf{E}\, Y = 100$, or, if such value is unavailable, an a posteriori value, say $\mathbf{E}\, Y = \frac{2}{3}x$ (because $\mathbf{E}\, X = \frac{3}{2}y$). Such information $\mathbf{E}\, Y = \mu$ (μ specified) can be incorporated elegantly by constructing f such that the entropy

$$I(f) = \begin{cases} -\sum_{y=1}^{\infty} f(y) \log f(y) & \text{(discrete case)} \\ -\int_0^{\infty} f(y) \log f(y)\, \mathrm{d}y & \text{(continuous case)} \end{cases}$$

is maximum under the restrictions that $\mathbf{E}\, Y = \mu$ and $\sum f = 1$ (discrete case) or $\int f = 1$ (continuous case). It is well known that the solution to this optimization problem is of the exponential form $f(y) = \exp(\theta y - \psi(\theta))$ where

$$\psi(\theta) = \begin{cases} \log(\sum_{y=1}^{\infty} \exp(\theta y)) = \theta - \log(1 - \mathrm{e}^{\theta}) & \text{(discrete case)} \\ \log(\int_0^{\infty} \exp(\theta y)\, \mathrm{d}y) = -\log(-\theta) & \text{(continuous case)} \end{cases}$$

and $\theta \in (-\infty, 0)$ is determined such that $\mathbf{E}\, Y = \psi'(\theta) = \mu$. This is easily obtained from the Lagrangian function $L(f) =$

$$= \begin{cases} -\sum f(u) \log f(u) + \lambda \left(\sum u f(u) - \mu\right) + \gamma \left(1 - \sum f(u)\right) & \text{(discrete case)} \\ -\int f(u) \log f(u) + \lambda \left(\int u f(u) - \mu\right) + \gamma \left(1 - \int f(u)\right) & \text{(continuous case)} \end{cases}$$

with optimality conditions $\frac{\partial L}{\partial f(u)} = \frac{\partial L}{\partial \lambda} = \frac{\partial L}{\partial \gamma} = 0$ (cf. GOLAN ET AL., 1986, JAYNES, 1996). If one already 'knows' the solution, then a proof can be obtained by using the positivity of the Kullback-Leibler information number.

The interesting case is that where, in absence of further information, $\mu = \frac{2}{3}x$ is used and the ratio $f(\frac{1}{2}x)/f(x) = \exp(-\frac{1}{2}\theta x)$ is compared with the value 2 in the discrete case when x is even (if x is odd, then swapping is always indicated), and with the value 4 in the continuous case.

In the discrete case we have that $\psi'(\theta) = (1 - \mathrm{e}^{\theta})^{-1}$ is equal to $\frac{2}{3}x$ if $\theta = \log(1 - \frac{3}{2x})$ and $f(\frac{1}{2}x)/f(x) = (1 - \frac{3}{2x})^{-\frac{1}{2}x}$ is larger than 2 for all $x \in \{2, 4, 6, \ldots\}$ and, hence, swapping if x is odd and not swapping if x is even is indicated.

In the continuous case we have that $\psi'(\theta) = -\frac{1}{\theta} = \frac{2}{3}x$ implies $\theta = -\frac{3}{2x}$ and that $f(\frac{1}{2}x)/f(x) = \exp(\frac{3}{4})$ is smaller than 4 for all x and, hence, swapping is always indicated.

Inadmissibility considerations
The above mathematical-statistical discussions were considerably manipulative. Two procedures were suggested namely

$$d(\xi) = \begin{cases} a_1 & \text{if } \xi \text{ is even} \\ a_2 & \text{if } \xi \text{ is odd} \end{cases}$$

in the discrete case (see Section 2.5 for an extensive discussion), and

$$d(\xi) \equiv a_2$$

in the continuous case. It is easy to establish that the latter procedure, as well as that where $d(\xi) \equiv a_1$, is inadmissible in the sense that other procedures exist with expected utility never smaller and often larger than the expected utility $\frac{3}{2}\mathbf{E}\ Y$ of these constant procedures. An example is

$$d_k(\xi) = \begin{cases} a_1 & \text{if } \xi > k \\ a_2 & \text{if } \xi \leq k \end{cases}$$

with k a predetermined constant in $\mathbb{R}^+$. To establish that d_k and many other procedures provide expected utilities above $\frac{3}{2}\mathbf{E}\ Y$, it is of interest to consider 'randomized procedures' or, equivalently, test functions $\varphi : \mathcal{X} \to [0,1]$ with the interpretation that $\varphi(x)$ is the probability of a_1 and $\alpha(x) = 1 - \varphi(x)$ that in favor of a_2. The utility of such randomized decision is

$$U(x,y,z;a_1)\varphi(x) + U(x,y,z;a_2)\alpha(x) = x\varphi(x) + (3y - x)\alpha(x)$$

if the outcome of (X,Y,Z) is $(x = yz, y, z)$. The expected utility is

$$\begin{aligned} \mathbf{E}\ (X\varphi(X) + (3Y - X)\alpha(X)) &= \mathbf{E}\ (2X - 3Y)\varphi(X) + \mathbf{E}\ (3Y - X) \\ &= \mathbf{E}\ (2X - 3Y)\varphi(X) + \tfrac{3}{2}\mathbf{E}\ Y \end{aligned}$$

and the excess over the expected utility $\frac{3}{2}\mathbf{E}\ Y$ based on $d(\xi) \equiv a_1$ or $d(\xi) \equiv a_2$ is

$$\begin{aligned} \mathbf{E}\ (2X - 3Y)\varphi(X) &= \tfrac{1}{2}\mathbf{E}\ (2Y - 3Y)\varphi(Y) + \tfrac{1}{2}\mathbf{E}\ (4Y - 3Y)\varphi(2Y) \\ &= \tfrac{1}{2}\mathbf{E}\ Y(\varphi(2Y) - \varphi(Y)) \end{aligned}$$

and this is strictly positive if φ is strictly increasing.
If $\varphi = \mathbf{1}_{(k,\infty)}$ corresponds to d_k, then the excess is equal to

$$\tfrac{1}{2}\mathbf{E}\ Y(\mathbf{1}_{(k,\infty)}(2Y) - \mathbf{1}_{(k,\infty)}(Y) = \tfrac{1}{2}\mathbf{E}\ Y\mathbf{1}_{(k/2,k)}(Y)$$

which is strictly positive if $\mathrm{P}(\frac{1}{2}k \leq Y \leq k) > 0$.

Conclusion
We still are not in the situation that the problem can be regarded as solved. In this respect it is interesting to recall that the context of the Protagoras paradox mentioned in Section 2.2 enforced a (temporary) solution. A contextual ingredient already available but not yet explored is that illustrated by Kraitchik's question at the beginning of Section 2.1: how can the *game* be to the advantage of both? The next section is concerned with this perspective.

2.5 Game theory

We restrict the attention to the formulation of the problem involving only two players: 'you/we' and the mysterious player who completes the envelopes. This situation is very similar to the ones studied by WALD (1964) in his decision-theoretic approach to the problems of statistics. Our approach is in line with FERGUSON (1967), and will use the notation style customary in the theory of statistical decision functions.

A game being defined as a triple, the first game $(\mathcal{A}, \mathcal{Y}, U)$ to consider is that where Player 1 chooses an action from $\mathcal{A} = \{a_1, a_2\}$, Player 2 (the rich eccentric or Nature) chooses y from $\mathcal{Y}$ (either $\mathbb{N}$ or $\mathbb{R}^+$) and

$$U(y, z; a) = \begin{cases} yz & \text{if } a = a_1 \\ y(3 - z) & \text{if } a = a_2 \end{cases}$$

goes from Player 2 to Player 1 ('you/we'). This payoff depends on the outcome z of Z. This formulation is inadequate in the sense that Player 2 cannot be regarded as the 'minimizing' player and also in the sense that the information $x = yz$ available to Player 1 has not been used. To allow for the last mentioned statistical input, the game $(\mathcal{A}, \mathcal{Y}, U)$ is replaced by the game $(D, \mathcal{Y}, U)$ where we now have that Player 1 chooses a decision procedure $d : \mathcal{X} \to \mathcal{A}$ from the class D of all (nonrandomized) rules of this kind. The payoff is defined as

$$\begin{aligned} U(y; d) &= \mathbf{E}\, U(y, Z; d(x, Z)) \\ &= \begin{cases} y & \text{if } d(y) = a_1,\ d(2y) = a_2 \\ 2y & \text{if } d(y) = a_2,\ d(2y) = a_1 \\ \frac{3}{2}y & \text{if } d(y) = d(2y) \end{cases} \end{aligned}$$

This formulation still is irrelevant in the sense that Player 2 will choose y as small as possible. If in the discrete case the specification $\mathcal{Y} = \{1, 2, \ldots, n\}$ is made (for some given n) then taking $y = 1$ is the minimax strategy of Player 2 and any rule with $d(1) = a_2$ and $d(2) = a_1$ is such that the minimum payoff is maximum. The two-person zero-sum game $(D, \mathcal{Y}, U)$ is 'determined' in the sense that $\max_d \min_y U(y; d) = \min_y \max_d U(y; d) = 2$ while the saddle points are characterized in the above. Randomization can be dispensed with.

In his review of WALD (1964), SAVAGE (1951) explained the preoccupation by losses, risks, errors, etcetera in the Neyman-Pearson-Wald approach to statistics. In this approach we are not discussing utilities but regrets; shortcomings with respect to the best. Given y, the best we can do is to swap if we observe y and to keep what we have if we observe $2y$. Hence

$$\begin{aligned} L(y, z; a) &= \max_{h=1,2} U(y, z; a_h) - U(y, z; a) \\ &= 2y - U(y, z; a) \\ &= \begin{cases} (2 - z)y & \text{if } a = a_1 \\ (z - 1)y & \text{if } a = a_2 \end{cases} \end{aligned}$$

and the expected loss, given y, is equal to

$$R(y;d) = \begin{cases} y & \text{if } d(y) = a_1,\ d(2y) = a_2 \\ 0 & \text{if } d(y) = a_2,\ d(2y) = a_1 \\ \frac{1}{2}y & \text{if } d(y) = d(2y) \end{cases}$$

After this *Umwertung aller Werte*, the rich eccentric has become Player 1 while 'we' are now Player 2, the minimizing player in the two-person zero-sum game where we are allowed to choose from the set D and pay the amount

$$R(y,d) = \mathbf{E}\, L(y, Z; d(X))$$

where $X = yZ$ and this risk depends on the value y chosen by Player 1, the equivalent of 'Nature' in Wald's approach to the problems of statistics, initiated in his review of VON NEUMANN AND MORGENSTERN (1944) (see WALD, 1947). The game $(\mathcal{Y}, D, R)$ is not determined because the lower value of the game

$$\underline{v} = \sup_{y\in\mathcal{Y}} \inf_{d\in D} R(y,d) = 0,$$

is less than the upper value

$$\overline{v} = \inf_{d\in D} \sup_{y\in\mathcal{Y}} R(y,d) > 0.$$

In regular situations, it is a necessity that $\mathcal{Y}$ is finite for $\overline{v} < \infty$. Exceptions (irregular situations) are special constructions such as $\mathcal{Y} = \{1, 3, 5, \ldots\}$ where, of course, the optimal solution

$$d(x) \begin{cases} a_1 & \text{if } x \text{ is even} \\ a_2 & \text{if } x \text{ is odd} \end{cases}$$

yields $R(y,d) = 0$ for all y.

When no restrictions are put on $\mathcal{Y}$ then $\overline{v} = \infty$ because $\max(R(\frac{n}{2}, d), R(n,d)) = \frac{n}{4}$ in the nonrandomized case, and $\max(R(\frac{n}{2}, \delta), R(n,\delta)) = \frac{n}{6}$ after randomization. To overcome this difficulty and to establish a positive result we will discuss the specific cases where the *upper bound* n is specified a priori. We will review both the continuous $(\mathcal{Y} = [0,n])$ and the discrete case $(\mathcal{Y} = \{1, 2, \ldots, n\})$.

The case $\mathcal{Y} = [0,n]$, n given

We are interested in constructing the minimax risk procedure d^* (if it exists) and start out by noting that Bayesian and Laplacian statisticians are attracted by the uniform prior with density $f(\eta) = n^{-1}$, $(0 \le \eta \le n)$. Section 2.3 provides that the corresponding Bayes rule prescribes

$$d^*(x) = \begin{cases} a_2 & \text{if } x \le n \\ a_1 & \text{if } x > n \end{cases}$$

which, quite surprisingly, is minimax as well, with respect to the class D of all nonrandomized Bayes rules. To establish this minimaxity of d^*, we note that

$$R(y,d^*) \;=\; \begin{cases} \frac{1}{2}y & \text{if } 0 < y \le \frac{1}{2}n \\ 0 & \text{if } y > \frac{1}{2}n \end{cases}$$

and that, hence, $\sup_{y\in[0,n]} R(y,d^*) = \frac{1}{4}n$. Next we concentrate the attention on the two possibilities $y = \frac{1}{2}n$ and $y = n$ having $x \in \left\{\frac{1}{2}n, n, 2n\right\}$ as consequence. Any nonrandomized rule d assigns values $d(\frac{1}{2}n)$, $d(n)$, $d(2n)$ to these outcomes. The most appropriate assignments (a_2, a_1, a_1) and (a_2, a_2, a_1) lead to maximum risks equal to $\frac{1}{2}n$ and $\frac{1}{4}n$, respectively. Hence $\sup_y R(y,d) \ge \frac{1}{4}n$ holds for all d. Equality holds for d^* which, hence, is minimax in the sense that

$$\sup_{y\in\mathcal{Y}} R(y,d^*) = \inf_{d\in D} \sup_{y\in\mathcal{Y}} R(y,d) = \overline{v}.$$

Note that minimaxity holds with respect to the class D of nonrandomized mixed rules. If randomization is allowed, the upper value $\overline{v}$ of the game can be decreased to $\frac{n}{6}$ by taking

$$\varphi(x) = \begin{cases} \frac{1}{3}, & x \le \frac{n}{2} \\ \frac{2}{3}, & \frac{n}{2} < x \le n \\ 1, & x > n^3 \end{cases}$$

The case $\mathcal{Y} = \{1,\ldots,n\}$, n given
Following the suggestions in FERGUSON (1967), we consider the mixed extension $(\mathcal{Y}^*, \mathcal{D}, r)$. Here $\mathcal{Y}^*$ is the class of all probability measures τ on $\mathcal{Y}$, each one characterizable by an element f from the unit simplex S_n in $\mathbb{R}^n$, the coordinates f_η corresponding to the probabilities assigned to the possibilities η for y. The class $\mathcal{D}$ is that of behavioral randomized rules $\delta : \mathcal{X} \to \mathcal{A}^*$ which, as indicated before, prescribe that an action is taken from $\mathcal{A} = \{a_1, a_2\}$ according to a random mechanism which chooses a_1 with probability

$$\varphi(x) = \delta(x)(\{a_1\}),$$

where the test function $\varphi : \mathcal{X} \to [0,1]$ with

$$\mathcal{X} = \begin{cases} \{1,2,\ldots,n,n+2,n+4,\ldots,2n\} & \text{if n is even} \\ \{1,2,\ldots,n,n+1,n+3,\ldots,2n\} & \text{if n is odd.} \end{cases}$$

The loss involved in such randomized decision is defined as

$$\begin{aligned} L(y,z,\delta(x)) &= \varphi(x)L(y,z;a_1) + (1-\varphi(x))L(y,z,a_2) \\ &= \varphi(x)(3y-2x) + x - y \end{aligned}$$

where, of course, $x = yz$. The risk (expected loss) given y is equal to

$$\begin{aligned} R(y,\delta) &= \mathbf{E}\,\varphi(yZ)(3y-2yZ) + y\mathbf{E}\,(Z-1) \\ &= \tfrac{1}{2}y(1+\varphi(y)-\varphi(2y)). \end{aligned}$$

with special cases (needed later on)

$$\begin{aligned} R(\tfrac{n}{2},\delta) &= \tfrac{1}{2}\tfrac{n}{2}(1+\varphi(\tfrac{n}{2})-\varphi(n)) = \tfrac{n}{6} \\ R(n,\delta) &= \tfrac{1}{2}n(1+\varphi(n)-\varphi(2n)) = \tfrac{n}{6}, \end{aligned}$$

if n is even. In case n odd, we have $R(\frac{n-1}{2},\delta) = R(n-1,\delta) = \frac{n-1}{6}$. Note that there is no reason not to choose $\varphi(2n) = 1$. To minimize $\max_{y\in\{1,\ldots,n\}} R(y,\delta)$ (and for some other purposes as well) it is interesting to consider the Bayes risk

$$r(\tau,\delta) = \sum_{y=1}^{n} R(y,\delta)f(y) = \tfrac{1}{2}f^T(a+A\varphi)$$

where $a = (1,2,\ldots,n)^T$ and A is equal to

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & 2 & -2 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 2 & 0 & -2 & 0 \\ 0 & 0 & 3 & 0 & -3 \end{pmatrix}, \quad \text{and}$$

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & -2 & 0 & 0 \\ 0 & 0 & 3 & 0 & -3 & 0 \\ 0 & 0 & 0 & 4 & 0 & -4 \end{pmatrix}$$

in the cases $n = 2$, 3, 4. Extensions to $n \geq 5$ are obvious.

Any of the general minimax risk theorems of NIKAIDÔ (1953, 1959; see SCHAAFSMA, 1971 for more references) implies that

$$\max_f \min_\varphi r(f,\varphi) = \min_\varphi \max_f r(f,\varphi) = \nu$$

and that a saddle point (f^*,φ^*) exists such that

$$r(f,\varphi^*) \leq \nu = r(f^*,\varphi^*) \leq r(f^*,\varphi)$$

holds for all $f \in S_n$ and $\varphi \in \Phi$. Here we used that r: $S_n \times \Phi \to \mathbb{R}$ defined by

$$r(f,\varphi) = \tfrac{1}{2}f^T(a+A\varphi)$$

is linear in f and affine-linear in φ. The theorems imply that the procedure φ^* is minimax and the distribution τ^* with density f^* is least favorable. The saddle point (f^*,φ^*) is not necessarily unique and in our problem it is not unique (if $n \geq 2$). We are now, fortunately, well equipped to characterize the minimax risk procedures.

Theorem 2.1 *If n is* even, *then $\varphi^*(\frac{n}{2}) = 0$ and $\varphi^*(n) = \frac{1}{3}$ is necessary to have $\max_\eta R(\eta,\varphi) \leq \frac{n}{6}$. If we choose φ^* such that*

$$\varphi^*(i) = \begin{cases} 0 & \text{if } 1 \leq i \leq \frac{n}{2} \\ \frac{1}{3} & \text{if } \frac{n}{2} < i \leq n \\ 1 & \text{if } n < i \leq 2n \end{cases}$$

accordingly, then we have $\max_\eta R(\eta,\varphi^*) = \frac{n}{6}$. *The class of all minimax risk procedures is characterized by*

$$\left\{\varphi = (\varphi_1,\varphi_2,\ldots,\varphi_n,\varphi_{n+2},\ldots,\varphi_{2n})^T \mid \max_\eta \tfrac{1}{2}\eta(1+\varphi(\eta)-\varphi(2\eta)) \le \tfrac{n}{6}\right\}.$$

If n *is* odd, *then* $\varphi^*(\frac{n-1}{2}) = 0$ *and* $\varphi^*(n-1) = \frac{1}{3}$ *is necessary to have* $\max_\eta R(\eta,\varphi) \le \frac{n-1}{6}$. *If we choose* φ^* *such that*

$$\varphi^*(i) = \begin{cases} 0 & \text{if } 1 \le i \le \frac{n-1}{2} \\ \frac{1}{3} & \text{if } \frac{n-1}{2} < i \le n-1 \\ 0 & \text{if } i = n \\ 1 & \text{if } n < i \le 2n \end{cases}$$

then we have $\max_\eta R(\eta,\varphi^*) = \frac{n-1}{6}$. *The class of all minimax risk procedures is characterized by*

$$\left\{\varphi = (\varphi_1,\ldots,\varphi_n,\varphi_{n+1},\varphi_{n+3},\ldots,\varphi_{2n})^T \mid \max_\eta \tfrac{\eta(1+\varphi(\eta)-\varphi(2\eta))}{2} \le \tfrac{n-1}{6}\right\}.$$

Proof
It follows from $\max(R(\frac{n}{2},\delta), R(n,\delta)) = \frac{n}{6}$ that $\min_{\delta\in\mathcal{D}} \max_\eta R(\eta,\delta) \ge \frac{n}{6}$. If we choose $\varphi = \varphi^*$ as mentioned, then indeed $\max_\eta R(\eta,\varphi^*) = \frac{n}{6}$. It is easily seen that

$$\begin{aligned} R(\eta,\varphi^*) &= \tfrac{1}{2}\eta(1+\varphi^*(\eta)-\varphi^*(2\eta)) \\ &= \begin{cases} \frac{1}{2}\eta & \text{if } \eta \le \frac{n}{4} \\ \frac{1}{3}\eta & \text{if } \frac{n}{4} < \eta \le \frac{n}{2} \\ \frac{1}{6}\eta & \text{if } \frac{n}{2} < \eta \le n \end{cases} \end{aligned}$$

The maximum risk will thus be $\frac{n}{8}$ in case $y \le \frac{n}{4}$, and $\frac{n}{6}$ in cases $\frac{n}{4} < y \le \frac{n}{2}$ and $\frac{n}{2} < y \le n$. Hence, $\max_\eta R(\eta,\varphi^*) = \frac{n}{6}$.
That this solution is not unique is quickly seen by taking $\tilde{\varphi}^*$ equal to φ^*, with the exception that $\tilde{\varphi}^*(1) = 1$. Then

$$R(\eta,\tilde{\varphi}^*) = \begin{cases} 1 & \text{if } \eta = 1 \\ R(\eta,\varphi^*) & \text{elsewhere} \end{cases}$$

and (if $n > 6$) no harm is done to the maximum.
In case of n odd, then the necessity of $\varphi^*(\frac{n-1}{2}) = 0$, $\varphi^*(n-1) = \frac{1}{3}$ is easily shown, in the same way as for the even case. That $\max_\eta R(\eta,\varphi^*) \le \frac{n-1}{6}$ is trivial after the observation that

$$\begin{aligned} R(\eta,\varphi^*) &= \tfrac{1}{2}\eta(1+\varphi^*(\eta)-\varphi^*(2\eta)) \\ &= \begin{cases} \frac{1}{2}\eta & \text{if } \eta \le \frac{n-1}{4} \\ \frac{1}{3}\eta & \text{if } \frac{n-1}{4} < \eta \le \frac{n-1}{2} \\ \frac{1}{6}\eta & \text{if } \frac{n-1}{2} < \eta \le n-1 \\ 0 & \text{if } \eta = n \end{cases} \end{aligned}$$

Analogue to the case n even, the solution and its non-uniqueness are trivial. □

2.6 Discussion: the limits of reason

Stripped to its logical essentials, the two-envelope problem is about choosing between the sure gain x and the unsure gain that is either $\frac{1}{2}x$ or $2x$. Issues like the one discussed in the two-envelope problem cannot be settled unless something additional, either factual or fictitional or contextual, is incorporated. The Protagoras paradox is of particular interest in this respect because, there, the *context* was such that a (temporary) solution was enforced. With respect to the two-envelope problem one has to go beyond the logical essentials because the *factual* knowledge $\mathrm{P}(Z = 1) = \mathrm{P}(Z = 2) = \frac{1}{2}$ has to be incorporated. The Bayesian solutions obtained, however, were not applicable because they depended on the unknown function f. In Section 2.4 an attempt was made to settle the issue mathematically. These attempts were not successful either. In Section 2.5 we tried to incorporate arguments from the theory of games and from Wald's theory of statistical decision functions. It was only after the specification of the 'upper bound' n for y, that we could arrive at 'something not unreasonable'.

We conclude that, if almost no information exists with respect to f, then it is wise *not* to try to estimate the posterior probability $\mathrm{P}(Z = 1|X = x)$. Note that half a century ago, Kraitchik made a similar statement, see Section 2.1. If an optimal decision or optimal procedure is advocated then a scrutiny will reveal that it largely is based on something fictitious. To put it in Whitehead's words, it would be a *fallacy of misplaced concreteness* if one completely accepts such result.

We are interested in the two-envelope problem, because there are some consequences for Statistical Science at large. It often happens that the statistician is asked to use data in order to compute some posterior probability, to make a distributional inference, or to suggest an optimal decision. Some, perhaps many, of these situations are such that the lack of relevant information is so large that it is wise *not* to try to settle the issue. This leaves us with the problem to draw a distinction between those situations where the factual information is too weak to make a meaningful statement and those where the factual information is sufficiently strong. The difficulty is, of course, in the area between.

The criticism mentioned here does not only refer to the Bayesian approach where it has to be considered unwise to specify a prior distribution unless relevant information is sufficiently abundant. Criticism refers also to the Neyman-Pearson-Wald approach where we are trying to discuss 'many possible worlds', one for every value of the parameter θ. Usually we assume that exactly one of these worlds is factually true, but we don't know which one. It is well-known from logic, and is exemplified by the two-envelope paradox, that paradoxes may appear if one compares arguments from different worlds, one factual, the other fictitious. We, perhaps, have to admit that our approaches are considerably manipulative and, as KARL PEARSON noticed in his paper about the Fundamental Problem of Practical Statistics (1920), *'in the nearness of an abyss'*.

In our approach to the two-envelope problem, the economist, psychologist and logician

did not get as much attention as the probabilist, mathematical statistician and game theorist, because we considered $\mathrm{P}(Z = 1) = \mathrm{P}(Z = 2) = \frac{1}{2}$ to be *factual knowledge* which should not be ignored.

The two-envelope problem is an extremely simple example of the complicated socio-economic and other issues to be settled in practice. Sciences like economy, sociology, and psychology try to clarify these issues by emphasizing various aspects of the managerial use of information. The factual basis of everything is, of course, that data are collected representing the actual state of reality. Unfortunately, such statistical data leave room for uncertainties of various kinds. The sciences indicated have much more to say than is suggested in the context of the two-envelope problem. What they have to say, however, is largely of a qualitative 'existential' nature. The economist may infer from empirical data that assessments of utilities and probabilities are considerable subjective, i.e. different from person to person. The psychologist may use his data to characterize people as systematically risk averse or risk prone. The sociologist will argue that economic behaviour is much less 'rational' than a decision-theorist might hope. The two-envelope problem provides a nice illustration of what Kant called *die Grenzen der Vernunft* — the limits of reason — and the need of factual information to make inferences and decisions 'reasonable'.

Part II

Statistical inference, distributional inference in particular

Chapter 3

Estimating a density by adapting an initial guess

'In fact, one of the few points of agreement between the antagonists R.A. Fisher and J. Neyman was that one would seldom have enough information about the unknown state of nature to assume a prior distribution for it and that lacking such prior possibilities the formula of Bayes would not be applicable.'

T. Ferguson[1]

Most of classical statistics is based on a statistical model $\mathcal{P}$ and the idea that nothing is known a priori about the location of the true distribution P in the family $\mathcal{P}$. A priori information about the location of P in $\mathcal{P}$ is almost always a factual reality. We are interested in situations where this information is expressed in the form of an initial guess P_0 of P. It goes without saying that this initial guess may be difficult to compose and that it should not dominate our inferences.

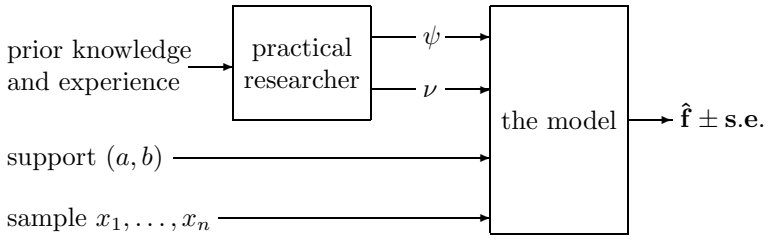
For a large variety of reasons, it is of interest to pay attention to nonparametric density estimation methods and to go beyond the familiar kernel methods. See, e.g., Silverman (1986) and De Bruin et al. (1999) for more specific motivation.

When initial knowledge is present, it should be incorporated by the model that is used for generating inferences. De Bruin et al. (1999) provide a unique method for density estimation, where such initial knowledge is taken into account. The method will be defined in Section 3.2 and discussed in the sections thereafter. These discussions show that (some) improvement is needed, amongst other reasons because the rate of convergence is too low to be satisfactory. Improvement is found in a U-statistic symmetrization, see Sections 3.6 and 3.7. Theoretical and simulation results suggest that our improved method is a good alternative to existing density estimation methods. In Chapter 5 an extensive case study is performed.

The current chapter is based on the article by Albers and Schaafsma with the same title that will appear in Computational Statistics and Data Analysis (2003a). Most results in Sections 3.3 and 3.4 were previously published as a technical report (Albers and Schaafsma, 2001a).

[1] Mathematical Statistics – a decision-theoretic approach, Academic Press, New York, 1967, p. 32.

Figure 3.1: Visualization of our method. The researcher's outputs, an initial distribution ψ and ν, a 'measure of belief in ψ', are inputs to the model.



3.1 Introduction

Many statisticians, epistemologists, and other scientists have discussed the fundamental problem of practical statistics, which is as follows. *Given* is the number m of successes in n independent and identical Bernoulli trials with unknown probability of success p. *Required* is a numerical assessment of the probability of s successes in a future series of r experiments. Of course, if n is large then the $\text{Bin}(r, \frac{m}{n})$-distribution will approximately yield the required posterior probabilities. The 'solution' for non-large n, provided by Bayes, Laplace, Pearson, Fisher and others was that in the absence of further information about p one should use the Bayesian approach with uniform prior. The De Finetti Representation Theorem (see, e.g., BILLINGSLEY, 1995) states that (probabilistic) coherency is impossible, unless the Bayesian approach is adopted. In this Bayesian approach one needs to express a priori knowledge about the statistical model in the form of a (subjective) prior distribution. This is a hazardous task, and serious 'misspecification' of the prior will lead to bad inferences. 'The Bayesian approach to statistical problems, though fruitful in many ways, has been rather unsuccessful in treating nonparametric problems. This is due primarily to the difficulty in finding workable prior distributions on the parameter space, which in nonparametric problems is taken to be a set of probability distributions on a given sample space.' (FERGUSON, 1973). Note that Ferguson is referring to nonparametric problems for which he introduced Dirichlet-process priors. These are convenient if the distribution function F has to be estimated, but not if a nonparametric density estimate is required. We concentrate the attention of such continuous analogue of the above, with $r = 1$, is as follows.

Given are the outcomes $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ of an independent random sample $X_1, \ldots, X_n$ from a probability distribution on $\mathbb{R}$ with a density f which is 'smooth' (at least continuous) and strictly positive on a given support (a, b), e.g. $(0, \infty)$, but further (largely) unknown. *Required* is an estimate $f_n(x)$ of $f(x)$ in a given point x or, more generally, an estimate f_n of f on (a, b), including a statement about the uncertainty involved, see Figure 3.1 for a visualization. Such $f_n(x)$ should be regarded as distributional inference for the next observation $x_{[n+1]}$.

The performance of density estimates is measured by specifying a difference between the true density and the estimate. Many dissimilarity coefficients can be considered (such as L_1 or L_2-distance between the densities, maximal absolute difference, L_1-distance between the quantile functions, or, equivalently, L_1-distance between the distribution functions, etcetera). We restrict the attention to the L_1-distance or total variation distance

$$||f_n - f||_1 = \int_a^b |f_n(x) - f(x)| \, \mathrm{d}x.$$

Note: various authors, e.g., DEVROYE (1987), define the total variation as

$$d_{TV} = \sup_A \left| \int_A f_n - \int_A f \right|,$$

which is half the L_1-distance. Although this definition has the advantage that the metric is 'normalized' (has values between 0 and 1), we use the first mentioned definition, taken from the field of functional analysis (see, e.g.DUNFORD AND SCHWARTZ, 1957).

We prefer the L_1-distance (and with that the total variation distance) above other dissimilarity coefficients because it is invariant under all piecewise continuous (differentiable) bijections and because it provides the metric which was preferred by a variety of other authors (see, e.g., DEVROYE AND GYŐRFI, 1985, DEVROYE, 1987, WAND AND DEVROYE, 1993, and DE BRUIN ET AL., 1999).

Working with the distribution function F has the disadvantage that the best unbiased (hence natural) estimate, the empirical distribution function, violates the differentiability requirement. Bayesian analogues derived by FERGUSON (1973), as indicated earlier, on the basis of Dirichlet-process priors, require the specification of a distribution function Ψ, but provide Bayes estimates of F which are not differentiable either. The total variation between these estimates and the true distribution is equal to 2 and consistency is precluded (for a review of consistency and convergence, see GHOSAL, 1997, and GHOSAL ET AL., 2000). A plethora of methods exists for specifying estimates of f such that the total variation converges to 0 in probability if $n \to \infty$ and f is sufficiently smooth (see, e.g., GHOSAL ET AL., 1999). In Ferguson's theory, a Dirichlet prior must be specified by providing a distribution function. The method of DE BRUIN ET AL. (1999) is also based on the specification of a distribution function Ψ, now as an initial guess of the true distribution function F. The derivative ψ of Ψ is an initial guess of f. The theory behind this method requires that ψ is chosen a priori. It is assumed that the 'support' $\{x : \ \psi(x) > 0\}$ is specified as the true interval $\{x : \ f(x) > 0\}$ (usually $(0, \infty)$ or $(-\infty, \infty)$). The method provides a special estimate f_n of f, as well as a statement about the standard errors involved.

The underlying rationale is that there are many ways to characterize a probability distribution. One can use the distribution function F, the density function $f = F'$, the quantile function $H = F^{-1}$, the characteristic function, the moment-generating

function, etcetera. Each of these characterizations entails its own estimation approach (see SILVERMAN, 1986). The empirical distribution function is not appropriate if one is interested in the estimation of $f = F'$. For that purpose some smoothing operation has to be applied. It is, of course, an interesting question whether the empirical distribution function F_n, the empirical quantile function $H_n = F_n^{-1}$, or some analogue characteristic has to be subjected to the smoothing operation.

We concentrate the attention on the estimation of H, which will turn out to be very convenient to estimate the quantile density $h = H'$ and the probability density $f = F' = (H^{-1})'$. Our estimation will be performed by fitting a Bernstein polynomial to the data (next section), and applying an additional smoothing by using U-statistic symmetrization (Section 3.6). Other perspectives of density estimation have their own advantages and disadvantages. Of particular interest in this context are:

1. The estimation of f directly through kernel estimation, see e.g. SILVERMAN (1986) and VAN ES (1988);
2. The estimation of f by means of a Bernstein polynomial (instead of estimating $H = F^{-1}$ like we do), see e.g. VITALE (1975), and GHOSAL (2001);
3. Estimation through orthogonal expansion schemes, wavelets, splines, etcetera.

The comparison between the density estimate introduced in the next section and the ones mentioned above will be discussed in Section 3.5.

3.2 The original density estimation method

The method is based on the characterization by means of the quantile function (see for extensive literature in this context GILCHRIST, 2000). Before discussing the arguments behind the method, we give the precise definition, which is as follows.

Definition 3.1 (De Bruin et al.) *Replace the ordered sample $x_{[1]}, \ldots, x_{[n]}$ by*

$$y_{[0]} = 0, \qquad y_{[i]} = \Psi(x_{[i]}) \quad (i = 1, \ldots, n), \qquad y_{[n+1]} = 1,$$

where $\psi = \Psi'$ is an a priori guess for $f = F'$. Then the estimator B_n of the quantile function $B = (F \circ \Psi^{-1})^{-1} = \Psi \circ F^{-1}$ of the distribution of $Y_1 = \Psi(X_1)$ is

$$B_n(p) = \sum_{i=0}^{n+1} y_{[i]} \binom{n+1}{i} p^i (1-p)^{n+1-i}.$$

The estimation of f is performed via back-transformation: the estimate B_n is used to define the estimate $H_n = \Psi^{-1} \circ B_n$. Next $F_n = H_n^{-1} = B_n^{-1} \circ \Psi$ is used to estimate $F = H^{-1}$ and, finally, the estimate $f_n = F_n'$ of f is obtained by numerical differentiation. The notation $B_n(p)$ is preferred over the usual notation $B_{n+1}(p)$ because the underlying sample size is n.

The most convincing argument in favor of $B_n(p)$ is as follows. Involving the Ψ-transformed quantile distribution function $B(p) = \Psi(F^{-1}(p))$ of Y, we have the physical probability

$$\begin{aligned} \mathrm{P}(B(p) \leq Y_{[i]}) &= \mathrm{P}(p \leq G(Y_{[i]}) = U_{[i]}) \\ &= \sum_{k=0}^{i-1} \binom{n}{k} p^k (1-p)^{n-k}, \end{aligned}$$

where $G = B^{-1}$ is the probability distribution function of Y, and $U_{[i]}$ is the i-th order statistic of a sample of size n from the standard uniform distribution $\mathcal{U}(0,1)$. This result was first derived by THOMPSON (1936). This physical probability is used to postulate the epistemic probability[2]

$$\mathrm{P}(B(p) \leq y_{[i]}) = \sum_{k=0}^{i-1} \binom{n}{k} p^k (1-k)^{n-k}.$$

Plugging in a sample statistic for a population statistic to obtain a probability statement is, of course, not new. It is quite similar to Fisher's Fiducial Argument (see, e.g., FISHER, 1930 and SALOMÉ, 1998). Thompson's rule easily provides

$$\mathrm{P}(Y_{[i]} < B(p) \leq Y_{[i+1]}) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n,$$

or, epistemologically equivalent, a distributional inference about $B(p)$ such that these binomial probabilities are assigned to the intervals $(y_{[i]}, y_{[i+1]})$. Taking the expectation[3] we obtain the L-estimator for $B(p)$

$$\tilde{B}_n(p) = \sum_{i=0}^{n} h(y_{[i]}, y_{[i+1]}) \binom{n}{i} p^i (1-p).^{n-i}, \quad p \in [0,1],$$

with $h(y_{[i]}, y_{[i+1]}) \in [y_{[i]}, y_{[i+1]})$ a measurable function. When $h(y_{[i]}, y_{[i+1]}) = (1-p)y_{[i]} + py_{[i+1]}$ the Bernstein polynomial $B_n(p)$ defined in Definition 3.1 is obtained, and for $h(y_{[i]}, y_{[i+1]}) = \frac{1}{2}y_{[i]} + \frac{1}{2}y_{[i+1]}$ the Kantorovič polynomial

$$\tilde{K}_n(p) = \sum_{i=0}^{n} \frac{y_{[i]} + y_{[i+1]}}{2} \binom{n}{i} p^i (1-p).^{n-i},$$

[2] In literature, sometimes the notation $\mathrm{Q}(\cdot)$ is used for epistemic probabilities instead of $\mathrm{P}(\cdot)$, which is used for 'physical' probabilities. We do not distinguish in notation, and let the interpretation of P be defined by the context.

[3] Note that

$$\binom{n+1}{i} \int_0^1 p^i (1-p)^{n+1-i} \, \mathrm{d}p = \frac{1}{n+1}.$$

follows. The Bernstein polynomial approximation $B_n(p)$ is very convenient because the corresponding derivative

$$b_n(p) = \sum_{i=0}^{n} (y_{[i+1]} - y_{[i]}) \binom{n}{i} (n+1) p^i (1-p)^{n-i}$$

is strictly positive and continuous. If one differentiates the Kantorovič polynomial a less pleasant expression is obtained.

Polynomial estimators B_n and K_n (without Ψ-transformation) were already given and analyzed in MUÑOZ PEREZ AND FERNÁNDEZ PALACÍN (1987), and similar formulas for the density function in VITALE (1975). MUÑOZ PEREZ AND FERNÁNDEZ PALACÍN (1987) prove that $\lim_n \mathbf{E}\, B_n(p) = B(p)$ uniformly on $[0,1]$ in case $B(p)$ is continuous with bounded derivative. Furthermore, concerning the integrated risk, they provide

$$\mathbf{E} \int |B_n - B| \sim o(n^{-1/2}), \quad \text{and} \quad \mathbf{E} \int |\mathbf{E}\, B_n - B| \sim o(n^{-1/2}),$$

and, when H is also twice differentiable on $(0,1)$ with continuous derivative on $[0,1]$,

$$\mathbf{E} \int |B_n - \mathbf{E}\, B_n| \sim o(n^{-1/2}).$$

In Groningen, the study of polynomial quantile estimators started with DEHLING ET AL. (1991) , involving a study of the Islamic Mean. The Islamic Mean is obtained by taking pairwise averages of the n order statistics, after that taking pairwise averages of the $n-1$ averages, etcetera, until a single number is obtained. This L-statistic

$$T_n = \sum_{i=1}^{n} x_{[i]} \binom{n-1}{i-1} 2^{-n+1},$$

is very similar to $B_n(\frac{1}{2})$. DEHLING ET AL. (1991) show that the Islamic Mean, and hence $B_n(\frac{1}{2})$ are interesting alternatives to the sample median for estimating the population median.

DE BRUIN ET AL. (1999) extended this suggestion to the case of the entire quantile function, using semi-Bayesian and Fisherian arguments. This article also introduced the use of an 'initial guess' to increase the performance (at least in case some a priori knowledge or intuition is present). Not the original quantile distribution function $H = F^{-1}$, but one transformed to the unit interval, $B = \Psi \circ F = (F \circ \Psi)^{-1}$, is estimated. The estimator is transformed back using Ψ to obtain an estimator for H, and eventually F and f. The estimator $B_n(p)$ thus obtained in Definition 3.1 is, as can be easily seen, a linear function of the order statistics (of the transformed sample), where Bernstein polynomials are used as smooth weight functions, or kernels.

The example in Figure 3.2 illustrates the performance of the density estimation method based on a sample of size $n = 100$ from the Beta$(3,2)$-density, and $\psi = 1$.

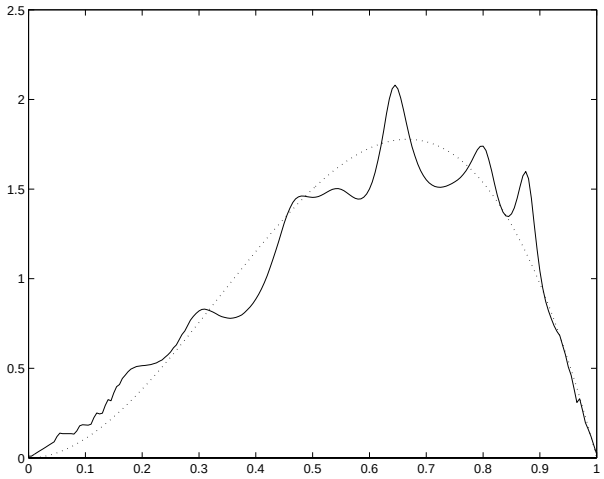


Figure 3.2: Visualization of the estimation by $f_n(x)$.

3.3 Asymptotic properties of the quantile function estimate

In this section it is established that the quantile density estimates $b_n(p)$ can be equipped with the standard error

$$\frac{b_n(p)}{\sqrt[4]{4\pi np(1-p)}}.$$

In Section 3.4, a similar result will be suggested for the density estimation function f_n which, however, does not hold in general. The proof involving the quantile estimation function (Theorem 3.1) requires two lemmas. An introduction is as follows[4].

The values $y_{[1]}, \ldots, y_{[n]}$ are ordered outcomes of an independent random sample from the distribution on $[0,1]$ with distribution function $F \circ \Psi^{-1} = B^{-1}$. If we define $u_{[i]} = F(\Psi^{-1}(y_{[i]})) = B^{-1}(y_{[i]})$, then we have that $(u_{[1]}, \ldots, u_{[n]})$ is the ordered outcome of an independent random sample from the uniform distribution on $[0,1]$. Hence, replacing $y_{[i]}$ by $B(u_{[i]})$, we obtain, by applying the Mean Value Theorem, that $v_i \in [u_{[i]}, u_{[i+1]}]$ $(i = 0, \ldots, n)$ exist such that

$$\begin{aligned} b_n(p) &= \sum_{i=0}^{n} \left(B(u_{[i+1]}) - B(u_{[i]})\right)(n+1)\binom{n}{i}p^i(1-p)^{n-i} \\ &= \sum_{i=0}^{n} b(v_i)\left(u_{[i+1]} - u_{[i]}\right)(n+1)\binom{n}{i}p^i(1-p)^{n-i}. \end{aligned}$$

[4]Reading the technicalities in this and the following section is not a necessity for a good understanding of the remainder of the chapter.

The expectations $b_{n,i} = \binom{n}{i}p^i(1-p)^{n-i}$ of the weights of the $b(v_i)$ are largest if $i \approx np$ and, hence, $v_i \approx p$. This suggests to replace $b(v_i)$ by $b(p)$ Using the approximation $\tilde{b}_n(p)$ of $b_n(p)$ thus obtained, we shall have to study the error $\tilde{b}_n(p) - b_n(p)$. This will be done in Lemma 3.2, for which we need the following result.

Lemma 3.1 *With respect to $\tilde{b}_n(p)$ just defined, we have*

$$\mathcal{L}\, n^{1/4}(\tilde{b}_n(p) - b(p)) \to \mathcal{N}\left(0, \frac{b(p)^2}{\sqrt{4\pi p(1-p)}}\right).$$

Proof
Note that the coverages $u_{[i+1]} - u_{[i]}$ are outcomes of random variables $U_{[i+1]} - U_{[i]}$ which have the same joint distribution as the r.v.'s $E_i/(E_0+\ldots+E_n)$ where $E_0, \ldots, E_n$ constitute an independent random sample from the standard negative-exponential distribution. Hence

$$\mathcal{L}\, \tilde{b}_n(p) = \mathcal{L}\left(\frac{b(p)\sum_{i=0}^n E_i\binom{n}{i}p^i(1-p)^{n-i}}{\sum_{i=0}^n E_i/(n+1)}\right).$$

The denominator in the right-hand side has a Gamma$(n+1, n+1)$-distribution. As it converges to 1 (in probability), it suffices to show that

$$\mathcal{L}\, n^{1/4}\left(\sum_{i=0}^n (E_i - 1)b_{n,i}\right) \to \mathcal{N}\left(0, \frac{1}{\sqrt{4\pi p(1-p)}}\right),$$

where $b_{n,i} = \binom{n}{i}p^i(1-p)^{n-i} = \mathrm{P}(S = s)$ if S has the binomial distribution $\mathrm{B}(n,p)$. The expectation of $\sum E_i b_{n,i}$ is obviously zero, the variance is equal to

$$\mathrm{Var}\left(\sum_{i=0}^n (E_i - 1)b_{n,i}\right) = \sum_{i=0}^n b_{n,i}^2 = \mathrm{P}(S_1 - S_2 = 0),$$

where $S_1, S_2 \sim \mathrm{B}(n,p)$ are independent and, as a consequence, $S_1 - S_2$ has a distribution on $\{-n, -n+1, \ldots, n\}$ which is asymptotically $\mathcal{N}(0, 2np(1-p))$. Using a local form of the Central Limit Theorem, we have that[5]

$$\mathrm{P}(S_1 - S_2 = 0) \approx \frac{1}{\sqrt{2\pi \cdot 2np(1-p)}}.$$

[5] An alternative derivation, is by considering S_1 and S_2 in a bivariate sense:

$$\begin{aligned}
\mathrm{P}(S_1 = S_2) &\approx \int_{-\infty}^{\infty}\int_{u-\frac{1}{2}}^{u+\frac{1}{2}} f_{S_1,S_2}(u,v)\,\mathrm{d}v\mathrm{d}u \\
&\approx \sqrt{2}\int_{-\infty}^{\infty} f_{S_1,S_2}(u,u)\,\mathrm{d}u \\
&= \frac{1}{\sqrt{2}\pi np(1-p)}\int_{-\infty}^{\infty}\exp\left\{-\frac{(u-np)^2}{np(1-p)}\right\}\mathrm{d}u \\
&= \frac{1}{\sqrt{4\pi np(1-p)}}.
\end{aligned}$$

Therefore, $\mathrm{Var}(\sum_{i=0}^{n}(E_i-1)b_{n,i}) \approx (4\pi np(1-p))^{-1/2}$, and the first step of the proof is complete.

Expectation and variance being dealt with, the asymptotic normality will follow from the Lyapunov Central Limit Theorem. This theorem implies

$$\mathcal{L}\left(\frac{\sum_{i=0}^{n} E_i b_{n,i} - 1}{\sqrt{\mathrm{Var}(\sum_{i=0}^{n} E_i b_{n,i})}}\right) \to \mathcal{N}(0,1)$$

if the condition

$$\frac{\mathbf{E}\ \sum_{i=0}^{n} |(E_i-1)b_{n,i}|^3}{(\mathrm{Var}(\sum_{i=0}^{n} E_i b_{n,i}))^{3/2}} \to 0 \quad \text{as} \quad n \to \infty$$

is satisfied. The asymptotic behaviour of the denominator has been obtained in the above as $(\sqrt{4\pi np(1-p)})^{-3/2}$. Hence it suffices to establish that the numerator is $o(n^{-3/4})$ or, equivalently, that $\sum b_{n,i}^3$ behaves that way. Note that, using notations similar to before,

$$\sum_{i=0}^{n} b_{n,i}^3 = \mathrm{P}(S_1 = S_2 = S_3) = \mathrm{P}(S_1 - S_2 = S_1 - S_3 = 0)$$

where $(S_1 - S_2, S_1 - S_3)'$ has a distribution on a subset of $\{(i,j) \mid i,j \in \{-n, -n+1, \ldots, n\}\}$, this distribution being asymptotically normal with vector of expectations $(0,0)'$, variances $2np(1-p)$ and covariance $np(1-p)$. Finally, using a local form of the multivariate Central Limit Theorem we obtain that

$$\begin{aligned} \mathrm{P}(S_1 - S_2 = S_1 - S_3 = 0) &\approx \left(4\pi^2 (np(1-p))^2 \det \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right)^{-1/2} \\ &= (2\pi np(1-p))^{-1} 3^{-1/2}. \end{aligned}$$

Hence the numerator behaves like n^{-1} which is obviously $o(n^{-3/4})$. □

Lemma 3.2 *Under weak regularity assumptions*

$$n^{1/4}(b_n(p) - \tilde{b}_n(p)) \to 0 \qquad \textit{in probability.}$$

Proof
Reformulate

$$n^{1/4}\left(b_n(p) - \tilde{b}_n(p)\right) = n^{1/4} \sum_{i=0}^{n} \left(b(V_i) - b(p)\right)\left(U_{[i+1]} - U_{[i]}\right)(n+1)b_{n,i},$$

where V_i is a random variable satisfying $U_{[i]} \le V_i \le U_{[i+1]}$. About f we know that it is smooth and strictly positive on a given interval (a,b). Almost equivalently, $b(p)$

will be assumed to be smooth and strictly positive. For the sake of convenience, the additional assumption is made that b is differentiable with $|b'(p)| \leq M$ for all p for some finite M. The Mean Value Theorem states that $n^{1/4}(b_n(p) - \tilde{b_n}(p))$ can be rewritten as

$$n^{1/4} \sum_{i=0}^{n} b'(W_i)(V_i - p)(U_{[i+1]} - U_{[i]})(n+1)b_{n,i}$$

with W_i between p and V_i. Hence

$$\begin{aligned} |n^{1/4}(b_n(p) - \tilde{b_n}(p))| &\leq M n^{1/4} \sum_{i=0}^{n} |V_i - p|(U_{[i+1]} - U_{[i]})(n+1)b_{n,i} \\ &\leq M n^{1/4} \sum_{i=0}^{n} \left(|U_{[i+1]} - p| + |U_{[i]} - p|\right)\left(U_{[i+1]} - U_{[i]}\right)(n+1)b_{n,i} \end{aligned}$$

because $V_i \in [U_{[i]}, U_{[i+1]}]$. To establish that

$$n^{1/4} \sum_{i=0}^{n} \left|U_{[i+1]} - p\right|\left(U_{[i+1]} - U_{[i]}\right)(n+1)b_{n,i} \to 0$$

in probability, it suffices to prove that the expectation of the left-hand side tends to 0. The other contribution follows similarly. The Cauchy-Schwartz inequality provides

$$\mathbf{E}\,|U_{[i+1]} - p|(U_{[i+1]} - U_{[i]}) \leq \sqrt{\mathbf{E}\,(U_{[i+1]} - p)^2}\,\frac{\sqrt{2}}{n+1},$$

because the 'coverage' $(U_{[i+1]} - U_{[i]}) \sim \text{Beta}(1, n)$ has expectation $\mathbf{E}\,(U_{[i+1]} - U_{[i]})^2 = \frac{2}{(n+1)(n+2)} < \frac{2}{(n+1)^2}$. Hence it suffices to establish that

$$n^{1/4} \sum_{i=0}^{n} \sqrt{\mathbf{E}\,(U_{[i+1]} - p)^2}\, b_{n,i} \to 0,$$

where $U_{[i+1]} \sim \text{Beta}(i+1, n-i)$ is such that

$$\mathbf{E}\,(U_{[i+1]} - p)^2 = \frac{(i+1)(n-i)}{(n+2)(n+1)^2}\left(\frac{i+1}{n+1} - p\right)^2.$$

To complete the proof we use the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for nonnegative a and b, and note that it suffices to establish that both

$$n^{1/4} \sum_{i=0}^{n} \sqrt{\frac{(i+1)(n-i)}{(n+2)(n+1)^2}}\, b_{n,i} = n^{1/4}\mathbf{E}\,\sqrt{\frac{(S+1)(n-S)}{(n+2)(n+1)^2}} \to 0,$$

and

$$n^{1/4}\sum_{i=0}^{n}\left|\frac{i+1}{n+1}-p\right|b_{n,i}=n^{1/4}\mathbf{E}\left|\frac{S+1}{n+1}-p\right|\to 0.$$

As both statements are trivial, the lemma holds. □

Theorem 3.1 *The quantile-density estimate defined in Section 3.2 has the following limiting density*

$$\mathcal{L}\, n^{1/4}(b_n(p)-b(p))\to\mathcal{N}\left(0,\frac{b(p)^2}{\sqrt{4\pi p(1-p)}}\right).$$

Proof
This follows immediately from Lemma 3.1 and Lemma 3.2. □

3.4 Asymptotic properties of the probability density function estimate

Applications in discriminant analysis and pattern recognition require estimates of $f(x)$, not of $b(p)$. The relationship between these population concepts is as follows. Starting from the distribution function F of X_1 and the corresponding quantile function $H=F^{-1}$, we have that $Y_1=\Psi(X_1)$ has distribution function $F\circ\Psi^{-1}$ and quantile function $B=\Psi\circ F^{-1}$. Hence $H=\Psi^{-1}\circ B$. For the densities $f=F'$, $h=H'$ and $b=B'$ we have that $f(x)=1/h(F(x))$ and $h(p)=1/f(H(p))$ while

$$h(p)=\frac{b(p)}{\psi(\Psi^{-1}(B(p)))}.$$

Hence

$$f(H(p))=\frac{\psi(\Psi^{-1}(B(p)))}{b(p)}$$

and

$$f(x)=\frac{\psi(\Psi^{-1}(B(F(x))))}{b(F(x))}=\frac{\psi(x)}{b(F(x))}.$$

In Section 3.2 sampling analogues were defined.

Theorem 3.2 suggests that $n^{1/4}(f_n(x)-f(x))$ might have a normal limiting distribution analogue to that of $n^{1/4}(b_n(p)-b(p))$. Unfortunately we were not able to establish such result in general. However, if $\Psi=F$, then the required result can be proven as follows.

Theorem 3.2 *If $\Psi = F$ then the difference between the density estimate defined in Section 3.2 and the true density $f = \psi$ satisfies*

$$\mathcal{L}\, n^{1/4}(f_n(x) - f(x)) \to \mathcal{N}\left(0, \frac{f(x)^2}{\sqrt{4\pi F(x)(1-F(x))}}\right).$$

Proof
Note that

$$\begin{aligned} f_n(x) - f(x) &= \frac{\psi(x)\,[b(F(x)) - b_n(F_n(x))]}{b_n(F_n(x))\; b(F(x))} \\ &\quad \frac{\psi(x)\,[b_n(F(x)) - b_n(F_n(x))] - \psi(x)\,[b_n(F(x)) - b(F(x))]}{b_n(F_n(x))\; b(F(x))}. \end{aligned}$$

We claim that the second part dominates the first one. To establish the asymptotic distribution of the second part, note that Theorem 3.1 implies

$$\mathcal{L}\, n^{1/4}\,(b_n(F(x)) - b(F(x))) \to \mathcal{N}\left(0, \frac{b(F(x))^2}{\sqrt{4\pi F(x)(1-F(x))}}\right).$$

Furthermore, not spelling out the details, we have

$$\frac{\psi(x)^2}{b_n(F_n(x))^2 b(F(x))^2}\,\frac{b(F(x))^2}{\sqrt{4\pi F(x)(1-F(x))}} \approx \frac{f(x)^2}{\sqrt{4\pi F(x)(1-F(x))}}.$$

Finally we complete the proof by, indeed, establishing that $b_n(F(x)) - b_n(F_n(x))$ is of smaller order of magnitude than $b_n(F(x)) - b(F(x))$, which is $O(n^{-1/4})$, as is shown in the following lemma. □

Lemma 3.3 *If $\Psi = F$ then*

$$b_n(F_n(x)) - b_n(F(x)) = o(n^{-1/4}).$$

Proof
The proof is similar to that of Lemma 3.1. A difficulty is that $b_n'(p)$ does *not* tend to $b'(p)$, the difference being $O(n^{1/4})$. Compare $b_n(p)$ with $b_n(r)$ $(0 \le p, r \le 1)$.

$$b_n(p) - b_n(r) = \sum_{i=0}^{n} (u_{[i+1]} - u_{[i]})(n+1)(\mathrm{P}(B_{n,p} = i) - \mathrm{P}(B_{n,r} = i))$$

where $B_{n,p} \sim \mathrm{Bin}(n,p)$ and the $u_{[i]}$ are ordered outcomes of the standard uniform distribution. This equation can be considered as the outcome of

$$\frac{\sum_{i=0}^{n} E_i(\mathrm{P}(B_{n,p} = i) - \mathrm{P}(B_{n,r} = i))}{\sum_{i=0}^{n} E_i/(n+1)}$$

with $E_0, \ldots, E_n$ a random sample from the standard negative exponential distribution and the denominator is practically 1. The expectation of the numerator is

$$\sum_{i=0}^{n} \mathrm{P}(B_{n,p} = i) - \sum_{i=0}^{n} \mathrm{P}(B_{n,r} = i) = 0,$$

and the variance is

$$\begin{aligned}
& \sum_{i=0}^{n} \left(\mathrm{P}(B_{n,p} = i) - \mathrm{P}(B_{n,r} = i)\right)^2 \\
= \;& \mathrm{P}(B_{n,p}^{(1)} = B_{n,p}^{(2)}) + \mathrm{P}(B_{n,p}^{(1)} = B_{n,r}^{(2)}) - 2\mathrm{P}(B_{n,p}^{(1)} = B_{n,r}^{(1)}) \\
\approx \;& \frac{1}{\sqrt{4\pi np(1-p)}} + \frac{1}{\sqrt{4\pi nr(1-r)}} - 2\frac{\exp\left[-\frac{1}{2}\left(\frac{p-r}{p(1-p)+r(1-r)}\right)^2\right]}{\sqrt{2\pi n(p(1-p)+r(1-r))}} \\
= \;& \frac{1}{\sqrt{4\pi n}}\left(\frac{1}{\sqrt{p(1-p)}} + \frac{1}{\sqrt{r(1-r)}} - \frac{2\sqrt{2}}{\sqrt{p(1-p)+r(1-r)}}\right) + \\
& + \frac{2}{\sqrt{4\pi n}} \frac{\exp\left[-\frac{1}{2}\left(\frac{p-r}{p(1-p)+r(1-r)}\right)^2\right] - 1}{\sqrt{p(1-p)+r(1-r)}}.
\end{aligned}$$

For the first terms, see the proof of Lemma 3.1. The last term follows from

$$B_{n,p}^{(1)} - B_{n,r}^{(1)} \sim \mathcal{N}\left(n(p-r),\ n(p(1-p)+r(1-r))\right).$$

As $p = F_n(x)$ tends to $r = F(x)$, the entire expression for the variance is, indeed, $o(n^{-1/2})$ □

If the restriction $\Psi = F$ is abandoned then the methods of proof no longer apply, since they depend on the interpretation of $b_n(p) - b_n(r)$ as outcome of the random variable specified.

Under certain regularity conditions for ψ (and f), it might be possible to prove a similar theorem for the general case, but we were unable to accomplish such result.

However, our theory is developed for situations where the practical researcher has some good intuition about f. Numerical analyses (Section 3.7) suggest that in those situations our method provides a useful contribution to existing density estimation theory. In these cases, and especially when Ψ and F are not too irregular, the behaviour of our density estimate will not be much different from that for the ideal situation, and the, now approximate, standard-error

$$\frac{f(x)}{\sqrt[4]{4\pi nF(x)(1-F(x))}}$$

will usually suffice.

3.5 Comparison with other methods

Comparison with kernel estimation methods
If one compares f_n with the 'usual' kernel estimates k_n (see, e.g., SILVERMAN, 1986), then there are some theoretical advantages of f_n but these are outweighed by practical advantages of k_n, even if the initial guess ψ of f is reliable (DE BRUIN ET AL., 1999). Some theoretical arguments are as follows.

The asymptotic distribution of f_n is studied via that of b_n (see Sections 3.3 and 3.4). This results in the approximation

$$\lim_{n\to\infty} \mathcal{L}\, n^{1/4}(f_n(x) - f(x)) \approx \mathcal{N}(0, \sigma^2(x))$$

where

$$\sigma(x) = \frac{f(x)}{\sqrt[4]{4\pi F(x)(1-F(x))}}.$$

The approximation-sign can be replaced by an equality-sign if $\Psi = F$. The rate of convergence $n^{-1/4}$ is not very satisfactory. Furthermore, the estimate f_n is not sufficiently smooth: the error ${f_n}' - f'$ in the derivative does not vanish, it is of order $O(n^{1/4})$ (Section 3.4). The expected L_1-error of f_n is of the order $n^{-1/4}$ because, using the notation $Z \sim \mathcal{N}(0,1)$, we have

$$\begin{aligned}
\mathbf{E} \int |f_n(x) - f(x)|\, \mathrm{d}x &= \int \mathbf{E}\, |f_n(x) - f(x)|\, \mathrm{d}x \\
&\approx n^{-1/4} \mathbf{E}\, |Z| \int \sigma(x)\, \mathrm{d}x \\
&= n^{-1/4} \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt[4]{4\pi}} \int \frac{\mathrm{d}F(x)}{\sqrt[4]{F(x)(1-F(x))}} \\
&= n^{-1/4} 2(\Gamma(3/4))^2 \pi^{-5/4} \\
&= .72 n^{-1/4}
\end{aligned}$$

Note that this asymptotic behaviour does neither depend on f nor on the initial guess ψ. DE BRUIN ET AL. (1999) provides simulation results in the case that f follows the Beta(2,1)-density. These results are compared with the above approximate values in Table 3.1. It follows from this table that standard errors based on $\sigma(x)$ are, perhaps, a bit conservative.

Kernel estimators seem to perform better. The rate of convergence of kernel estimates with optimal bandwidths (this requires knowledge of the unknown f) is such that $n^{2/5}(k_n(x) - f(x))$ has a limiting distribution, while ${k_n}' - f'$ vanishes in probability as $n \to \infty$, at least under regularity conditions (see, e.g., SILVERMAN, 1978, and SCHUSTER, 1969). As knowledge of f cannot be presupposed, DE BRUIN ET AL. (1999, Figure 4) performed an extensive numerical comparison between f_n and

n	10	13	25	50	75	100	150	200
DE BRUIN ET AL.	.31	.31	.27	.23	.22	.20	.19	.18
$.72n^{-1/4}$	.41	.38	.32	.27	.24	.23	.21	.19

Table 3.1: Comparison between the numerical results and the theory

kernel estimates k_n with bi-weight kernels and bandwidths based on likelihood cross-validation. This provided (for a special density f) that the L_1-error of k_n is significantly smaller than that of f_n, on the average. If the L_1-error of f_n can be decreased by about 15% then k_n and the modification of f_n would be about equally good, on the average.

Comparison with Bernstein polynomial estimations of the density function
VITALE (1975) introduced a Bernstein polynomial approach to estimate the density function directly (instead of via the quantile function). Basically, his method provides a Bernstein expansion of the histogram. After transformation to the unit interval, the estimate of Vitale is defined by

$$\tilde{f}_n^{(m)}(x) = \frac{(m+1)}{n} \sum_{j=0}^{m} h_n^{(m)}(j) \binom{n}{j} x^j (1-x)^{m-j}$$

where $h_n^{(m)}(j) = \#\{x_i \in [\frac{j}{m+1}, \frac{j+1}{m+1}]\}$. This representation is a linear combination of beta densities with random coefficients based on the observations. The parameter m is used as the degree of the polynomial expansion, and thus works as a smoothing parameter. In the next section we shall introduce a different method for smoothing, U-statistic symmetrization. Vitale shows that, under the weak assumptions that f is bounded and $f''(x)$ exists, the optimal choice (in the sense of minimal average mean squared error) of m is $m \sim n^{2/5}$. This follows from

$$\begin{aligned} \text{Bias}(x) &\sim \frac{1}{m} \frac{\tilde{f}'(x)(1-2x) + \tilde{f}''(x)x(1-x)}{2} \\ \text{Var}\,(x) &\sim \frac{m^{1/2}}{n} \frac{\tilde{f}(x)}{2\sqrt{\pi x(1-x)}} \end{aligned}$$

and trying to minimize the rate of $\text{bias}^2(x) + \text{Var}\,(x)$, yielding a convergence rate of the mean squared error of $n^{-4/5}$ for $0 < x < 1$ and $n^{-3/5}$ in the endpoints.

PETRONE (1999) suggested Bayesian methods using a Bernstein polynomial prior with alternative weight functions $h_n^{(m)}$ and GHOSAL (2001) derived the asymptotic results that if the real density f is a Bernstein polynomial, then the convergence rate of the posterior is equal to $n^{-1/2} \log n$, otherwise this rate is $n^{-1/3}(\log n)^{5/6}$ (under the same weak assumptions as in VITALE, 1975).

3.6 Improvements through *U*-statistic symmetrization

How to improve f_n, that is the key question of this section. Theoretical arguments behind B_n, b_n and, hence, f_n are spelled out in Section 3.2 and DE BRUIN ET AL. (1999). A practical argument in favor of these estimates is the strict positivity of b_n. This should certainly not be given up.

A breakthrough appeared when we tried to compute B_n, b_n and f_n for large values of n. To overcome numerical difficulties it was decided to split the sample into two subsamples of size $\frac{1}{2}n$ and to take the arithmetic average of these estimates. This average turned out to be smoother and more accurate than the estimate based on the entire sample. This experience is in line with the asymptotic distribution of f_n and it formed the basis of the U-statistic symmetrization explained below, after some preparations.

A natural way of smoothing is to partition the sample into $k = m^{-1}n$ sub-samples of size m each (for the sake of simplicity, assume $n = km$), and to take the arithmetic average of the resulting estimates: for each subsample $y_1^{(h)}, \ldots, y_m^{(h)}, (h = 1, \ldots, k)$ the density estimate $f_{m,h}$ is derived from

$$B_m(p|y_1^{(h)}, \ldots, y_m^{(h)}), \quad h = 1, \ldots, k$$

in the same way as f_n was derived from B_n. Define $\bar{f}_{n,m}$ as the average of the k estimates $f_{m,h}$. The approximation

$$\mathcal{L}\, m^{1/4}(f_{m,h}(x) - f(x)) \approx \mathcal{N}(0, \sigma^2(x))$$

with $\sigma(x)$ defined in the beginning of Section 3.5, implies that

$$\mathcal{L}\, m^{1/4}(\bar{f}_{n,m}(x) - f(x)) \approx \mathcal{N}(0, k^{-1}\sigma^2(x)) = \mathcal{N}(0, n^{-1}m\sigma^2(x))$$

or, equivalently, that the standard error of $\bar{f}_{n,m}(x)$ is $m^{-1/4}k^{-1/2}\sigma(x)$, which equals $m^{1/4}n^{-1/2}\sigma(x)$, i.e. $k^{-1/4}$ times that of $f_n(x)$. This suggests that the standard error is reduced by 15% or more if $(\frac{m}{n})^{1/4} < .85$ or, equivalently, if m is less than $.53n$ (see the remark in the previous section). Note that a special case appears when the sample is split into $m = \sqrt{n}$ equal parts, the resulting estimate $\bar{f}_{\sqrt{n},\sqrt{n}}$ is consistent and converges with rate $n^{-3/8}$ (if $f = \psi$).

To remove the permutation dependence (and the assumption that $k, m \in \mathbb{N}$), a U-statistic symmetrization will be applied. It will have the effect that the degree n of the polynomial expression $b_n(p)$ is lowered to the degree m of the estimate $b_n^{(m)}(p)$ to be derived. The positivity of the quantile density estimate will not be affected.

Let $B_m(p|y_1, \ldots, y_m)$ denote the polynomial approximation introduced in the previous section, applied to the m observations $y_1, \ldots, y_m$ (with $y_i = \Psi(x_i)$). Note that the degree of this polynomial is $m + 1$. The U-statistic

$$B_n^{(m)}(p) = \binom{n}{m}^{-1} \sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} B_m(p|y_{\alpha_1}, \ldots, y_{\alpha_m})$$

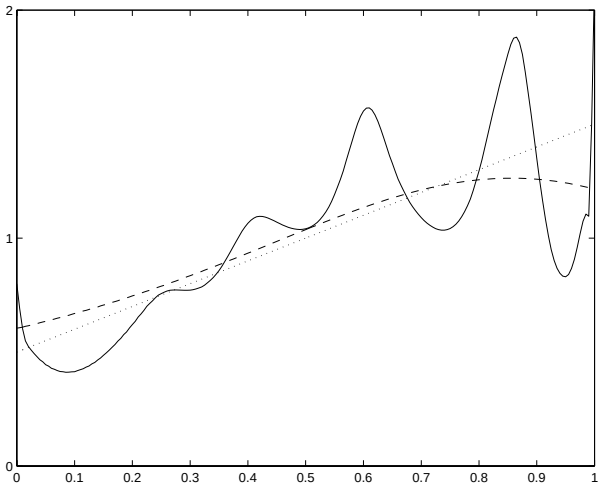


Figure 3.3: Visualization of the improvement discussed in Section 3.6.

is the estimate of $B(p)$ of interest. It can be rewritten as the L-statistic

$$p^{m+1} + \sum_{j=1}^{m} \binom{m+1}{j} p^j (1-p)^{m+1-j} \sum_{i=j}^{n-m+j} \frac{\binom{i-1}{j-1}\binom{n-i}{m-j}}{\binom{n}{m}} y_{[i]}.$$

So, $B_n^{(m)}$ is the average of the quantile function estimates based on all size-m subsets of the sample. From this we get $F_n^{(m)} = {B_n^{(m)}}^{-1} \circ \Psi$, $f_n^{(m)} = {F_n^{(m)}}'$, etcetera. For $m \to \infty$, the $B_m(p|\ldots)$ are consistent estimators of $(\Psi^{-1} \circ H)(p)$. For m fixed, $B_n^{(m)}(p)$ is an unbiased estimate of $\mathbf{E}\,(B_m(p|X_1, \ldots, X_m))$ which, of course, depends on the underlying distribution function of X_1. In principle the (exact) theory of HOEFFDING (1948) about U-statistics is applicable but the complexity of the formulas and the dependence on the unknown f, F, H, etcetera, of the asymptotic variances precludes application in practice. It is reasonable to conjecture that the rate of convergence is better than $n^{-3/8}$ if $k = n^{1/2}$ is taken.

The bias of $f_n^{(m)}$ will be larger than that of the original estimator $f_n(x)$. For a good comparison between this and other methods, one has to look at, e.g., the Mean (Integrated) Squared Error. We rely on the computational experiences to be discussed in Section 3.7, because theoretical analyses are both too complicated and too much dependent on 'irrelevant asymptotics'. In practice, m is chosen to be less than, say 10 or 20, and asymptotic theory becomes questionable in such cases.

The example in Figure 3.3 illustrates the errors involved in using f_n and $f_n^{(m)}$ and suggests that improvement indeed is gained by lowering the degree of the estimating

	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.000	.050	.100	.150	.200	.250	.300	.350	.400	.450	.500
1	.025	.043	.067	.100	.132	.165	.196	.233	.272	.299	.336
2	.040	.047	.057	.078	.100	.124	.146	.177	.210	.225	.256
3	.051	.051	.057	.068	.086	.103	.119	.146	.176	.181	.208
4	.059	.056	.060	.067	.080	.094	.104	.128	.156	.155	.179
5	.066	.006	.063	.068	.078	.091	.096	.118	.144	.137	.160
6	.072	.064	.067	.070	.079	.090	.093	.112	.136	.125	.147
7	.078	.068	.070	.073	.080	.090	.091	.108	.130	.117	.139
8	.082	.071	.073	.075	.081	.092	.091	.107	.127	.113	.133
9	.086	.074	.076	.078	.084	.094	.092	.106	.125	.110	.130
10	.090	.077	.078	.081	.086	.096	.093	.107	.124	.108	.128
11	.094	.080	.081	.084	.088	.098	.095	.107	.124	.107	.127
12	.097	.083	.084	.087	.090	.100	.097	.108	.124	.107	.126
13	.100	.086	.086	.089	.092	.103	.099	.109	.124	.108	.125
14	.103	.089	.089	.092	.094	.105	.101	.110	.125	.108	.125
15	.105	.091	.091	.094	.096	.107	.103	.110	.126	.109	.125

Table 3.2: Average total variation distances for 50 replications, for different choices of a (horizontal axis) and m (vertical), with a sample size of $n = 100$. Underlined are the optimal choices for m.

polynomial. A sample of size $n = 100$ is taken from the distribution on $(0, 1)$ with density $f(x) = \frac{1}{2} + x$ (the dotted line). Using the uniform density as initial guess, the densities $f_n^{(m)}$ $(m = 0, \ldots, 100)$ were computed and those corresponding to $m = 6$ (dashed line) and $m = 100$ (solid line) are displayed.

3.7 Simulation studies

In the current context it is obvious that m should depend on the sample size and on the reliability of the initial guess. To investigate the structure of this dependence, extensive Monte Carlo simulations have been performed. Results will be presented and discussed in this section.

Given some density f, consider (estimates of) the expected value of the total variation distance $||f_n^{(m)} - f||_1$ as a function of m. The optimal m, i.e. the value for which this expected L_1-distance is minimum, depends on the sample size and on the reliability of ψ. This optimal m is determined using the special densities

$$f_a(x) = (1 - a) + 2ax \quad |a| \leq 1, 0 \leq x \leq 1,$$

and the standard-uniform density $\psi(x) \equiv 1$ as initial guess. Figure 3.3 provides a visualization for $a = \frac{1}{2}$ (for explanation, see the last paragraph of the previous section).

	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
25	0	0	1	2	2	3	4	5	5	6	7
50	0	1	2	3	4	5	6	7	8	9	10
75	0	1	2	3	4	6	7	9	10	11	11
100	0	1	2	4	5	7	8	10	11	13	14
150	0	2	3	5	7	8	10	12	14	16	17
200	0	2	4	6	8	10	12	14	16	18	20
250	0	2	4	6	9	11	13	15	17	19	21

Table 3.3: Optimal m for the cases f_a with various a (horizontal axis) and n (vertical).

Computations were performed for a variety of values of n, as well of a. Variations in a lead to variations in the reliability of the initial guess through the relation $||f_a - \psi||_1 = \frac{1}{2}|a|$.

Table 3.2 displays results involving the choice $n = 100$. For all a in the table, a random sample of size 100 is drawn from f_a. Density estimates $f_{100}^{(1)}, \ldots, f_{100}^{(15)}$ are constructed, and total variation distances $||f_a - f_{100}^{(\cdot)}||_1$ are calculated. This is repeated 49 times, and reported are, for each pair (m, a), the average L_1-errors for the $N = 50$ replications. For example, for $f_{.4}(x) = .6 + .8x$ and a sample of size 100, a choice of $m = 5$ seems best, since the average total variation distance is lowest for $m = 5$.

Similar analyses have been performed for a variety of other choices of n. For each pair (n, a) in Table 3.3, a random sample (of size n) is drawn from f_a. For all possible values of m $(m = 0, 1, \ldots, n)$, $||f_a - f_n^{(m)}||_1$ is calculated. This process is replicated 249 times and Table 3.3 displays those m for which the average total variation distance between true and estimated density was smallest. These simulations suggest that the optimal m is approximately proportional to $n^{1/2}$ and $||f_a - \psi||_1$. The expression $m^* = 1.3n^{1/2}a = 2.6n^{1/2}||f_a - \psi||_1$ seems reasonable. The densities f_a studied were rather regular with only one sign change of $f_a - \psi$.

Note that the quantile densities corresponding to the f_a are not polynomials. In case $f_{1/2} = \frac{1}{2} + x$ we have a corresponding quantile density $2/\sqrt{1 + 8p}$, which is not a polynomial though, of course, (low degree) polynomial approximations can be fairly accurate. To study the choice of m in alternative situations, a second simulation study was carried out, involving five examples, all with the unit interval as support, $\psi(x) = 1$, and $||f - \psi||_1 = .25$. The examined densities are

$$\begin{aligned} f_{\mathrm{I}}(x) &= .5 + x \\ f_{\mathrm{II}}(x) &= 1.5 - |2x - 1| \\ f_{\mathrm{III}}(x) &= 1.616 - 2.464x(1 - x^2) \\ f_{\mathrm{IV}}(x) &= 1.091 - .427\sin(4.712x) \\ f_{\mathrm{V}}(x) &= .741 + 1.301\ \mathrm{e}^{-5x}, \end{aligned}$$

m	0	4	5	6	7	8	10	15	20	30
f_{I}	.250	.096	.091	.089	.089	.090	.093	.097	.106	.122
f_{II}	.250	.182	.167	.155	.145	.138	.129	.122	.124	.134
f_{III}	.250	.184	.168	.156	.146	.138	.128	.123	.125	.134
f_{IV}	.250	.147	.138	.132	.127	.124	.119	.116	.119	.130
f_{V}	.250	.104	.096	.092	.090	.089	.090	.098	.107	.124
theor.	.250	.102	.108	.113	.117	.121	.128	.141	.152	.169

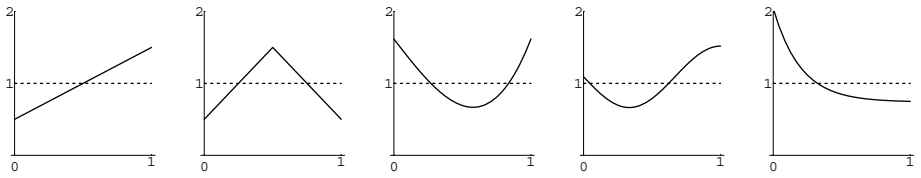
Table 3.4: Performance of $f_{\mathrm{I}}, \ldots, f_{\mathrm{V}}$ and a theoretical approximation. Displayed values are $\frac{1}{R}\Sigma||f_{\cdot} - f_{\cdot,n}^{(m)}||_1$, the underlined values are the minima for each f.

see Figure 3.4. We believe that these densities are 'representative' for the distributions occurring in practice. Analyzing these five densities will give insight in the effect of the type of density on the optimal m. For cases I and V, deviations $\psi(x) - f(x)$ are more regular (involving only one crossing of 0) than for the cases II to IV where two crossings occur. This interpretation of 'representativity' is in line with NEYMAN (1937) where 'smooth alternative probability density functions were [...] presented [...] in terms of probability density functions that have few intersections with the null probability density function' (RAYNER AND BEST, 1989, p. 45). It is reasonable to expect that the optimal m is smallest in cases I and V.

To quantify the optimal m in these 5 cases, a sample of size $n = 100$ is drawn from each density, and the corresponding $f_{100}^{(m)}$ and L_1-errors $||f_{100}^{(m)} - f||_1$ are computed for all possible values of m ($m = 0, 1, \ldots, 100$). This process is performed $R = 500$ times. The average L_1-error, for a deliberate choice of values of m, is reported in Table 3.4. The last row of this table contains the values $.72m^{1/4}n^{-1/2}$, mentioned in Table 3.1. These values provide a crude approximation to the expected L_1-error, especially those for the optimal m (the italicized ones), since they approximate the average L_1-error of the obviously less accurate estimates $\bar{f}_{n,m}$ studied in Section 3.6. The case $m = 0$ involves a degeneracy. If $m = 0$ is actually used, then the value .250 appears as the constant true value.

It follows from the first row of Table 3.4 that, in case $f = f_{\mathrm{I}}$, the expected L_1-error is minimum if m is 6 or 7 (the fourth row of Table 3.3 provides the value 7, take

Figure 3.4: From left to right: densities f_{I} up to f_{V}, discussed in Section 3.6



$a = \frac{1}{2}$). The second, third and fourth row provide that m should be about twice as large if the behaviour of f is less regular. Row five shows, as expected, that m must be chosen substantially smaller if these f's are replaced by the more regular one f_{V}. Both simulation studies taken together indicate that m should depend on the sample size n, the value v of $||f - \psi||_1$ where optimality is required, and the (expected) regularity of f. The latter will be quantified by the symbol w, where w is equal to 1 if the number of sign changes of $\psi - f$ is equal to 1, like in the cases of f_a, f_{I} and f_{V}, and is equal to 2 if two or more sign changes are occurring (cases f_{II} to f_{IV}).

Conclusion
Our simulation studies suggest that the optimal m is approximately equal to

$$m^* = 2.6\, n^{1/2}\, v\, w$$

Note that $m^* = 6.5$ in the case $n = 100$, $v = .25$, $w = 1$ studied in the fourth row of Table 3.3 and in the first row of Table 3.4. The choice $m^* = 13$ is indicated for f_{II}, f_{III} and f_{IV}. In practice f is unknown and the research worker will have to choose the value v of $||f - \psi||_1$ and the value $w \in \{0, 1\}$ where 'optimality is required'.

Robustness
As a final remark, we mention that $||f_n^{(m_1)} - f_n^{(m_2)}||_1$ is very small if m_1 and m_2 are 'not much different', as is exemplified in Figure 3.5 and in Table 3.5. This is, of course, of considerable importance, because it would have been unfortunate if minor variations in the smoothing parameter m induce major variations in the resulting density estimate.

Figure 3.5 indicates this: using one random sample of size $n = 100$ from the Beta$(3, 2)$-distribution and the initial guess $\psi \sim$ Beta$(3, 3)$, density estimates $f_{100}^{(9)}, \ldots, f_{100}^{(13)}$ are constructed. These estimates with smoothing parameters $m = 9, \ldots, 13$ are 'practically identical' in the light of the true values which are $O(1)$.

Another validation of our suggestion is in Table 3.5. A random sample of size $n = 100$ is drawn from $f(x) = \frac{1}{2} + x$ on $(0, 1)$, with $\psi(x) = 1$. On basis of this sample, estimates $f_{100}^{(m)}$, with m=4, 6, 7, 8, 12, 20 are constructed. This process is repeated 249 times, the averages of the R=250 L_1-distances $||f_{100}^{(m_1)} - f_{100}^{(m_2)}||_1$ are shown.

Table 3.5: Indication of the robustness. See the text for full details.

	4	6	7	8	12	20
4	.000	.025	.033	.041	.063	.090
6		.000	.009	.017	.039	.068
7			.000	.008	.031	.060
8				.000	.023	.054
12					.000	.031

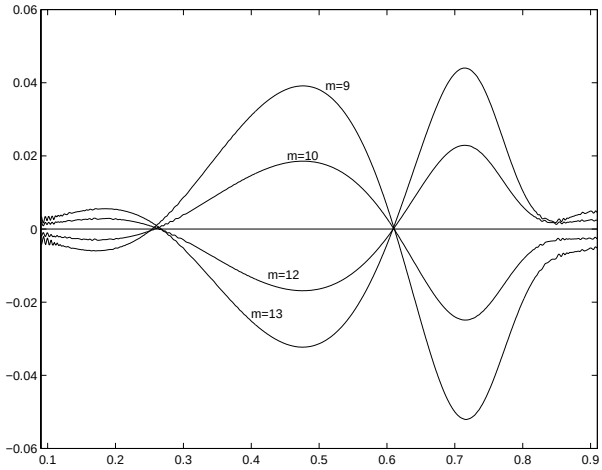


Figure 3.5: Indication of the robustness of the choice of m. Displayed are the differences $f_{100}^{(9)} - f_{100}^{(11)}, \ldots, f_{100}^{(13)} - f_{100}^{(11)}$.

Another desired robustness property is that minor deviations in *initial guess* should not lead to major variations in the induced density.

3.8 Specifying the initial guess

Application of the previous sections requires the specification of an initial guess ψ and the choice of m. Our theory is based on the assumption that the initial guess ψ of the true but unknown density f is made a priori. This assumption is not necessarily unrealistic. Suppose, for example, that the distribution under 'standard conditions' is fairly well known and that the distribution under certain non-standard conditions is required. If the deviations are expected to be fairly small, then the distribution under standard conditions can be used as initial guess.

In other situations some 'data peeping' cannot be dispensed with. A standard situation is that where the distribution has support $(a, b) = (-\infty, \infty)$ and classical statisticians like to make normality assumptions but do not have reliable prior information about expectation μ and variance σ^2. These true values will then be estimated from the sample, preferably by using the sample mean $\bar{x}$ and the sample variance s^2 because the underlying estimators are best unbiased if the normality holds.

The theory of nonparametric density estimation is motivated by the idea that, in practice, parametric models like that discussed above, are never exactly true. They may provide useful approximations but one would like to adapt these, especially if the data suggest that the assumptions are false. That is what we tried to do in the preceding sections by 'fine-tuning' the initial guess ψ.

Though one should not forget that the theory was based on the assumption that the initial guess was made up a priori, it is a matter of 'good statistical practice' to discuss some methods to have a look at the data and to adapt the initial guess if necessary. We already discussed the possibility to use the $\mathcal{N}(\bar{x}, s^2)$ distribution as initial guess. It is well known that this choice maximizes the Shannon entropy $-\int f(x) \log f(x)\, \mathrm{d}x$ under the restriction that f is a probability density on $(-\infty, \infty)$ with $\int x f(x)\, \mathrm{d}x = \bar{x}$ and $\int (x - \bar{x})^2 f(x)\, \mathrm{d}x = s^2$. This *maximum entropy* approach allows some generalization which we like to discuss.

Maximum Entropy approach

For explanatory works on maximum entropy, see JAYNES (1996), GOLAN ET AL. (1996), ZELLNER AND HIGHFIELD (1988) (and Section 2.4). Shannon's entropy is 'an axiomatic method to define a unique function to measure the uncertainty of a collection of events' (GOLAN ET AL., 1996). and is defined as the measure

$$H(f) \equiv -\int_a^b f(u) \log(f(u))\, \mathrm{d}u$$

in the continuous case. When no initial information is present other than the finiteness and specifications of a and b, the entropy measure reaches its maximum when $f(x) = 1/(b-a)\mathbf{1}_{(a,b)}(x)$. If $a = -\infty$ or $b = \infty$ (or both) then, obviously, $\max_f H(f) = \infty$. When a priori knowledge is presented in the form of postulated values for expectations such as $\int x f(x)\, \mathrm{d}x = \bar{x}$ and $\int x^2 f(x)\, \mathrm{d}x = s^2 + \bar{x}^2$, the maximum entropy approach will yield the distribution f that maximizes the entropy $H(f)$ subject to the constraints given by the knowledge presented. We already mentioned that the density of $\mathcal{N}(\bar{x}, s^2)$ appears. According to Jaynes, this distribution 'can be realized in the greatest number of ways consistent with what we know' (GOLAN ET AL., 1996).

Example

Suppose the following information is available. The support of the unknown density of interest f is $(0, 10)$, and the first moment is $\mathbf{E}\ f(x) = 6.3$. We like to take the maximum entropy estimate of f as initial guess ψ. It is constructed by maximizing

$$-\int_0^{10} \psi(x) \log(\psi(x))\, \mathrm{d}x$$

subject to constraints

$$\begin{aligned}
\psi(x) &\geq 0 && \text{nonnegativity of } \psi \\
\int_0^{10} \psi(x)\, \mathrm{d}x &= 1 && \text{total probability 1} \\
\int_0^{10} x\psi(x)\, \mathrm{d}x &= 6.3 && \mathbf{E}\ \psi(x) = 6.3.
\end{aligned}$$

The 'MaxEnt' density $\psi(x) = \mathrm{e}^{\cdot 162x}/25$ obtained (by constructing the Lagrangian function, and setting the partial derivatives zero, see Section 2.4) may be used as initial guess of the density f.

Conclusion
The maximum entropy principle and, especially, the explicit way its solution can be obtained for a large variety of cases makes it a valuable tool (see, e.g., GOLAN ET AL., 1996, JAYNES, 1996, and ZELLNER AND HIGHFIELD, 1988). In practice, however, if one has a distribution at some interval, say $[0, 1]$, with its first 2 moments prescribed then a standard family, say $\beta(f, g)$, may also be used (it would correspond to the maximum entropy method if $\beta_1(u) = \log u$ and $\beta_2(u) = \log(1-u)$, and $\mathbf{E}(\log X)$ and $\mathbf{E}(\log(1-X))$ are prescribed). On Jaynes and maximum entropy, ZABELL (1988) states:

> 'The right-wing totalitarians [...] believe that once an axiom system is adopted, one must accept without question every consequence that flows from it. One searches within one's heart, discovers the basic properties of belief and inference, christens them axioms, and then all else follows as logical consequence. [...] One representative of this position is E.T. Jaynes, who dates his adherence to Bayesianism to the time when he encountered Cox's axiomatization of epistemic probability, and who views the Shannon axioms for entropy as an unanswerable foundation for his method of maximum entropy. This position errs in giving the axioms too distinguished a position.'

What if the true density is normal?
Returning to the general issue of specifying the 'inputs' (ψ and m) of our nonparametric density estimate (Section 3.6) and restricting the attention to $a = -\infty$ and $b = \infty$, the idea of using a normal density as initial guess sounds plausible.

The sample values $x_{[1]}, \ldots, x_{[n]}$ can be used to construct an initial guess Ψ in the form of a normal distribution with parameters estimated from the sample. The density of $\mathcal{N}(\bar{x}, s^2)$ seems natural in this respect, though alternative estimates[6] are not necessarily worse. The issue of interest is whether our estimates $f_n^{(m)}$ provide a satisfactory method for fine-tuning such normal a priori guess. The choice of m is of particular interest in this case. To make this choice, a specification of $v = ||f - \psi||_1$ is needed. It is interesting to study the L_1 distance if f *truly* is a $\mathcal{N}(\mu, \sigma^2)$ density and ψ is its estimate based on $\hat{x}$ and s^2. In practice the true density f will not be of this normal kind and, hence, the values of m to be suggested for the normal case, should be regarded as lower bounds of the choice of m to be made in practice.

A partial analytic result is provided by using

$$\begin{aligned} ||\psi - f||_1 &= ||\mathcal{N}(\bar{x}, s^2) - \mathcal{N}(\mu, \sigma^2)||_1 \\ &\leq ||\mathcal{N}(\bar{x}, s^2) - \mathcal{N}(\mu, s^2)||_1 + ||\mathcal{N}(\mu, s^2) - \mathcal{N}(\mu, \sigma^2)||_1, \end{aligned}$$

[6] e.g. such that $x_{[1]}$ corresponds to the $(n+1)^{-1}$th quantile and $x_{[n]}$ to the $1-(n+1)^{-1}$th quantile, or such that a normal density is fitted to the quantiles through Normal Probability Plot Regression (see Chapter 5). Of particular interest in our context are the estimators $\hat{f}$ that, given the data, try to minimize $||f - \hat{f}||_1$. We, however, content ourselves with $\mathcal{N}(\bar{x}, s^2)$ as indicated.

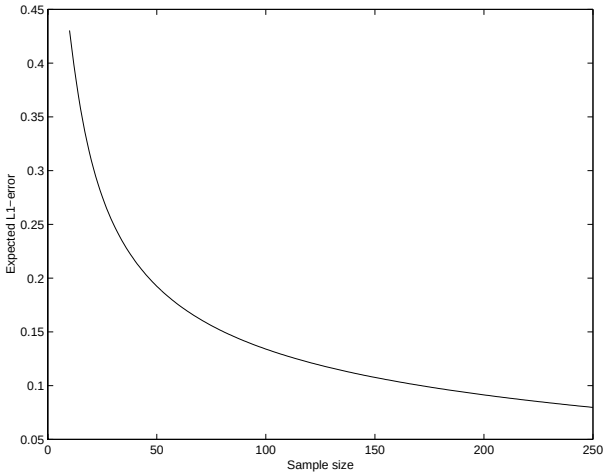


Figure 3.6: Expected L_1-error versus sample size

on account of the triangular inequality, and it is, theoretically, possible to obtain explicit results. We do have the following partial results: in case of unknown μ and known σ^2, the exact result

$$\mathcal{L}\ ||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\hat{\mu},\sigma^2)||_1 = 2\Phi\left(\tfrac{|\hat{\mu}-\mu|}{2\sigma}\right)-1,$$

where, e.g., $\hat{\mu}=\bar{x}$, applies. A similar but more complicated formula holds for $||\mathcal{N}(\mu,s^2)-\mathcal{N}(\mu,\sigma^2)||_1$

$$\mathcal{L}\ ||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\mu,\hat{\sigma^2})||_1 = 2\left|\Phi\left(\frac{a(\sigma^2,\hat{\sigma^2})}{\sigma}\right)-\Phi\left(\frac{a(\sigma^2,\hat{\sigma^2})}{\hat{\sigma}}\right)\right|$$

where $a(\sigma^2,\hat{\sigma^2})=\sqrt{\log\frac{\hat{\sigma^2}}{\sigma^2}\frac{\hat{\sigma^2}\sigma^2}{|\hat{\sigma^2}-\sigma^2|}}$ and $\hat{\sigma^2}$ is, e.g., the sample variance.

More useful might be the asymptotic approximation (BROWN, 1995)

$$\lim_{n\to\infty}\mathrm{P}\left(||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\hat{\mu},\hat{\sigma^2})||_1<\frac{2z_{1-\alpha}}{\sqrt{n\pi\mathrm{e}}}\right)=1-\alpha.$$

When, like in our case, both μ and σ^2 are unknown, BROWN (1995) provides the asymptotic inequality

$$\lim_{n\to\infty}\mathrm{P}\left(||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\hat{\mu},\hat{\sigma^2})||_1<\sqrt{\frac{-2\log\alpha}{n}}\right)\geq 1-\alpha.$$

n	10	15	25	50	75	100	250	500	1000
$v_{\min}$	1.39	1.13	.86	.60	.49	.43	.27	.19	.14

Table 3.6: Expected L_1-error versus sample size

and establishes that this holds for *all* two parameter densities with both parameters unknown.

A precise, exact, and explicit result is not attainable. That is why we will establish a value for the total variation numerically (which is quite easily possible).

Taking, without loss of generality, $\mu = 0$ and $\sigma^2 = 1$, standard sampling theory provides $\bar{X} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ and $(n-1)S^2 \sim \chi^2_{n-1}$, ($\bar{X}$ and S^2 being statistically independent). We computed $||\mathcal{N}(0,1) - \mathcal{N}_(a,b)||_1$ for all (a,b) in a sufficiently precise lattice on $\mathbb{R}$. Following that, we computed $\mathbf{E}_{a=\hat{X}} \mathbf{E}_{b=S^2} ||\mathcal{N}(0,1) - \mathcal{N}(a,b)||_1$. See Figure 3.6 and Table 3.6 for a display and note that the expected L_1-error decreases when the sample size increases. These results are very important. Note that in this case where the initial guess ψ corresponds to a normal density estimated from the data, we have the peculiarity that $v \approx 4.3n^{-1/2}$. The crucial input v in the rule of thumb $m = 2.6\sqrt{n}\ v\ w$ (Section 3.10) should depend on n as well. This provides a lower bound $11w$ for m, independent of the sample size. Recall that $w = 1$ if only one zero of $\psi - f$ is expected, and 2 if the behaviour is less regular Two normal densities, in general, intersect at two points, hence $w = 2$ seems the appropriate choice.

These lower bounds for m and v correspond to the situation where f is truly normal. The researcher has to 'add a value on top of v, or m, describing his confidence in the notion that f is normal'.

3.9 Extending the theory to the bivariate case

It is not straightforward to generalize our univariate density estimate $f_n^{(m)}(x)$ to some bivariate estimate $f_n^{(m)}(x,y)$. The reason is that the concept of quantile distribution does not translate easily to the bivariate case. Associated to this in a semi-parametric way are *copula functions*. For the joint distribution H of X and Y with marginals F and G, there is a so called copula function $C : [0,1]^2 \to [0,1]$ such that

$$H(x,y) = C(F(x), G(y)), \quad \forall x, y.$$

References in this field are SCHWEIZER AND SKLAR (1983), MARSHALL (1996), and SKLAR (1996). ALBERS (1998) developed a semi-parametric model to estimate bivariate densities under the assumption of some form of positive association, e.g. positive quadrant dependence (see LEHMANN, 1966), based on the techniques in DE BRUIN ET AL. (1999). It is possible to construct a copula such that the $H(x,y)$ inferred is indeed positive quadrant dependent, e.g. $H(x,y) = \min(F(x), G(y))$ (MARSHALL,

1996). This method was not very successful, partly due to the wiggling behaviour of the (marginal) density estimates, and partly because the over-simplification in taking only one parameter to model dependence.

Another approach is that of TENBUSCH (1994) who extended the uni-variate Bernstein polynomial estimation method of VITALE (1975, see Section 3.5) to the bivariate case. When (X, Y) is defined on the unit square, the density estimate is

$$\tilde{f}_n^{(m)}(x,y) = \frac{(m+1)^2}{n} \sum_{i=0}^{m} \sum_{j=0}^{m} h_n^{(m)}(i,j) \binom{n}{i} x^i (1-x)^{m-i} \binom{n}{j} y^j (1-y)^{m-j}$$

with $h_n^{(m)(i,j)} = \#\{(x_i, y_i) \in [\frac{i}{m+1}, \frac{i+1}{m+1}] \times [\frac{j}{m+1}, \frac{j+1}{m+1}]\}$. Tenbusch claims that for this method the quality of the estimators is comparable with the quality of the usual bivariate kernel estimators. A drawback of his method is that it is seriously affected by order-preserving transformations of the data.

To overcome the drawbacks of Albers's original method and of Tenbusch's method, the following approach might be of interest. Consider that the problem is reduced to that of estimating the densities of a large number of well-chosen linear combination. To be specific, if the bivariate data suggests a normal initial guess, then we would prefer to transform the data linearly such that the initial guess corresponds to the $\mathcal{N}_2(0_2, I_2)$ distribution. Next consider J linear combinations

$$(\bar{x}_i, \bar{y}_i) \mapsto (\cos\varphi_j)\bar{x}_i + (\sin\varphi_j)\bar{y}_i \quad \text{where } \varphi_j = j\tfrac{\pi}{n}, \quad j = 0, 1, \ldots, J-1, \quad \forall i.$$

For each linear combination $\phi = \mathcal{N}(0,1)$ is used as the initial guess. Next, applying the procedures of this chapter, we obtain J estimated densities for these linear combinations. These marginal densities can then be combined into one estimate $h_n^{(m)}(x,y)$ of $h(x,y)$, e.g., by using numerical methods familiar to MRI-scan specialists.

3.10 Discussion

The nonparametric density estimates f_n and $f_n^{(m)}$ require that an initial guess ψ of the true density f is made. In principle, the initial guess should be made *a priori* because the theory is based on this assumption. In practice some data peeping may be necessary, see Section 3.8. The subjectivity involved in the specification of the initial guess remains visible if m is small. If one chooses $m = n$ such that $f_n^{(n)} = f_n$ corresponds to the estimator studied in DE BRUIN ET AL. (1999) , then the data speaks almost exclusively and the information provided by ψ is almost completely ignored. That is why f_n cannot compete with estimates that take more information into account. The situation becomes more interesting, but also more complicated, if one accepts the idea that ψ is 'reliable' in the sense that the difference $||\psi - f||_1$ between ψ and the true but unknown f may be expected to be less than some constant c which is substantially smaller than the upper bound 2. If, e.g., one is willing to believe that $||\psi - f||_1$ is less than .50 then one will try to choose m such that m is

'optimal' if $v = ||\psi - f||_1$ is .25. The rule of thumb $m^* = 2.6n^{1/2}\ v\ w$ mentioned at the end of Section 3.6 will then provide $m^* = 6.5w$ if $n = 100$ and, hence, $m^* = 6.5$ if one sign change of $\psi - f$ is expected. This implies that the 'uniqueness' of f_n is preserved, at least to some extent. This, however, depends on the belief that v is something like .25. In an interesting case from practice (see Section 3.8), values of v 'not less than $4.3n^{-1/2}$ are suggested which in the case $n = 100$ provides that $v \geq .43$ and $m^* = 11$, respectively 22, if $w = 1$, respectively 2. The difference between the m^* values suggested is less important than one might expect. Table 3.5 suggests that the L_1-errors are more affected by the sample one has to evaluate than by the degree m of the estimate $b_n^{(m)}$ of the quantile density.

The 'unique' nonparametric density estimate $f_n^{(m)}$ thus defined has an expected L_1-error which is considerably smaller than that of f_n if the true density f is not too much different from the initial guess ψ. It is natural that $f_n^{(m)}$ will, under these conditions, be a better estimate than any kernel estimate if the bandwidth is based on likelihood cross-validation (because of the additional knowledge incorporated in $f_n^{(m)}$). This was supported by simulations. The situation may change if the kernel and bandwidth are also chosen on the basis of the initial guess ψ. It will depend on the true density f whether the estimate $f_n^{(m)}$, with $m = 2.6n^{1/2}v\ w$ (or if ψ has been obtained by a preliminary parametric analysis of the data, $m = 10w$), has the smallest expected L_1-error or some kernel estimate with a bandwidth determined such that it is optimal for $f = \psi$.

To summarize, we claim (on basis of our simulation experiments) that $f_n^{(m)}$ with $m = .65n^{1/2}w$ (the case $v = .25$), possibly with $w = 1$, is quite reasonable in practice and that the explicit and usable formula

$$m^{1/4}n^{-1/2}\hat{\sigma}(x) = \frac{(.65\ w)^{1/4}}{n^{3/8}} \frac{f_n^{(m)}(x)}{\sqrt[4]{4\pi F_n^{(m)}(x)(1 - F_n^{(m)}(x))}}$$

for the corresponding standard error is not unreasonable.

The theory of this chapter is of particular interest because it tries to combine that what is supposed to be 'good' in the schools of the classical statisticians, the nonparametric statisticians and the Bayesians. Such *multi-modal approach* is always somewhat questionable because arguments can be weighted differently. A peculiar feature is that instead of a restrictive Bayesian a priori distribution, a less demanding initial guess is exploited. This suggests that a relatively new type of statistics is involved.

Chapter 4

A goodness of fit test, smoother than smooth

> *'There are two points which must be borne in mind when deciding upon the order of the test to be applied in any particular case. This, of course, is a question entirely beyond the limits of statistical theory which concerns itself only with the properties of different tests. When these properties are elucidated it is a problem for the practical statistician to choose the test, the properties of which correspond most closely to the circumstances of the problem he considers. My personal feeling is that in most practical cases, there will be no need to go beyond the fourth order test. But this is only and* opinion *and not any mathematical result.'*
>
> J. Neyman[1]

The theory of the previous chapter will be used to construct a test for the null hypothesis H_0: $f = \psi$. For such a test an estimate $\hat{f}$ of f is compared with ψ. The total variation distance between $\hat{f}$ and ψ is used as test statistic. As $\hat{f}$ we will use $f_n^{(m)}$ defined in Chapter 3. The choice of the smoothing parameter m is discussed and the distribution of the test statistic under H_0 is studied theoretically by using asymptotic theory, and numerically by performing simulation experiments to specify critical values. Relations with other proposals from the literature are discussed. Most results of this chapter also appear in Albers and Schaafsma (2003b).

4.1 Goodness of fit testing

After Karl Pearson's breakthrough paper about the χ^2 test was published in 1900, many improvements were made and modifications were suggested. Neyman (1937), for example, considered continuous analogues of Pearson's problem. We concentrate the attention on such analogue.

Problem
Given are the outcomes $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ of an independent random sample $X_1, \ldots, X_n$ from a probability distribution on $\mathbb{R}$ with a 'smooth' density f, 'not unlike a given density ψ'. *Required* is a statement about the truth or falsity of the hypothesis of equality: $H_0 : f = \psi$.

[1] 'Smooth' test for goodness of fit, *Skandinavisk Aktuarietidskrift*, **20**, 1937.

The statistician who has to solve this problem may be appalled by the abundance of proposals. PEARSON's test (1900) requires a classification of the data, the choice of the number of classes in particular, as well as the class borders. NEYMAN's smooth test (1937) provides the freedom (but also the task) to specify an orthonormal basis for an L_2 space, e.g. the normalized Legendre polynomials. KOLMOGOROV's test (1933)[2] is yet another possibility. The freedom to choose has a dark side, which we like to illustrate as follows. An investigator of human growth used Kolmogorov's test to see whether a normal distribution could explain his data. The test did not lead to rejection of the hypothesis. Subsequent analyses were made under normality assumptions. The investigator was criticized because an application of Pearson's χ^2 test displayed the existence of significant deviations from normality. This illustration by means of an example from practice can, of course, be supported by theoretical power considerations from a Neyman-Pearson perspective. Maximum shortcomings of level-α tests are equal to $1 - \alpha$ for a variety of nonparametric alternatives. In spite of this discomforting background, we like to study the problem because of its relevance. The present-day statistician may choose a computational approach by using some estimate of f, e.g. a kernel estimate. He will then compare the estimate and the postulated density ψ. Our method fits within this approach: we concentrate the attention on the idea to reject H_0 if the area $||\hat{f} - \psi||_1$ between the graph of ψ and that of a special nonparametric density estimate $\hat{f}$ is too large. This estimate $\hat{f}$ will be constructed by adapting the theory of Chapter 3 to the present situation, see Section 4.2. The null distribution of the test statistic $||\hat{f} - \psi||_1$ is studied to determine P-values and tables of critical values. The approach can be embedded in the general theory of goodness of fit tests which includes the testing of parametric models, e.g. that where ψ is a normal density with unspecified parameters, and the testing of nonparametric models, e.g. that f is a concave function on some finite interval or that $\log f$ is a concave function (strong unimodality), see, e.g. DÜMBGEN AND SPOKOINY (2001).

Another alternative to 'standard goodness of fit testing theory' is to use *pre-test procedures*, which may be described as follows

> 'Suppose we want to test a given hypothesis, but we doubt whether we may assume a restricted, but possibly incorrect model, or have to resort to a larger and thus less precise model. To settle this issue, we perform a preliminary test on the adequacy of the restricted model. If this test fails to reject, then we feel free to stick to the restricted model and use a test specifically suitable for that model. Otherwise, we use an alternative main test which is more general and appropriate for the large model, but less powerful than the first test when the restricted model holds.'

(BOON, 1999, see also W. ALBERS ET AL. ,2000, 2001). We shall not go into further details and restrict ourselves to 'standard' goodness of fit problems.

[2]In literature also referred to as the Kolmogorov-Smirnov test, following SMIRNOV's extension (1939)of Kolmogorov's test to the two-sample problem.

The estimate $\hat{f}$ to be used in our approach depends on the sample size n and on the degree m of a polynomial to be specified. That is why the notation $\hat{f} = f_n^{(m)}$ is used, together with $t_n^{(m)} = ||f_n^{(m)} - \psi||_1$ for the outcome of the test statistic $T_n^{(m)}$. The P-value $\mathrm{P}_0(T_n^{(m)} \geq t_n^{(m)}) = \alpha(x)$ will be used as degree of belief in H_0. Here P_0 refers to the distribution of $T_n^{(m)}$ under H_0. The complement $\varphi(x) = 1 - \alpha(x)$ of $\alpha(x)$ is the degree of belief in the falsity of H_0. SALOMÉ ET AL. (1999) suggests that the degree of belief interpretation is questionable. If, however, H_0 is rejected for $\alpha(x)$ smaller than some nominal level, then one is acting according to the Neyman-Pearson theory. In practice this Neyman-Pearson approach is quite natural: subsequent analyses are based on the assumption that $f \equiv \psi$ if H_0 is maintained; if H_0 is rejected, then some estimate, in our case $f_n^{(m)}$, will be used. This in contrast to using, e.g., Pearson's test which has the drawback that 'if the null hypothesis is rejected then there is no alternative distribution indicated.' (RAYNER AND BEST, 1989, p. 33).

Similar to Section 3.2, the probability transform $x_i \to u_i = \Psi(x_i)$ is applied to the original observations giving us

$$u_{[0]} = 0, \qquad u_{[i]} = \Psi(x_{[i]}) \quad (i = 1, \ldots, n), \qquad u_{[n+1]} = 1,$$

as the basis of the analysis. Note that $\Psi(X_i)$ has distribution function $G = F \circ \Psi^{-1}$, quantile function $B = G^{-1} = \Psi \circ F^{-1}$, density function $g(u) = f(\Psi^{-1}(u))/\psi(u)$, quantile density $b(p) = B'(p)$, etcetera. The null hypothesis H_0: $f = \psi$ is equivalent to H_0: $g \equiv 1$ and to H_0: $b \equiv 1$.

The notations $u_i = \Psi(x_i)$ and, of course, $u_{[i]} = \Psi(x_{[i]})$ are used, instead of $y_i = \Psi(x_i)$ of the previous chapter, because

1. To emphasize that under the null hypothesis the u_i constitute a sample from the standard uniform distribution;
2. To emphasize the different context we are in now.

The example

Let the data $x_{[1]}, \ldots, x_{[20]}$ be given as:

3.89,	7.44,	8.65,	9.40,	10.00,	11.27,	11.52,
14.23,	15.52,	15.63,	16.39,	17.33,	18.37,	21.12,
21.76,	22.54,	23.29,	23.36,	24.17,	24.57.	

The information is provided that the underlying density f ($f(x) = (5^{-4}x + 50^{-1})\mathbf{1}_{[0,25]}$) is such that the support $\{x; f(x) > 0\} = (0, 25)$. We wish to test H_0: $f = \psi$ where $\psi(x) = .04$ is the density of the uniform distribution on the support indicated. Using $\Psi(x) = x/25$ as notation for the distribution function of this null distribution, it is natural to apply the probability transform where x_i is replaced by

$$u_i = \Psi(x_i) = \tfrac{1}{25}x_i.$$

The true distribution of $U_i = \Psi(X_i)$ has distribution function $G = F \circ \Psi^{-1}$ where $G(u) = F(25u)$ and

$$g(u) = \frac{f(\Psi^{-1}(u))}{\psi(\Psi^{-1}(u))} = 25f(25u).$$

We concentrate the attention on the formulations H_0: $g \equiv 1$ and especially H_0: $b \equiv 1$.

If a simple alternative is considered, e.g. H_1: $g(u) = 2u$, then we can apply the Neyman-Pearson Fundamental Lemma. For this special alternative H_1 we reject H_0 if $\prod_{i=1}^{n} u_i$ is sufficiently large, or, equivalently, if $-2\sum \log(u_i)$ is sufficiently small. It is well known that the distribution of $-2\sum \log(U_i)$ is χ^2_{2n} if H_0 is true. The P-value

$$\mathrm{P}\left(\chi^2_{2n} \leq -2\sum \log(u_i)\right) = \mathrm{P}\left(\chi^2_{40} \leq 21.62\right)$$

thus obtained, for the example, is equal to .0078. Hence H_0 is rejected at all levels of significance $\alpha \geq .0078$.

In practice we do not know which simple alternative to choose and we are in need of an omnibus test. Using the density estimate $f_n^{(m)}$, with smoothing parameter $m = 1$, we obtain $||f_{20}^{(1)} - \psi||_1 = |\bar{u} - \frac{1}{2}| = .141$ for the example mentioned, see Figure 4.2 (page 85, left) and Section 4.3. With this choice of m the distribution of the test statistic under H_0 is obvious. $U_1, \ldots, U_n$ being an independent random sample from $\mathcal{U}(0,1)$ and using a normal approximation we obtain the P-value

$$\mathrm{P}\left(\left|\tfrac{1}{n}\sum U_i - \tfrac{1}{2}\right| > .141\right) \approx 2\Phi\left(-.141\sqrt{20 \cdot 12}\right) = .029.$$

Note that Karl Pearson's test requires the specification of the number of cells such that the χ^2_k distribution applies. If we take $k = 1$, we arrive at the two-sided sign test which, for our data, provides P = .263. If we take $k = 2$, then we have to work with the exact null distribution of Pearson's statistic. Computations provided P = .14. About the possibility of a further increase of k, KALLENBERG ET AL. (1985) wrote

> 'In a classical paper by MANN AND WALD (1942), a rule is given to let k increase with n roughly at the rate $n^{2/5}$ when using intervals with equal probability under H_0. More recent numerical work, however, has shown that for particular alternatives, a small fixed value of k often gives much better power (cf. BEST AND RAYNER, 1981)'.

Besides the choice of the number of classes k, the class borders $[l_1, r_1), \ldots, [l_k, r_k)$ have to be specified. KENDALL AND STUART (1973) suggest the borders are chosen such that one obtains the 'equiprobable χ^2 test' with the probabilities $\mathrm{P}(l_i \leq X < r_i) = k^{-1}$, $(i = 1, \ldots, k)$ under H_0, because this test is locally unbiased, in case of a simple alternative (cf. BEST AND RAYNER, 1981). Furthermore, 'when not only the frequencies ν_i but also the original observations X_i are available, reduction of the data through grouping results in tests that tend to be less efficient than those based on the Kolmogorov or related statistics' (LEHMANN, 1986, p. 480).

4.2 Specifications

To test H_0: $f = \psi$, consider the area

$$||f_n^{(m)} - \psi||_1 = ||g_n^{(m)} - 1||_1 = ||b_n^{(m)} - 1||_1$$

between the graphs of ψ and $f_n^{(m)}$. Note that the first equality follows from the fact that the L_1-norm corresponds to the total variation norm which is invariant under bimeasurable bijections (cf. DEVROYE, 1987 and DUNFORD AND SCHWARTZ, 1957) . The second equality can be established by noting that, in general, $||b-1||_1$ is equal to

$$\begin{aligned}
||b-1||_1 &= \int_0^1 |B'(p)-1| \, \mathrm{d}p \\
&= \int_0^1 \left|(G^{-1})'(p)-1\right| \, \mathrm{d}p \\
&= \int_0^1 \left|\frac{1}{g(G^{-1}(p))}-1\right| \, \mathrm{d}p \\
&= \int_0^1 \left|\frac{1}{g(u)}-1\right| \, \mathrm{d}G(u) \\
&= ||g-1||_1.
\end{aligned}$$

The definition of $B_n^{(m)}$ (and, hence, of $b_n^{(m)}$, $g_n^{(m)}$, $f_n^{(m)}$, etcetera) was given in Section 3.6. Recall that the U-statistic $B_n^{(m)}$ can be written as the L-statistic

$$B_n^{(m)}(p) = p^{m+1} + \sum_{j=1}^{m} \binom{m+1}{j} p^j (1-p)^{m+1-j} \sum_{i=j}^{n-m+j} \frac{\binom{i-1}{j-1}\binom{n-i}{m-j}}{\binom{n}{m}} u_{[i]}.$$

Differentiation provides $b_n^{(m)}$. This is a convex combination of the densities of the Beta$(i+1, m+1-i)$ distributions $(i=0,\ldots,m)$, which, of course, are polynomials of degree m. Note that the $\hat{g} = g_n^{(m)}$ and $\hat{f} = f_n^{(m)}$ obtained are genuine probability density functions: they are nonnegative everywhere and integrate up to one. In the density estimation case it was suggested that m should be proportional to $\sqrt{n}$. In the present context of testing H_0: $b=1$ some further smoothing is indicated. We recommend a choice of m not larger than 4 or 5 if an omnibus test is required, see the end of Section 4.1.

Though we are primarily interested in using $T_n^{(m)}$ with outcome

$$t_n^{(m)} = ||f_n^{(m)} - \psi||_1 = ||b_n^{(m)} - 1||_1$$

as test statistic, one can discuss some other test statistics as well, e.g. the Kolmogorov distance with outcome

$$\tilde{t}_n^{(m)} = ||F_n^{(m)} - \Psi||_\infty = ||B_n^{(m)} - p||_\infty.$$

A comparison between tests based on different dissimilarity coefficients is not performed though it might be of interest.

4.3 The extreme case $m = 1$

Ignoring the degenerate case $m = 0$ where the smoothing is so strong that $B_n^{(0)}(p) = p$ does not depend on the data, we start with $m = 1$ where

$$B_n^{(1)}(p) = 2\bar{u}p(1-p) + p^2 = (1-\bar{u})p^2 + \bar{u}(2p-p^2)$$

is a convex combination of the quantile function $2p-p^2$ of the Beta$(1, \frac{1}{2})$ distribution and the quantile function p^2 of the Beta$(\frac{1}{2}, 1)$ distribution. This, of course, does not mean that the inverse $G_n^{(1)}$ of $B_n^{(1)}$ is a convex combination as well. Note that for $\bar{u} = \frac{1}{2}$ the uniform distribution appears.

Theoretical intermezzo

It is of some theoretical interest to consider the quantile functions $B_\theta(p) = (1-\theta)p^2 + \theta(2p-p^2)$ for arbitrary $\theta \in [0,1]$. Here $B_n^{(1)}(p)$ corresponds to $B_\theta(p)$ if $\theta = \bar{u}$. An elementary analysis provides

$$G_\theta(u) = \begin{cases} (\theta - \sqrt{\theta^2 - (2\theta-1)u})/(2\theta-1) & \text{if } \theta \neq 1/2 \\ u & \text{if } \theta = 1/2 \end{cases}$$

with density

$$g_\theta(u) = \frac{1}{2\sqrt{\theta^2 + (1-2\theta)u}} \qquad (0 < u < 1)$$

(for $\theta = 0$ the Beta$(1, \frac{1}{2})$ distribution is obtained, for $\theta = 1$ it is Beta$(\frac{1}{2}, 1)$).

It is possible to extend this family $\{g_\theta \mid \theta \in [0,1]\}$ of densities by allowing negative values of θ, as well as values of $\theta > 1$. This extension, however, serves no practical purpose because we are interested in the testing of H_0: $g = 1$ and, hence, in the densities 'not too far from $g_{1/2}$'. If X_θ is a random variable with density function g_θ, then (for arbitrary $\theta \in \mathbb{R}$)

$$\begin{aligned} \mathbf{E}\, X_\theta &= \int_0^1 u g_\theta(u)\, \mathrm{d}u \\ &= \int_0^1 B_\theta(p)\, \mathrm{d}G_\theta(B_\theta(p)) \\ &= \int_0^1 B_\theta(p)\, \mathrm{d}p \\ &= \tfrac{1}{3}\theta + \tfrac{1}{3}. \end{aligned}$$

In a parametric approach to the testing of H_0: $g \equiv 1$ the attention might be concentrated on level-α tests which are 'optimal' if g belongs to some parametric family $\{g_\theta \mid \theta \in \Theta\}$. The densities just considered to test H_0: $g \equiv 1$ or, equivalently, H_0: $\theta = \frac{1}{2}$, constitute a possibility but not as convenient as the exponential densities

$$\tilde{g}_\theta \propto \mathrm{e}^{\theta u - \chi(u)},$$

because in the latter case a uniformly most powerful unbiased size-α test exists for H_0: $\theta = 0$ whereas in the earlier case only locally most powerful tests exist, and a less satisfactory criterion has to be chosen. However, the choice of this one-parameter exponential family is a convenient but subjective 'rationalization': if $t : [0, 1] \to \mathbb{R}$ is any increasing function of u, e.g. $t(u) = \log u$, then another exponential family $\{\tilde{\tilde{g}}_\theta | \theta \in \Theta\}$ is obtained by specifying $\log\left(\tilde{\tilde{g}}_\theta(u)\right) = c(\theta) + \theta t(u)$. The uniformly most powerful unbiased size-α test is then based on $n^{-1}\sum t(u_i)$.

The definition of g_θ was based on the idea that $b_\theta(p) = (1-\theta)2p + \theta(2-2p)$ is a linear function of p if $0 \le \theta \le 1$. It is of similar interest to study the assumption $g \in \{\bar{g}_\theta | \theta \in \Theta\}$ where $\bar{g}_\theta(u) = 2\theta u + (1-\theta)(2-2u)$ and, hence, $\bar{b}_\theta = g_\theta$. If one now wants to test H_0: $\theta = \frac{1}{2}$ versus A: $\theta \neq \frac{1}{2}$ then, again, $|\bar{u} - \frac{1}{2}|$ is obtained as the basis of a locally MP unbiased but not uniformly MP test. It is interesting, from a theoretical viewpoint, to study this testing problem in detail (there will exist a most stringent size-α test, this can be compared with the most stringent somewhere MP invariant size-α test, etcetera). This, however, serves almost no practical purpose. (*End of intermezzo.*)

In Section 4.2 the test statistics $T_n^{(m)}$ and $\tilde{T}_n^{(m)}$ were defined. For $m = 1$ we have

$$\begin{aligned}
t_n^{(1)} &= ||b_n^{(1)} - 1||_1 \\
&= \int_0^1 |2(1-\bar{u})p + \bar{u}(2-2p) - 1| \, \mathrm{d}p \\
&= |2\bar{u} - 1| \int_0^1 |1 - 2p| \, \mathrm{d}p \\
&= |\bar{u} - \tfrac{1}{2}|, \quad \text{and} \\
\tilde{t}_n^{(1)} &= \sup_p \left| B_n^{(1)}(p) - p \right| \\
&= \sup_p |2\bar{u} - 1| p(1-p) \\
&= \tfrac{1}{2}|\bar{u} - \tfrac{1}{2}|.
\end{aligned}$$

Conclusion

If one agrees upon $m = 1$, then there is consensus about using the deviation of $\bar{u}$ from $\frac{1}{2}$ as test statistic. The corresponding P-value is approximately given by

$$\mathrm{P}(\chi_1^2 \le 12(\bar{u} - \tfrac{1}{2})^2) = 2\Phi(-\sqrt{12n}|\bar{u} - \tfrac{1}{2}|).$$

This test corresponds to that of NEYMAN (1937) if a polynomial of degree 1 is used. The test rejects H_0 if the deviation of $\bar{u}$ from $\frac{1}{2}$ is sufficiently large. The test has no power if the true density g is such that $\int ug(u) \, \mathrm{d}u = \frac{1}{2}$. For all densities with this expectation different from $\frac{1}{2}$, the power of the level-α-test is governed by the asymptotic normality of $\bar{U}$ around this expectation.

The example revisited

See Section 4.1 and note that the density $f_{20}^{(1)}$ in Figure 4.2 (left) corresponds to

$g_n^{(1)}(u) = g_\theta(u)$ with $\theta = \bar{u} = .641$ (see the beginning of this section and note that $f_{20}^{(1)}(x)$ is equal to $g_{20}^{(1)}(\Psi(x))\psi(x) = \frac{1}{25} g_{20}^{(1)}(\frac{x}{25})$). The P-value

$$\mathrm{P}_1 = 2\Phi(-\sqrt{12n}|\bar{u} - \tfrac{1}{2}|) = .0290,$$

was already mentioned in Section 4.1.

4.4 The case $m = 2$

The exact equivalence with a Neyman smooth test vanishes if $m = 2$ because then

$$\begin{aligned} B_n^{(2)}(p) &= p^3 + \frac{3p(1-p)}{\binom{n}{2}} \sum_{i=1}^{n} (n - i + p(2i - n - 1)) u_{[i]} \\ &= p^3 + 3p(1-p)\bar{u} + 3p(1-p)(p - \tfrac{1}{2})g \\ &= p + 3p(1-p)\varepsilon + 3p(1-p)(p - \tfrac{1}{2})\delta \end{aligned}$$

where

$$\begin{aligned} \bar{u} &= \tfrac{1}{n} \sum_{i=1}^{n} u_i && (U\text{-statistic}) \\ &= \tfrac{1}{n} \sum_{i=1}^{n} u_{[i]} && (L\text{-statistic}) \\ &= \tfrac{1}{2} + \varepsilon, && \text{and} \end{aligned}$$

$$\begin{aligned} g &= \tfrac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |u_i - u_j| && (U\text{-statistic}) \\ &= \tfrac{2}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1) u_{[i]} && (L\text{-statistic}) \\ &= \tfrac{1}{3} + \delta \end{aligned}$$

are respectively the sample mean and Gini's mean difference. We introduced $\varepsilon = \bar{u} - \frac{1}{2}$ and $\delta = g - \frac{1}{3}$ because, under H_0, $\mathbf{E}\ \bar{U} = \frac{1}{2}$ and $\mathbf{E}\ G = \mathbf{E}\ |U_i - U_j| = \frac{1}{3}$. Many years ago there was a discussion between Fisher, Edgeworth and Gini about the appropriateness of measures of spread like the sample standard deviation s and Gini's mean difference g. It is interesting to note that in case $m = 2$ the statistics $\bar{u}$ and g (or ε and δ) appear if one expresses $b_n^{(2)}$ as a linear combination of the polynomials 1, p, and p^2.

Note: it would be of some theoretical interest to study the distributions on $[0, 1]$ with quantile functions $\{B_\theta | \theta \in \Theta\}$ where B_θ is obtained from the expression of $B_n^{(2)}$ by replacing (ε, δ) by $\theta \in (\theta_1, \theta_2)$. As a similar study for $m = 1$ suggests (see the theoretical intermezzo in Section 4.3) no results of practical interest are expected, and that is why this study is not performed here. To continue, we need the following result:

Theorem 4.1 *Under* H_0*:* $f = \psi$ *we have*

$$\mathcal{L}\, n^{1/2} \begin{pmatrix} \varepsilon \\ \delta \end{pmatrix} \longrightarrow \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2 & \rho \\ \rho & \tau^2 \end{bmatrix} \right),$$

with $\sigma^2 = \frac{1}{12}$, $\tau^2 = \frac{1}{45}$, *and* $\rho = 0$ *and exact equalities for* $\mathrm{Var}\ (\varepsilon) = \frac{1}{n}\sigma^2$ *and* $\mathrm{Cov}\ (\varepsilon, \delta) = 0$.

For the proof see Appendix A. It follows from this theorem that under H_0

$$\mathcal{L}\, n \left(12\varepsilon^2 + 45\delta^2\right) \to \chi^2_2 = \mathrm{Gamma}(1, \tfrac{1}{2}),$$

and that, hence, using any positive multiple of $12\varepsilon^2 + 45\delta^2$ as test statistic, the approximate P-value

$$\mathrm{P}_2^{(1)} = \mathrm{P}(\chi^2_2 \leq n(12\varepsilon^2 + 45\delta^2)) = \exp(-n(3\varepsilon^2 + 11.25\delta^2))$$

is obtained. Though it is convenient to use $12\varepsilon^2 + 45\delta^2$ as test statistic, we prefer to concentrate the attention on the test statistic $T_n^{(2)}$ with outcome

$$t_n^{(2)} = ||b_n^{(2)} - 1||_1 = 3 \int_0^1 \left| -3\delta p^2 + (3\delta - 2\varepsilon)p + (\varepsilon - \tfrac{1}{2}\delta) \right| \, \mathrm{d}p.$$

In practice, the computer will be used to obtain $t_n^{(2)}$. To increase the understanding, an approximation is obtained by using

$$t_n^{(2)} = \begin{cases} 3^{-1/2}|\delta| & \text{if } \varepsilon = 0 \\ \frac{3}{2}|\varepsilon| & \text{if } \delta = 0 \\ \frac{16|\varepsilon|}{9} = \frac{3|\varepsilon|}{2} + \frac{5|\delta|}{36} & \text{if } \delta = 2\varepsilon \\ \frac{8|\varepsilon|}{3\sqrt{3}} = \frac{3|\varepsilon|}{2} + \frac{(16\sqrt{3}-27)|\delta|}{12} & \text{if } \delta = \frac{2\varepsilon}{3} \end{cases}$$

to suggest the approximation

$$t_n^{(2)} \approx \sqrt{\tfrac{9}{4}\varepsilon^2 + \tfrac{1}{3}\delta^2}$$

which, actually, is an upper bound, see Figure 4.1. We conclude that in this statistic much more weight is assigned to ε than in the χ^2_2-statistic derived earlier because

$$\tfrac{9}{4}\varepsilon^2 + \tfrac{1}{3}\delta^2 = \tfrac{1}{3}(\tfrac{27}{4}\varepsilon^2 + \delta^2)$$

while

$$12\varepsilon^2 + 45\delta^2 = 45(\tfrac{12}{45}\varepsilon^2 + \delta^2).$$

and $\frac{27}{4} \gg \frac{12}{45}$.

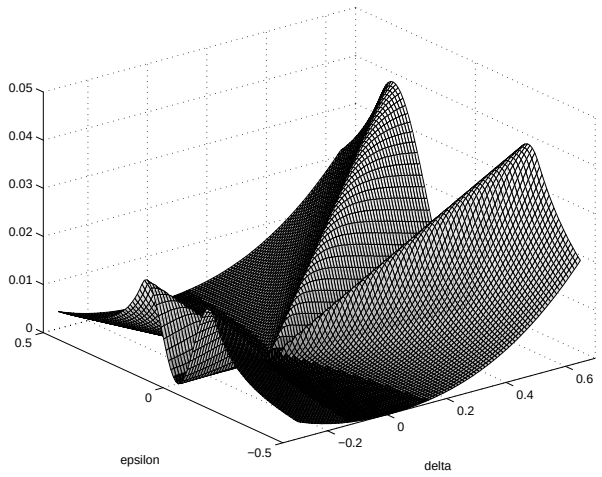


Figure 4.1: Difference $\sqrt{\frac{9}{4}\varepsilon^2 + \frac{1}{3}\delta^2} - t_n^{(2)}$*, for all theoretical possibilities of* (ε, δ)*.*

Note that the difference between its approximate value and $t_n^{(2)}$ is always nonnegative and that the approximation is very good if (ε, δ) is in the neighborhood of its expected value $(0, 0)$ under H_0.

The computation of the P-value $\mathrm{P}_0(T_n^{(2)} \geq t_n^{(2)})$ will be done on basis of an extensive numerical simulation as well as on the basis of the approximation where $\frac{9}{4}\varepsilon^2 + \frac{1}{3}\delta^2$ is used as test statistic and, for the sake of convenience, its distribution under H_0 is approximated by that of $c_n\chi_1^2$ with $c_n = (\frac{9}{4}\frac{1}{12} + \frac{1}{3}\frac{1}{45})n^{-1} = .195n^{-1}$ such that the (asymptotic) expectations coincide. The χ_1^2 distribution is used for reasons of simplicity and because $\frac{9}{4}\varepsilon^2$ is the dominant term. This provides the approximate P-value

$$\mathrm{P}_2^{(2)} = \mathrm{P}(\chi_1^2 \leq c_n^{-1}(\tfrac{9}{4}\varepsilon^2 + \tfrac{1}{3}\delta^2)) = 2\Phi\left(-\sqrt{c_n^{-1}(\tfrac{9}{4}\varepsilon^2 + \tfrac{1}{3}\delta^2)}\right).$$

The exact distribution of $T_n^{(2)}$ under H_0 is simulated as explained in Appendix B. Table B.1 provides critical values for $t_{n,\alpha}^{(2)}$. A numerical elaboration is presented in the following section.

The example, revisited
With respect to the example of Section 4.1 we have $\bar{u} = .64$, $g = .30$, $\varepsilon = .141$, $\delta = -.038$. The χ_2^2-test provides the approximate P-value

$$\mathrm{P}_2^{(1)} = .0231.$$

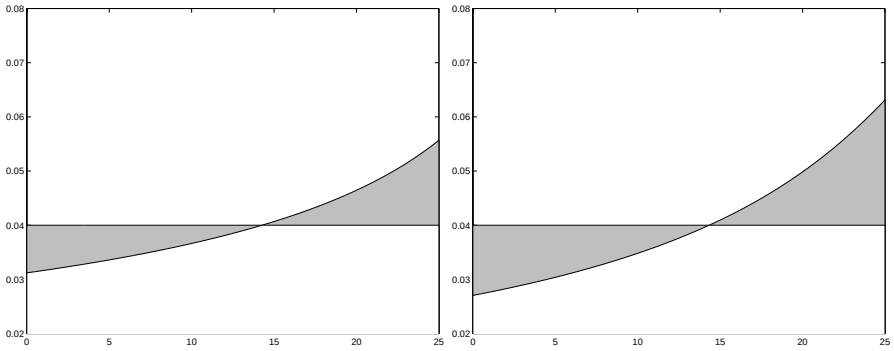


Figure 4.2: Density estimates (curves) of the data in the example with $m = 1$ (left) and $m = 2$ (right) and the uniform null-distribution. The test areas indicated are .141 ($m = 1$) and .212 ($m = 2$).

The test based on $\frac{9}{4}\varepsilon^2 + \frac{1}{3}\delta^2$ provides

$$\mathrm{P}_2^{(2)} = 2\Phi\left(-\sqrt{\tfrac{20}{.195}(\tfrac{9}{4}.141^2 + \tfrac{1}{3}.038^2)}\right) = .0313,$$

while the approximation $(\frac{9}{4}\varepsilon^2 + \frac{1}{3}\delta^2)^{1/2} = .212$ is very close to the shaded area $t_n^{(2)} = .212$ in Figure 4.2 (right). Note that $\varepsilon = .141$ is more than twice as large as its standard deviation $(\frac{1}{20} \times \frac{1}{12})^{1/2} = .065$ whereas $\delta = -.038$ (in absolute value) is comparable to its standard deviation $(\frac{1}{20} \times \frac{1}{45})^{1/2} = .033$.

Using Table B.1 in the appendix, it follows from $t_n^{(2)} = .212$ that the corresponding P-value $\mathrm{P}_2^{(3)}$ is slightly larger that .025. Note that $t_{20,.025}^{(2)} = .216$ while $t_{20,.05}^{(2)} = .190$; a linear interpolation provides

$$\mathrm{P}_2^{(3)} \approx .029.$$

4.5 The general case $m \geq 2$

In principle, the results of the previous two sections can be generalized to $m = 3, 4, \ldots$. In Section 4.4 exact representations, and approximations, were given in terms of $\varepsilon = \bar{u} - \frac{1}{2}$ and $\delta = g - \frac{1}{3}$, but simulations were needed to study the exact distribution of the test statistic $T_n^{(2)}$ under H_0. In the case $m = 3$ it is possible to define a special weighted sum h of $\sum i^2 u_{[i]}$, $\sum i u_{[i]}$, and $\sum u_{[i]}$, and deviation $\eta = h - \mathbf{E}\, h$, such that $\frac{1}{n}(\varepsilon, \delta, \eta)$ has a trivariate asymptotically normal distribution, probably without

correlations, with $b_n^{(3)}$ a function of $\varepsilon, \delta, \eta$. To be specific,

$$\begin{aligned} B_n^{(3)} &= p^4 + 2\binom{n}{3}^{-1} p(1-p) \times \Bigg[(5p^2 - 5p + 1) \sum_{i=0}^{n+1} i^2 u_{[i]} \\ &\quad + (-5(n+1)p^2 + (7n+1)p - 2n + 1) \sum_{i=0}^{n+1} i u_{[i]} \\ &\quad + ((n^2 + 2n + 2)p^2 - n(2n-1)p + n(n-1)) \sum_{i=0}^{n+1} u_{[i]} \Bigg]. \end{aligned}$$

Clearly, the results are too complicated to be useful. Anyway, the distribution of the test statistic has to be studied via simulation and we suggest to rely on the simulation results in Tables B.2 and B.3, regardless the complicatedness of $B_n^{(3)}$. Therefore, this lack of concreteness is a nuisance, but it does not affect the general line of thought.

Theoretical considerations
In Section 4.2 the symmetrized $B_n^{(m)}(p)$ (U-statistic) has been rewritten as an L-statistic. It was noted that the statistic $T_n^{(m)}$ we are interested in has as its outcome

$$t_n^{(m)} = ||f_n^{(m)} - \psi||_1 = ||b_n^{(m)} - 1||_1$$

This implies that stochastic theory about the distribution of $b_n^{(m)}(p) - 1$ is needed, especially under H_0: $b \equiv 1$. In principle, theory on U- and L-statistics is applicable, for p fixed. The results of Section 3.3 suggest that the distribution of $m^{-1/4} n^{1/2} T_n^{(m)}$ is 'fairly stable' under H_0. This can be used to predict results of the simulation experiments.

For $m = 1$ it is known from Section 4.3 that the distribution of $m^{-1/4} n^{1/2} T_n^{(1)} = n^{1/2} |\bar{u} - \frac{1}{2}|$ is, under H_0, approximately (asymptotically) like $\frac{1}{\sqrt{12}} \chi = |\mathcal{N}(0, \frac{1}{12})|$. For $m = 2$ it is known from Section 4.4 that, under H_0, the distribution of $m^{-1/4} n^{1/2} T_n^{(2)} \approx 2^{-1/4} n^{1/2} (\frac{9}{4}\varepsilon^2 + \frac{1}{3}\delta^2)^{1/2}$ is not much different from that of $\sqrt{T}$ where T has the negative-exponential distribution. These approximations provide confidence in the simulation results reported in Appendix B.

Numerical elaboration
Figure 4.3 shows for $m = 2$, 3, 4 that the relation between the average (logarithm of) $t_{n,\alpha}^{(m)}$ and the (logarithm of the) sample size is approximately linear, with slopes independent of the choice of m and α. Postulating a common slope β, this slope can be estimated, as well as the intercepts. After that, approximate P-values for can be derived using these values. However, since the intercepts depend on both m and α in a complicated way, Tables B.1, B.2, and B.3 cannot be dispensed with.

The choice of m
The tables provided in Appendix B (and the obvious result for $m = 1$) enable us to test H_0 on the basis of $T_n^{(m)}$ ($m = 1, 2, 3, 4$). In principle a choice of m has to be made

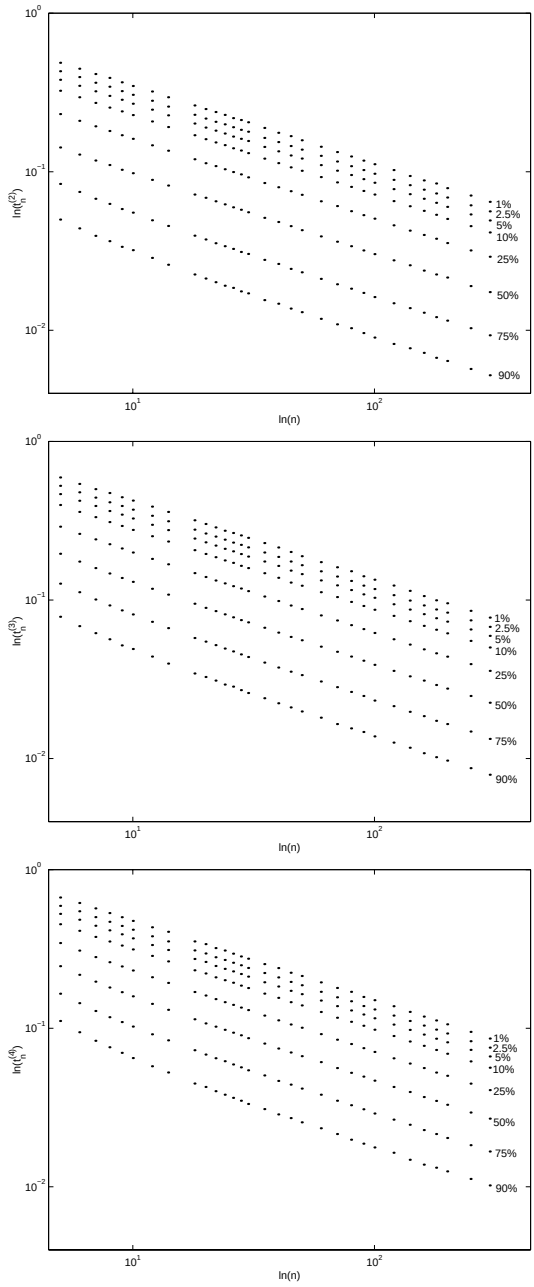


Figure 4.3: Display of simulation results, each dot represents an average $\log\left(t_{n,\alpha}^{(m)}\right)$ *(see Appendix B) for a specific choice of* $\log n$ *and* α*, with respectively* $m = 2$*, 3, 4.*

if one wants to comply with the Neyman-Pearson approach. At the end of Section 4.1 reference has been made to suggestions in the literature that (for similar problems) a small value of m is indicated. For the sample size $n = 20$ of our example, perhaps $m = 1$ or 2 is appropriate. For $n = 100$ the choice $m = 3$ or 4 may be attractive. We content ourselves with such suggestions because, in practice, the statistician will not worry too much about the effect of applying various values of m. In our opinion it is reasonable to discuss the deviations $\varepsilon = \bar{u} - \frac{1}{3}$ and $\delta = g - \frac{1}{3}$ as well as $t_n^{(1)}$, $t_n^{(2)}$, and, perhaps, if $m = 1$ and $m = 2$ do not lead to rejection of H_0, $t_n^{(3)}$ and $t_n^{(4)}$ as well. Further rationalization will not be effective.

4.6 Relation with Neyman's smooth tests

The quantile density estimate $b_n^{(m)}(p) = \frac{\partial}{\partial p} B_n^{(m)}(p)$ is a *positive* polynomial function on $[0, 1]$ of degree m. In fact, $b_n^{(m)}$ is a *convex* combination (with positive weights expressible as L-statistics) of the densities of the Beta$(i + 1, m + 1 - i)$ distributions $(i = 0, \ldots, m)$ which, of course, are polynomials of degree m. This representation is very fortunate, because it implies that the $b_n^{(m)}$ and, hence, the density estimates $g_n^{(m)}$ and $f_n^{(m)}$ are genuine probability densities.

Nevertheless, it may be appealing to the mathematician to discuss an alternative basis of $L_2[0, 1]$, e.g. that of orthogonal polynomials or of other functions (e.g. trigonometric polynomials). This is of particular interest if we are discussing the density $g = G'$ of $U_1 = \Psi(X_1)$ (with $G = F \circ \Psi^{-1} = B^{-1}$) and the way it 'should be estimated'.

The discussion about the choice of basis is included, partly because it is related to the specification of NEYMAN's smooth tests (1937). For the terminology and motivation behind this method, we refer to RAYNER AND BEST (1989). Let $\varphi_0 \equiv 1, \varphi_1, \varphi_2, \ldots$ be any system of linearly independent squared integrable functions on $[0, 1]$. The Gram-Schmidt orthogonalization process provides the orthonormal basis $\psi_0, \psi_1, \psi_2, \ldots$ of (a subspace of) $L_2[0, 1]$

$$\varphi_0 \equiv \psi_0 \equiv 1, \quad \varphi_1 = \frac{\psi_1 - (\psi_1, \psi_0)\psi_0}{||\psi_1 - (\psi_1, \psi_0)\psi_0||}, \quad \ldots$$

If a function $h \in L_2[0, 1]$ (a quantile density or a probability density or some other function) can be written as a linear combination of $\varphi_0, \ldots, \varphi_k$ then it can equally well be written as a linear combination of $\psi_0, \ldots, \psi_k$. Now a variety of related paradigms can be discussed.

Paradigm 1

Focussing on the quantile densities, and starting from the estimate $b_n^{(m)}$ with $m = k$, we can consider $\varphi_h(p) = p^h \quad (h = 0, \ldots, k)$ and determine the weights $w_{n,h} \quad (h = 0, \ldots, k)$ such that $b_n^{(k)}(p) = \sum_{h=0}^{k} w_{n,h} p^h$. Section 4.3 shows that, in case $k = 1$,

$$w_{n,0} = 2\bar{u}, \qquad w_{n,1} = 2(1 - 2\bar{u}).$$

Section 4.4 provides for $k = 2$ that

$$w_{n,0} = 1 + 3(\varepsilon - \tfrac{1}{2}\delta), \qquad w_{n,1} = -6\varepsilon + \tfrac{9}{4}\delta, \qquad w_{n,2} = -9\delta.$$

For higher m expressions can be obtained as well using the Gram-Schmidt process. The formulas are too complicated to be given. Under H_0: $b \equiv 1$, the ideal weight $w^*_{n,0} = 1$ and the other weights are 0. The deviations from the ideal weights are 2ε and -42ε in the case $k = 1$; $-3(\varepsilon - \frac{1}{2}\delta)$, $-6\varepsilon + \frac{9}{4}\delta$, and -9δ in the case $k = 2$, etcetera. The first deviation is a linear combination of the other ones. Hence it is reasonable to use $w_{n,1}, \ldots, w_{n,k}$ as the basis of a test. This can be a χ^2_k test, but there are arguments to pay more attention to $w_{n,1}$ than to $w_{n,2}$, etcetera. That is what has been done in Section 4.5.

Paradigm 2

Focussing on the probability densities, NEYMAN (1937) provided a unified approach by choosing a number k and corresponding basis functions $\varphi_0, \ldots, \varphi_k$, preferably the corresponding orthonormal ones $\psi_0, \ldots, \psi_k$. Important examples are that where $\varphi_h(y) = y^h \quad (h = 0, \ldots, k)$ and that where $\varphi_1, \ldots, \varphi_k$ are the indicator functions of the subintervals $(p_0, p_0 + p_1], (p_0 + p_1, p_0 + p_1 + p_2], \ldots, (1 - p_k, 1]$ of $[0, 1]$ to test H_0: $g \equiv 1$, the last approach providing Pearson's χ^2_k test, with approximate P-value

$$\mathrm{P}\left(\chi^2_k \geq \sum_{j=0}^{k} \frac{(n_j - np_j)^2}{np_j}\right)$$

where n_j is the number of observations in cell j. Regarding the choice of the number of components k in Neyman's test, RAYNER AND BEST (1989) states that '$k \leq 4$ will usually suffice'. See INGLOT ET AL. (1990, 1994), and KALLENBERG ET AL. (1985) for extensive analyses in this respect.

Given such system $\varphi_0, \ldots, \varphi_k$ and being interested in the testing of H_0: $g \equiv 1$, we can consider the random variables $(\varphi_0(U_i) \equiv 1), \varphi_1(U_i), \ldots, \varphi_k(U_i)$ and their expectations and (co)variances under H_0. We have

$$\begin{aligned} \mu_j &= \mathbf{E}_0 \varphi_j(U_i) = \int_0^1 \varphi_j(u)\,\mathrm{d}u \\ \sigma_{jh} &= \mathrm{Cov}_0(\varphi_j(U_i), \varphi_h(U_i)) = \int_0^1 (\varphi_j(u) - \mu_j)(\varphi_h(u) - \mu_h)\,\mathrm{d}u \end{aligned}$$

$(j, h = 1, \ldots, k)$. The random variables $U_1, \ldots, U_k$ actually observed have density g not necessarily equal to that of $\mathcal{U}(0, 1)$. The true expectations $\mathbf{E}\ \varphi_j(U_i)$ may deviate from their values μ_j under H_0. Unbiased estimates are provided by the sample means $T_j = n^{-1} \sum_{i=1}^n \varphi_j(U_i)$. The Multivariate Central Limit Theorem implies that $(T_1, \ldots, T_k)'$ is asymptotically normally distributed, under H_0, with expectation $\mu = (\mu_1, \ldots, \mu_k)$ and covariance matrix $\frac{1}{n}\Sigma = (\frac{1}{n}\sigma_{jh})$. (Note that this covariance matrix is equal to the inverse of Fisher's information matrix.) Hence

$$n(T - \mu)'\Sigma^{-1}(T - \mu) \sim \chi^2_k$$

and an approximate P-value is given by $\mathrm{P}\left(\chi^2_k \geq n(t-\mu)'\Sigma^{-1}(t-\mu)\right)$.

In the situation of classified data considered by him, Karl Pearson regarded this P-value as a 'fairly reasonable criterion of the probability that the deviations $t_j - \mu_j$ can be supposed to have arisen from random sampling', i.e. of the probability than H_0 is true. This formulation allows some modification of P. In fact, it suggests that a recalibration might be considered. In this respect SALOMÉ ET AL. (1999) is of interest: if P is very small, say 10^{-4}, then it would not be wise to bet on the basis of the odds ratio $(1-p)/p$ that H_0 is false, there is a tendency to display 'overconfidence'. If one uses the orthonormal system $\psi_0, \ldots, \psi_k$ instead of $\varphi_0, \ldots, \varphi_k$ then one has

$$\begin{aligned} \mu_j &= \mathrm{E}_0\ \psi_j(U_i) = 0 \\ \sigma_{jh} &= \mathrm{Cov}_0\ (\psi_j(U_i), \psi_h(U_i)) = (\psi_j, \psi_h) = \mathbf{1}_{j=h} \end{aligned}$$

One will then obtain the P-value

$$\mathrm{P}\left(\chi^2_k \geq \sum_{j=1}^{k} \frac{1}{n}\left(\sum_{i=1}^{n} \psi_j(U_i)\right)^2\right)$$

which, of course, is in *exact* agreement with the expression derived before.

As it is natural to expect that the 'deviations' $\frac{1}{n}\sum_{i=1}^{n} \phi_j(U_i)$ are of decreasing importance if j increases, it is also natural to replace the test statistic used by some weighted sum

$$\sum_{j=1}^{k} \frac{w_{n,j}}{n}\left(\sum_{i=1}^{n} \psi_j(U_i)\right)^2$$

the weights $w_{n,1} > w_{n,2} > \ldots$ being decreasing. It has been established in SCHAAFSMA AND STEERNEMAN (1981) that this may lead to a qualitative improvement. The choice of weights, however, is a delicate issue. That is why we keep contenting ourselves with the discussion presented at the end of Section 4.5.

Paradigm 3

Returning to the problem of estimating the probability density function, Neyman played with the idea that any density g on $[0,1]$ can be approximated by its projection on the $k+1$ dimensional subspace spanned by $\varphi_0, \ldots, \varphi_k$ or, equivalently, by $\psi_0, \ldots, \psi_k$. His line of thought was, more or less, that

$$\hat{g} = 1 + \sum_{j=1}^{k}\left(n^{-1}\sum_{i=1}^{n} \psi_j(u_i)\right)\psi_j$$

is a nice estimate of g, especially because it is an unbiased estimate of g if g itself corresponds to its projection $1 + \sum_{j=1}^{k}(g, \psi_j)\psi_j$ on the $k+1$ dimensional subspace indicated. A serious drawback of these estimates is that negative values may arise.

This difficulty does not appear if we use the estimates of the form $b_n^{(m)}$, neither will it appear if we follow, e.g., the *maximum entropy* approach of JAYNES (1996) which, in the present context, is as follows (for a short discussion on the method of maximum entropy, see Section 3.8).

Suppose that we have estimates

$$\hat{\mu}_j = n^{-1} \sum_{i=1}^{n} \psi_j(u_i)$$

of the expectations $\mu_j = \mathbf{E}\ \psi_j(U_i)$ and that we are interested in the true density g of U_i $(i = 1, \ldots, n)$. Our estimate $\hat{g}$ of g 'should' satisfy the restrictions $\int_0^1 \psi_j(u) g(u)\, \mathrm{d}u = \hat{\mu}_j,\ (j = 1, \ldots, k)$ and be such that the Shannon entropy

$$-\int_0^1 g(u) \log(g(u)) \mathrm{d}u$$

is maximum. The solution to this optimalization problem is, somewhat surprisingly, that $\hat{g} = g_{\hat{\theta}}$ where

$$g_\theta(u) = \exp\left(\theta_1 \psi_1(u) + \ldots + \theta_k \psi_k(u) - c(\theta)\right)$$

defines an exponential family and $\hat{\theta}$ is the maximum likelihood estimate of θ. The P-value thus obtained for testing H_0: $g \equiv 1$, or 'equivalently' for testing H_0: $\theta = 0_k$, is given by

$$\mathrm{P}\left(\chi_k^2 \geq \tfrac{1}{n} \hat{\theta}' \hat{\theta}\right).$$

Conclusion

Though the arguments are varying, most methods discussed have in common that H_0: $g \equiv 1$ is tested by choosing k linearly independent functions $\varphi_1, \ldots, \varphi_k : [0, 1] \to \mathbb{R}$ in addition to $\varphi_0 \equiv 1$, considering the corresponding orthonormal basis $\psi_0 \equiv 1, \psi_1, \ldots, \psi_k$, and by applying the χ_k^2 test indicated. Many authors have discussed the choice of the number k. KARL PEARSON (1900) himself stated

> 'Thus, if we take a very great number of groups our test becomes illusory. We must confine our attention in calculating P to a finite number of groups, and this is undoubtedly what happens in actual statistics. The number k of degrees of freedom will rarely exceed 30, often not greater than 12.'

Later generations of statisticians dealing with Neyman's smooth tests have made recommendations that k should not be larger than 2 or 3 (see the citation by Kallenberg in Section 4.1 and the remark and references earlier in this section).

The discussion here is about whether or not some additional basis vectors, say $e_{k+1}, \ldots, e_{k+l}$ should be involved in the analysis. If they are involved then they get equal rights and the χ_k^2 test is replaced by a χ_{k+l}^2 test. Authors like SCHAAFSMA AND STEERNEMAN (1981) indicated that a weighted combination, with more weight on ψ_i than on ψ_{i+1} is appropriate if considerable deviations ψ_j are less likely if j is larger.

4.7 Relations with other goodness of fit tests

We are fascinated by the total-variation (or L_1) distance $||f-\psi||_1$ and the Kolmogorov distance $||F-\Psi||_\infty$ because the total-variation distance is invariant under bijective mappings while the Kolmogorov distance is invariant under monotonous transformations. Under certain monotonicity assumptions we have that $||f-\psi||_1 = 2||F-\Psi||_\infty$. We always have $||f-\psi||_1 \leq 2||F-\Psi||_\infty$ (see, e.g., LOÈVE, 1955). Both distances are such that they do not change if the distribution functions $G = F \circ \Psi^{-1}$ are replaced by the corresponding quantile functions. Our test statistic $\tilde{T}_n^{(m)} = ||F_n^{(m)} - \Psi||_\infty$ has been obtained by replacing the unknown true quantile function B in $||F-\Psi||_\infty = ||B-1||_\infty$ by the corresponding estimate $B_n^{(m)}$ which is a continuous and increasing analogue of the empirical quantile function. KOLMOGOROV's test (1933) is based on $||\hat{B}-1||_\infty$ where $\hat{B}$ is the empirical quantile function. As the true quantile function is smooth, the estimates $B_n^{(m)}$ will be closer to the truth, on the average, than the discontinuous functions $\hat{B}$ on which they are based. It is also reasonable to expect that the power properties of the tests based on $||f_n^{(m)} - \psi||_1$ and $||F_n^{(m)} - \Psi||_\infty$ are better than those based on Kolmogorov's test. Much will depend, however, on the alternative hypotheses to be considered and on the choice of m to be made. A possibility for further research is to perform, by using extensive simulation studies, a concrete power comparison between the $||f_n^{(m)} - \psi||_1$-statistics ($m = 1,\ 2,\ 3$) in the line of the power comparison in INGLOT ET AL. (1994).

A more delicate issue is as follows. If one accepts that the context asks for a test statistic of the form $||\hat{f} - \psi||_1$ then the question arises which nonparametric density estimate $\hat{f}$ one should use. In DE BRUIN ET AL. (1999) it was made very clear that the estimator f_n studied there is 'not unreasonable though some further improvement is possible'. Such improvement can be achieved

1. By using $f_n^{(m)}$ instead of $f_n = f_n^{(n)}$; or
2. By using a kernel estimator k_n, preferably with the bandwidth determined such that the method is optimal for estimating ψ itself. (Note that ψ is given.)

The comparison between the tests based on the statistics $||f_n^{(m)} - \psi||_1$ and $||k_n - \psi||_1$ will depend on a large number of arguments, e.g. the choice of m and of the basic kernel, the alternative hypotheses for which power comparisons are made, etcetera. *An argument in favor of* $||f_n^{(m)} - \psi||_1 = ||h_n^{(m)} - 1||_1$ *is that the distribution of the test statistic under* H_0 *does not depend on* ψ (critical values can be found in Appendix B). For the test statistics $||k_n - \psi||_1$ additional simulation studies would be needed.

Part III

Applications

Chapter 5

Analyzing water-quality data

> *'To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.'*
>
> Sir R.A. Fisher[1]

The Advancement of Science is largely a matter of combining rational argument and empirical evidence. In areas like philosophy and pure mathematics rationalization is in the forefront whereas the empirical sciences deal with the design of experiments, collection of empirical evidence, etcetera. In mathematical statistics we try to combine these arguments. In principle the development of theory should follow the needs from practice while experiments should follow the needs of theory. Fisher's statement is therefore also justified in the reversed case where *practice is called in after the theory has been developed*. Interaction between theory and practice is needed. That is why the author of this thesis spent two months at the Centre for Quantitative Methods in Eindhoven (The Netherlands) adapting the (original) theory of Chapter 3 such that it can be applied to water quality data available at CQM.

5.1 Description of the data

The Institute for Inland Water Management and Waste Water Treatment , RIZA[2], is the research and advisory body for inland water for the Dutch Directorate-General for Public Works and Water Management, *Rijkswaterstaat*. The Institute for Coast and Sea, RIKZ[3], is its research and advisory body on the subject of sea and coastal water. These institutes collect data on and conduct research into quality and quantity of water. On the basis of the data, RIZA and RIKZ make recommendations concerning the management of water in The Netherlands and abroad. Both institutes participate in Rijkswaterstaat's MWTL-project[4], a large-scale long-running project concerning the topics mentioned.

RIKZ and RIZA are interested in concentrations of physical, chemical and biological substances in inland and sea water. For that purpose measurements of concentrations

[1] *Indian Statistical Congress*, Sankhyā, ca. 1938
[2] RIZA: Rijksinstituut voor Integraal Zoetwaterbeheer en Afvalwaterbehandeling
[3] RIKZ: Rijksinstituut voor Kust en Zee
[4] MWTL: Monitoring Waterstaatkundige Toestand des Lands

in surface water are made. Those concentrations include chemicals, metals, radioactivity levels, and biological entities. A considerable number of observations falls below a given detection-, determination- or report-threshold Δ.

The level of the threshold depends on the measured substance. This censoring to the left generates difficulties, at least when concentrations below Δ were encountered in the sample. If only very few values are below Δ, then different methods will not lead to too different results. For example, any 'reasonable' method to generate a median estimate will not be effected if (much) less than 50% of the data are censored. However, in the data collected by RIZA and RIKZ, these fractions are often high and a reliable method for dealing with such observations is needed.

The MWTL data set used in the analysis is of size $N = 47850$. Measurements are performed for 35 different substances, and the data spread a range of approximately 10 years. All kinds of water-systems (sea water, small and large rivers, lakes, etcetera) in The Netherlands are covered; measurements were taken at 64 locations (at most locations only a few of the 35 substances were measured). For the different substances, the locations and time period of measurement vary. Each entry in the data set consists of the following parameters: the name of the substance, chemical or quantity measured, the location where this is measured, the date of measurement (only for RIKZ-measurements), and, of course the measured concentration.

On average, 1350 measurements have been performed per substance. The MWTL data set is a sample in space (64 locations) and time (a span of approximately ten years). It is a questionable approach to regard the 1350 measurement sessions as a representative sample of the distribution of concentrations of the specified substance. Nevertheless, we will start out by using this assumption in our analysis. With a more sophisticated time-series analysis, the time factor might be taken into account, see Section 5.5 for more considerations.

As stated above, concentrations of 35 different 'substances' are measured. In 34 cases real concentrations were involved. One case deals with the presence of bacteriological entities, and the measured values are frequencies. This case will not be analyzed. The 34 cases can be classified according to whether

1. All measurements are above Δ (Section 5.2);
2. Some measurements are below Δ (Section 5.3);
3. Two or more thresholds are involved (Section 5.4).

The third situation occurs for instance when, during the measurement period, the measuring equipment is improved, thus lowering the threshold. Another possibility is that different equipment is used at different locations for measuring the same substance, not all equipments being equally sensitive.

5.2 All measurements are above the threshold

SWAVING AND DE VRIES (2000) investigated how these data should be dealt with

such that bias and uncertainty in estimates for the first four population moments are smallest. In this section we consider the cases where censoring is absent.

We want to use the estimates $f_n^{(m)}$ introduced in Chapter 3 to make inferences concerning all substances individually. The question now is how to choose the initial guess ψ, and the confidence ν in it. Given these ψ and ν, the smoothing factor m follows from the rule of thumb $m = 5.2\sqrt{n}\, v$ (see Section 3.7), where v is the L_1-distance $||\psi - f||_1$ where optimality is required (and the choice $w = 2$ has been incorporated). Simulations suggested that the method performs well if a choice of $v < \frac{1}{2}$ is justified. Using the approximate standard error

$$n^{-1/4} m^{1/8} \sigma_n^{(m)}(x) \approx \frac{m^{1/8} f_n^{(m)}(x)}{\sqrt[4]{4\pi n F_n^{(m)}(x)(1 - F_n^{(m)}(x))}}$$

confidence bands around $f_n^{(m)}(x)$ will be constructed via

$$f_n^{(m)}(x) \ \pm \ n^{-1/4} m^{1/8} \Phi^{-1}\left(\tfrac{\alpha}{2}\right) \sigma_n^{(m)}(x).$$

SWAVING AND DE VRIES (2000) state that the 34 substances with continuous measurement scale can be described reasonably well with suitable lognormal densities. That is why we will use (log)normal densities as initial guesses and transform the data via $t = \log x$. The parameters μ en σ^2 will be estimated from the data, so $\psi \sim \mathcal{N}(\bar{t}, s_t^2)$. The theory prescribes how to specify ψ a priori. Some data peeping is necessary and allowed if this is taken into account in specifying m (see Section 3.8 for a theoretical motivation). That the (transformed) data follow a normal distribution might be questionable. In many cases, a superposition of (normal) distributions should be expected (e.g. when measurements are performed at different locations, each location having its own distribution). However, as initial guess, a normal distribution is expected to suffice.

Because theoretical arguments are lacking, the specification of the confidence in the initial guesses will be based on the data. In this section, the quantification of v, is based on the following inputs:

1. the sample size n. Section 3.8 provides a contribution $v_* = v_*(n)$;
2. how well the data fits a normal density. The squared correlation coefficient r^2 in a normal probability plot regression (explained in the next section) provides information about the confidence in the initial guess. The higher the correlation, the higher the confidence in the assumption of (log)normality. This is formalized by the contribution $v_{**}(r) = 2(1 - r^2)$. Alternative contributions, mainly decreasing functions of r or another measure of correlation (e.g. $2\sqrt{1 - r^2}$) are also possible, but we consider the specification given as good enough.

In this section the data-dependent choice of v is defined through[5]

$$v = v(n, r) = v_*(n) + v_{**}(r).$$

[5]This is rounded off to the nearest multiple of .05 because of numerical arguments.

Swaving and De Vries were computing estimates of some characteristics, most importantly population moments. Using the estimated densities, alternative estimates for the first four (central) moments as characteristics of the estimated density $f_n^{(m)}$ can be constructed. Calculations for average, variance, skewness, and kurtosis, respectively, were done as follows:

$$\begin{aligned}
\bar{t} &= \int t f_n^{(m)}(t)\,\mathrm{d}t \\
s^2 &= \int (t-\bar{t})^2 f_n^{(m)}(t)\,\mathrm{d}t \\
g_1 &= \int \frac{(t-\bar{t})^3}{(s^2)^{3/2}} f_n^{(m)}(t)\,\mathrm{d}t, \\
g_2 &= \int \frac{(t-\bar{t})^4}{(s^2)^{4/2}} f_n^{(m)}(t)\,\mathrm{d}t.
\end{aligned}$$

Results for the substances with all measurements uncensored, were given in ALBERS (2001). In Figure 5.1 some results are visualized. The figures display the estimated concentration density and the 95%-confidence bands of

`pb50`, the concentration of led in particle dust suspended in water (measured in mg per kg 'dry-weight' particle dust;

`pb60`, the concentration of led in the meat of mussels (measured in mg per kg 'dry-weight' meat);

`Cl`, the concentration of chloride (measured in mg per liter);

`SiO2`, the concentration of silicate (measured in mg per liter).

The two graphs for led concentrations show similarities with normal densities, but also clear dissimilarities. This can also be seen from the estimated population characteristics (e.g. the estimated skewness for `pb60` is $-.34$, differing significantly from zero). The extreme skewness in the graph for silicate could indicate either the existence of a physical maximum concentration of `SiO2` in water or a superposition of densities. The latter explanation is the case, since silicate a substance that is needed for growth by a certain kind of fytoplankton, and these micro-algae appear only in spring and summer, hence in winter there is no 'thread' to silicate. The typical, almost bimodal character for chloride is explained by looking at locations separately: the concentration densities at fixed locations are approximately normal whereas their mixture is not (this is visualized in ALBERS, 2001).

5.3 Some measurements below threshold

For the substances with one fixed detection threshold Δ (and some observations below this Δ) the procedure used is as follows. The number of censored observations

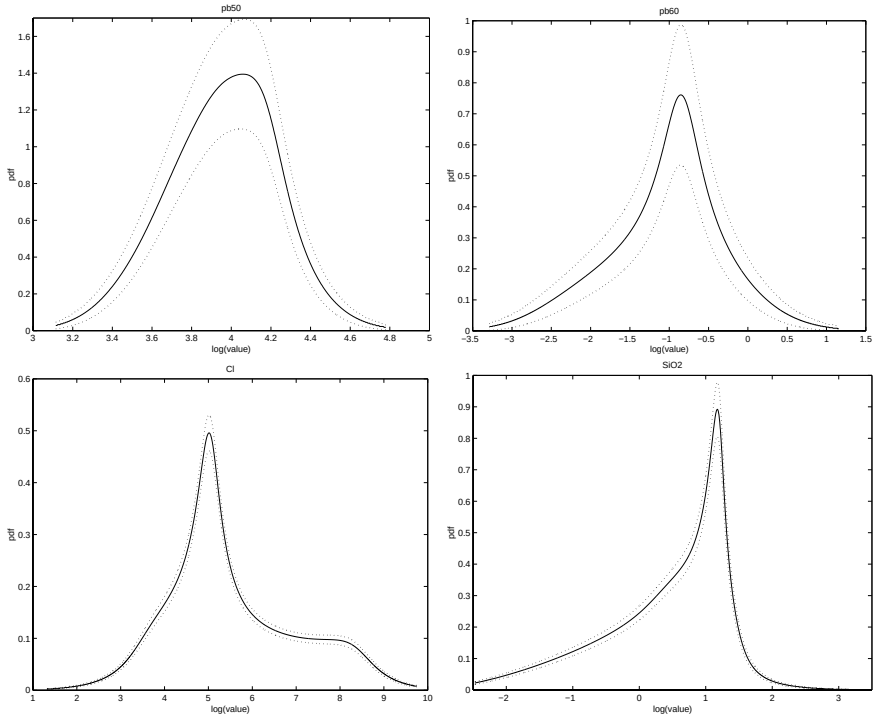


Figure 5.1: Above-left: `pb50` *($n = 213$, $m = 19$), above-right:* `pb60` *($n = 99$, $m = 16$), below-left:* `Cl` *($n = 2827$,$m = 41$), below-right:* `SiO2` *($n = 2206$, $m = 61$) and their 95% confidence bounds.*

k, the height of the threshold Δ, and, of course, the values of the $(n-k)$ uncensored measurements are known. Different methods exist for treating censored values. When the fraction of censored values is low, say $k \leq 2$, using an appropriate method is of course less important than when the fraction is considerable. The fraction of observations below the threshold varies between the substances from approximately 0.1% to 67%, and some sophistication will be necessary.

The method we will use, based on normal probability plot regression, is in line with SWAVING AND DE VRIES (2000). The censored measurements will be replaced by estimated values. Using ordinary least squares, the ordered observations $t_{[k+1]}, t_{[k+2]}, \ldots, t_{[n]}$ are fitted to quantiles $\Phi^{-1}(\frac{k+1}{n+1}), \Phi^{-1}(\frac{k+2}{n+1}), \ldots, \Phi^{-1}(\frac{n}{n+1})$ from a normal distribution. On basis of the resulting regression line, the estimator for $t_{[i]}$, $(i = 1, \ldots, k)$ is constructed by using the value corresponding to $\Phi^{-1}(\frac{i}{n+1})$, $(i = 1, \ldots, k)$. A disadvantage of this method is that estimated values are sometimes larger than Δ. When this occurs, the estimate is replaced by Δ.

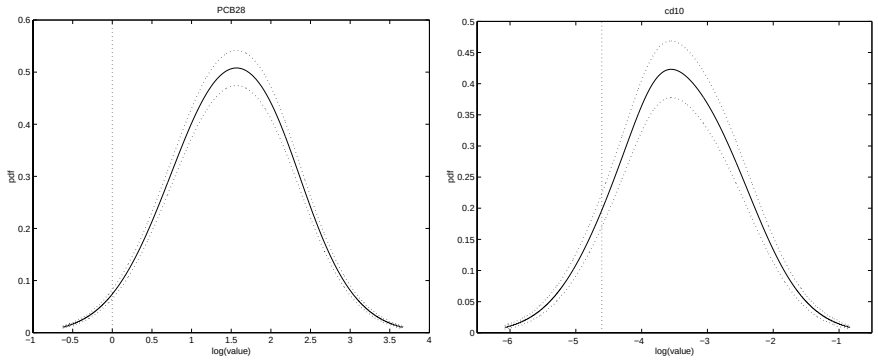


Figure 5.2: `PCB28` *($n = 1558$, $k = 34$, $\Delta = \log(1)$, $m = 10$) (left) and* `cd10` *($n = 706$, $k = 61$, $\Delta = \log(0.01)$, $m = 14$) (right).*

In SWAVING AND DE VRIES (2000) it was assumed that Δ was at a fixed point, and measurement errors were all zero. We continue these assumptions, in Section 5.5 we will reconsider this.

Using this adapted data set as 'the' data set, we continue in the same way as in Section 5.2. Of course, the consistency of the estimator for values below Δ has disappeared (unless $\Psi = F$), and the confidence bands are only exactly of level $(1 - \alpha)$ when $\psi = f$.

More precisely: for small values of concentration t ($t \ll \Delta$), we have replaced the latent observations by 'observations' from $\psi(t)$, and therefore $\mathbf{E}\ f_n^{(m)}(t) \approx \psi(t)$. For values larger than the threshold, the observations constitute a sample from the (unknown) true density f, providing $\mathbf{E}\ f_n^{(m)}(t) \approx f(t)$. Since our estimator is always a continuous estimator, the expected value of $f_n^{(m)}(t)$ with $t \approx \Delta$ will be 'somewhere between $\psi(t)$ and $f(t)$'. 'Where exactly' cannot easily be defined. Of course in the case that $\Psi = F$, practically impossible, consistency still holds. When f is truly distributed normally, ψ is a consistent estimator of f and, eventually, our estimator will be consistent as well. When f is not normal, a systematic error is made. In general, every method will generate systematic errors when the tail assumptions are incorrect. Making tail assumptions is a necessity, since no information about the tail behaviour is present.

Results are given in ALBERS (2001). Some examples are in Figure 5.2. The substances displayed are `PCB28`, a component of poly-chlorinated biphenyls; a type of organic micro-pollution (measured in μg per kg 'dry-weight' particles), and `cd10`, the concentration of cadmium (measured in μg per liter). Again, our method of fine-tuning a normal initial guess seems quite appropriate.

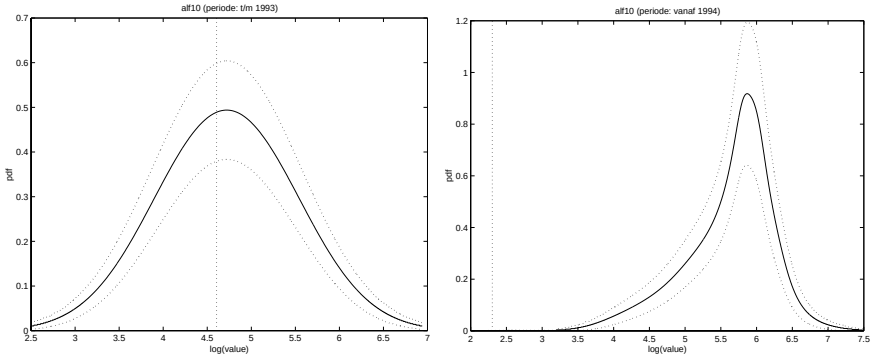


Figure 5.3: `alf10` *(left: until 1993, right: from 1994). A significant time effect is clearly visible.*

5.4 Substances involving multiple thresholds

Multiple thresholds may occur because of technological improvements, measuring with different types of equipment at different locations, etcetera. Two cases can be distinguished: those with and those without recording of the measurement-dates[6]. The case without date recording is not of interest to us.

After analyzing the data it is possible to classify each observation in a group with one Δ. Classification can be performed according to time, location, etcetera. When there are d thresholds, the classification will provide d groups, each with one threshold that matters, and the previous theory can be applied. Finally, with their sample sizes as weights, the separate estimates can be joined to construct one density estimate. This is, of course, only relevant if there are no significant differences between the d groups. Figure 5.3, for example, displays density estimates corresponding to the substance `alf10`, the amount of radiation in the water (measured in mili-Becquerel per liter). Until 1993, measurements were performed with equipment with a threshold of $\Delta_{d=1} = \log(100 \text{ mBq}/l)$, after that measurements were performed with a corresponding threshold of $\Delta_{d=2} = \log(10 \text{ mBq}/l)$. Kolmogorov's goodness of fit test (see Section 4.1 and KOLMOGOROV, 1933) clearly showed that the two (estimated) densities were significantly different. Joining them is unreasonable: it is better to infer about the two sub-populations separately.

5.5 Complications

Discrete measuring
Our density estimation method is used to estimate the distribution of the concen-

[6]It is unclear why the dates for RIKZ-measurements are missing, since they actually were recorded.

trations as measured. It is assumed that this is a continuous distribution, though measurements cannot be performed with infinite precision. When the precision is good enough, as is the case for most substances, the violation of the continuity assumption has almost no effect. However, for a few substances the precision is so low, that our method does not generate usable estimates, and a histogram seems the best possible thing.

Measurement errors
Until now, emphasis was put on descriptive statistics: the data were used to describe the population's behaviour. Focus was on the distributions of the measurements, not on the 'true' underlying concentration values (though even their definition is not more than operational).

The measurements are compositions of the true values, and systematic and random disturbances. To make inferences about the distribution of the true values y instead of the observed t, information has to be gathered about the measurement error $e = t - y$. The distribution $\mathcal{L}\ (e|Y = y)$ has to be known in order to transform $\mathcal{L}\ t$ into $\mathcal{L}\ Y$, which is the distribution of true interest.

With repeated measurements, information can be gathered about the measurement errors. Sometimes this error (or its variance) is neglectable, and the previously presented theory is applicable. This is not generally the case, especially for low concentrations the measurement error can be (relatively) high.

Point mass at zero
As mentioned earlier, we assumed continuity. It is not impossible that a point mass occurs for concentrations of zero. This would be the case for substances that only occur in water when the water is polluted, but not in general. Then, the concentration is zero with unknown probability λ, and larger than zero with probability $1-\lambda$. Before using the density estimation theory, an estimate $\hat{\lambda}$ has to be constructed.

Time effects
One of the major aspect of the data is unaccounted for in the analysis up till now: the measurements are performed in a time-frame. Reasons behind not discussing this earlier are that in our data set there was no record of date of measurement for RIZA-measurements while the density estimation method of Chapter 3 is not suitable, in its current form, for time series analysis.

In the analyses of this chapter the data were interpreted as being measured at one fixed time-point. This wrong assumption mainly influences the trustworthiness of the inferences for substances with a structural trend in its concentration, because the true density involved is constantly shifting and changing. Effects of the interpretation for substances that are assumed to behave fairly stable in time are (much) smaller. Previous experience (SWAVING AND VRIES, 2000) shows that most substances could be described quite well by lognormal distributions, this gives us confidence in the assumption that the time-effect is usually small compared to other effects.

Chapter 6

A problem from the interface between time series and multivariate analysis

> *'Whence, finally, this one thing seems to follow: that if observations of all events were to be continued throughout all eternity, (and hence the ultimate probability would tend toward perfect certainty), everything in the world would be perceived to happen in fixed ratios and according to a constant law of alternation, so that even in the most accidental and fortuitous occurrences we would be bound to recognize, as it were, a certain necessity and, so to speak, a certain fate.*
> *For all I know that is what Plato had in mind when, in the doctrine of the universal circle, he maintained that after the passage of countless centuries everything would return to its original state.'*
>
> JACOB BERNOULLI[1]

The density estimation method of Chapter 3, applied in Chapter 5, refers to situations where an *independent* random sample is considered. In practice, dependencies are often present, because time (or space) is involved, see also the remark in Section 5.5. Almost all theoretical approaches in time series analysis start from stationarity assumptions. If one is studying only *one* time series then it's difficult to think of anything else. In the study of growth curves the statistical perspective of a 'sample' from a 'population' of growth curves is relevant and the stationarity assumption can be dispensed with. In this chapter an issue from the analysis of stationary time series data will be treated by interpreting it as a problem from multivariate statistical analysis. The concept of strong similarity from the theory of distributional inference will be applied. The exact but suboptimal results obtained are contrasted with the asymptotic results some of which obtained earlier by DAVIS ET AL.(1995, 1996, 1998).

6.1 Introduction to the unit root problem

Modern time series analysts tend to remove long-term effects by differencing their time series. At a colloquium in Groningen (April 22th 1999) Richard Davis presented

[1] Jacob Bernoulli, **Ars Conjectandi**, Basileae, 1713; taken from SCHAAFSMA, Laudatio Sir David Cox, *Nieuw Archief voor Wiskunde*, **15**:3, 1997.

a lecture with the title *Recent Developments in the Unit Root Problem for Moving Averages*[2]. Largely summarizing DAVIS ET AL. (1988, 1995, 1996), he studied the following problem. The outcomes $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ are available and have to be used to test the hypothesis that the time series $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$ satisfies $X_i = E_i - E_{i-1}$, where $\ldots, E_{-1}, E_0, E_1, \ldots$ are independently $\mathcal{N}(0, \sigma^2)$ distributed. As a context, Davis suggested the moving-average model $X_i = E_i + \psi E_{i-1}$, with the assumption $-1 \leq \psi < 1$. This assumption is natural because when $|\psi| < 1$ we have stationarity, a desirable property of a moving average process; for $\psi = -1$ and $\psi = +1$ the distributions of $(X_1, X_2, \ldots, X_n)$ generated 'coincide' (time series analysts call such a MA(1) process with $|\psi| < 1$ an invertible process, and with $|\psi| = 1$ a noninvertible process). Another motivation is from DAVIS AND DUNSMUIR (1996)[3]: 'the maximum likelihood estimator in the $|\theta_0| < 1$ case is asymptotically normal with mean θ_0 and variance $(1 - \theta_0^2)/T$. However, the normal limit provides a particularly inaccurate approximation for values of θ_0 close to ± 1'. More arguments in favour of studying this problem are provided in DAVIS AND DUNSMUIR (1996, p. 3) and DAVIS AND MIKOSCH (1998, p. 100). The last authors claim that the problem is of some interest, because 'for example, a test that a time series has been over-differenced to achieve stationarity is equivalent to testing for the presence of a unit root in the moving average polynomial'. Davis has derived an intriguing asymptotic level-α test, with $\alpha = .045$, by studying the likelihood function $l_x(\theta)$ with respect to the parameter θ which has (ψ, σ^2) as its true value. He established that H_0: $\psi = -1$ should be maintained if and only if $\hat{\psi} = -1$, where $\hat{\psi}$ is the maximum likelihood estimator of ψ (under the restriction $-1 \leq \psi < 1$).

We shall return to Davis' asymptotic approach and other existing methods from literature in Section 6.5. Our interest in the problem has its origin in the relationship with multivariate analysis, especially with that part of multivariate analysis where the attention is concentrated on structured covariance matrices (intra-class correlation, random-effects models, Rao's growth curve models, factor analysis, etcetera).

The problem
Given is the outcome x of $X \sim \mathcal{N}_n(0_n, \Sigma)$ with $\Sigma = (\sigma_{i,j})$ such that

$$\sigma_{i,j} = \begin{cases} (1 + \psi^2)\sigma^2 & \text{if } i = j \\ \psi\sigma^2 & \text{if } |i - j| = 1 \\ 0 & \text{if } |i - j| \geq 2 \end{cases}$$

where (ψ, σ^2) is restricted to $[-1, 1) \times (0, \infty)$. Required are

1. a Neyman-Pearson level-α test for $\mathrm{H}_0 : \psi = -1$ against $\mathrm{A} : \psi > -1$;
2. estimates of ψ, σ^2 and, perhaps, some other true values, either in the form of point estimates, or in the form of point estimates with standard errors, confidence intervals or, preferably, distributional inferences.

[2]Transparencies of an earlier lecture with approximately the same content are available from `http://www.stat.colostate.edu/ rdavis/lectures/nber97.pdf`

[3]Davis and Dunsmuir use θ_0 as notation for ψ and T as notation for the sample size.

Note that the correlation matrix (ρ_{ij}) is given by

$$\rho_{i,j} = \begin{cases} \rho = \psi/(1+\psi^2) & \text{if } |i-j| = 1 \\ 0 & \text{if } |i-j| \geq 2 \end{cases}$$

where $\rho \in [-\frac{1}{2}, \frac{1}{2})$ is the first-order auto-correlation. This suggests a parametrization where (ψ, σ^2) is replaced by (ρ, τ^2) with ρ as defined and where $\tau^2 = (1+\psi^2)\sigma^2$ is the variance of X_i $(i = 1, \ldots, n)$. The restrictions now are $(\rho, \tau^2) \in [-\frac{1}{2}, \frac{1}{2}) \times (0, \infty)$ while H_0: $\rho = -\frac{1}{2}$ is the hypothesis of interest.

Yet another parametrization will appear in the context of Section 6.2. Here the parametrization (ν^2, η^2) consists of the two variances

$$\nu^2 = 2\sigma^2(1 - \psi + \psi^2) = 2\tau^2(1-\rho)$$
$$\eta^2 = 2\sigma^2(1 + \psi + \psi^2) = 2\tau^2(1+\rho)$$

of $X_i - X_{i+1}$ and $X_i + X_{i+1}$, respectively. Here the restrictions are

$$\frac{1}{3} \leq \frac{\eta^2}{\nu^2} = \frac{1+\rho}{1-\rho} < 3$$

and H_0: $\eta^2/\nu^2 = \frac{1}{3}$ is the null hypothesis to be tested.

Remark 1

The restrictions $\psi \in [-1, 1)$, $\rho \in [-\frac{1}{2}, \frac{1}{2})$ and $\eta^2/\nu^2 \in [\frac{1}{3}, 3)$ have been imposed to make the parameters identifiable in the sense that if X is generated as prescribed and its distribution is given, then the parameter is uniquely determined. Note that the restriction $-1 \leq \psi < 1$ has its origin in the idea that the contribution of E_i should be 'more important' than that of E_{i-1} in $X_i = E_i + \psi E_{i-1}$ (in the sense that Var $E_i \geq$ Var (ψE_{i-1})). Starting from $-1 \leq \psi < 1$, the correlation $\rho = \psi(1+\psi^2)^{-1}$ is in $[-\frac{1}{2}, \frac{1}{2})$. The inverse transformation is

$$\psi = \frac{1 - \sqrt{1-4\rho^2}}{2\rho}$$

(The other root, $(1 + \sqrt{1-4\rho^2})/2\rho$, provides values beyond the interval $[-1, 1)$ if $\rho \in [-\frac{1}{2}, \frac{1}{2})$.)

Remark 2

The situation is also of general methodological interest because the parametrizations with

$$\begin{array}{lll} (\psi, \sigma^2) & \text{as the true value of} & \theta^{(1)} = \left(\theta_1^{(1)}, \theta_2^{(1)}\right) \in \Theta^{(1)} \\ (\rho, \tau^2) & \text{as the true value of} & \theta^{(2)} = \left(\theta_1^{(2)}, \theta_2^{(2)}\right) \in \Theta^{(2)} \\ (\nu^2, \eta^2) & \text{as the true value of} & \theta^{(3)} = \left(\theta_1^{(3)}, \theta_2^{(3)}\right) \in \Theta^{(3)} \end{array}$$

are all well-motivated. Some methods of inference depend on the parametrization. The situation will then require a specific analysis. The domains $\Theta^{(1)}$, $\Theta^{(2)}$, and $\Theta^{(3)}$

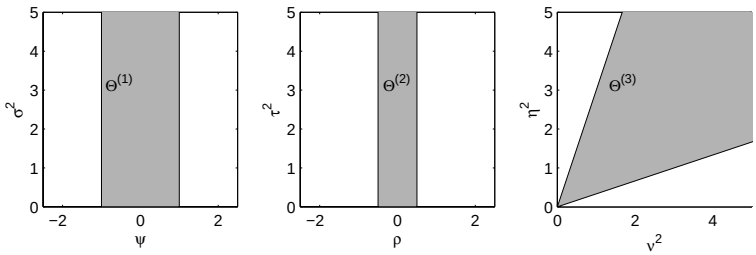


Figure 6.1: Visualization of the restricted area for each parametrization

are visualized in Figure 6.1. Other parametrizations, e.g. $\theta^{(4)} = (\eta^2/\nu^2, \tau^2)$, are of course also possible.

Preview

The case $n = 2$ to be discussed in Section 6.2 is of methodological but of no practical interest because nobody should consider the testing of H_0 or the estimation of parameters if only 2 outcomes are available. More than a few observations are necessary before it is reasonable to make any statistical inference about the true value t of any θ. Section 6.3 is about $n \geq 3$. The methodological aspects will not be treated as satisfactorily as in Section 6.2 but the exact methods to be developed are of practical interest, especially if n is neither too small, nor too large (because the normality assumption becomes questionable if n is very large while, moreover, asymptotic methods can be incorporated).

6.2 The case $n = 2$

Testing

To test H_0 in case $n = 2$, the parametrization with (ν^2, η^2) as the true value of $\theta^{(3)} \in \Theta^{(3)}$ is most convenient because the $1:1$ transformation

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 - X_2 \\ X_1 + X_2 \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \nu^2 & 0 \\ 0 & \eta^2 \end{bmatrix} \right)$$

makes it obvious that H_0: $\eta^2/\nu^2 = \frac{1}{3}$ should be tested against A: $\eta^2/\nu^2 > \frac{1}{3}$ by rejecting H_0 iff $3Y_2^2/Y_1^2 \geq F_{1,1;\alpha}$. Note that (Y_1^2, Y_2^2) is a complete sufficient statistic in the general situation and that $3Y_1^2 + Y_2^2$ is a complete sufficient statistic for the null hypothesis. The theory of multi-parameter exponential families (LEHMANN, 1986) provides that the F-test discussed is UMP invariant level α (p. 290), UMP similar level α (p. 199), and UMP unbiased level α (p. 122).

Estimating

The first part of the problem being solved, we can discuss the second part, the estimation of various parameters, preferably by means of confidence intervals or distribution functions of distributional inferences.

Point estimates
Continuing with the third parametrization, we have that Y_1^2 and Y_2^2 are best unbiased estimates of ν^2 and η^2. As $F_{1,1} \sim \text{Cauchy}^2$ does not have a finite expectation, η^2/ν^2 does not allow an unbiased estimate. The maximum likelihood method is complicated by the fact that Y_1^2 and Y_2^2 are *not* necessarily ML estimates of ν^2 and η^2 because the restriction $\frac{1}{3} \leq \eta^2/\nu^2 < \frac{1}{3}$ has to be complied with. The ML estimates $\hat{\eta}^2, \hat{\nu}^2$ are such that

$$\begin{array}{lll} y_2^2/y_1^2 \leq \frac{1}{3} & \text{implies } \hat{\eta}^2/\hat{\nu}^2 = \frac{1}{3} & (\text{in fact } \hat{\nu}^2 = \frac{1}{6}(y_1^2 + 3y_2^2),\ \hat{\eta}^2 = \frac{1}{3}\hat{\nu}^2) \\ \frac{1}{3} \leq y_2^2/y_1^2 \leq 3 & \text{implies } \hat{\eta}^2/\hat{\nu}^2 = y_2^2/y_1^2 & (\text{in fact } \hat{\nu}^2 = y_1^2,\ \hat{\eta}^2 = y_2^2) \\ y_2^2/y_1^2 \geq 3 & \text{implies } \hat{\eta}^2/\hat{\nu}^2 = 3 & (\text{in fact } \hat{\nu}^2 = \frac{1}{6}(3y_1^2 + y_2^2),\ \hat{\eta}^2 = 3\hat{\nu}^2) \end{array}$$

(where, in principle, the latter value is not allowed but this complication is of no practical interest).

Since $(1 + \rho)/(1 - \rho) = \eta^2/\nu^2$ we have that

$$\hat{\rho} = \begin{cases} -\frac{1}{2} & \text{if } r \leq -\frac{1}{2} \\ r & \text{if } -\frac{1}{2} \leq r \leq \frac{1}{2} \\ \frac{1}{2} & \text{if } r \geq \frac{1}{2} \end{cases}$$

where $r = 2x_1x_2/(x_1^2 + x_2^2)$ can vary between -1 and $+1$. From the relation $\tau^2 = \frac{1}{4}(\nu^2 + \eta^2)$ we directly obtain the ML estimate

$$\hat{\tau}^2 = \begin{cases} (x_1 + x_2)^2 & \text{if } y_2^2/y_1^2 \leq \frac{1}{3} \\ \frac{1}{2}(x_1^2 + x_2^2) & \text{if } \frac{1}{3} \leq y_2^2/y_1^2 \leq 3 \\ (x_1 - x_2)^2 & \text{if } y_2^2/y_1^2 \geq 3 \end{cases} .$$

Similarly, the relationship $\rho = \psi/(1 + \psi^2)$ and $\tau^2 = (1 + \psi^2)\sigma^2$ can be used in constructing ML estimates of (ψ, σ^2). This method is appealing to generate point estimates, but to express statistical uncertainties, other methods are needed.

Bayesian methods (likelihood inference included)
Given some parametrization we can discuss the likelihood function, incorporate a weight function w and derive the posterior density

$$q(\theta_1, \theta_2) = \frac{L_x(\theta_1, \theta_2)w(\theta_1, \theta_2)}{\int\int_\Theta L_x(\theta_1, \theta_2)w(\theta_1, \theta_2)\ \mathrm{d}\theta_1\ \mathrm{d}\theta_2}.$$

Choosing the Jeffreys prior $w(\theta) = |I(\theta)|^{1/2}$ is of some interest because it implies that the posterior densities are probabilistically coherent (JEFFREYS, 1932, and 1939, Ch. 3). Although Jeffreys's prior is widely accepted amongst Bayesians for single parameter models, in multi-parameter models its is more controversial (see, e.g., DATTA ET AL., 2000). At a deeper level even this choice can be criticized as Jeffreys himself seems to have done. 'In almost all cases we can approach this [Jeffreys] prior as the limit of a sequence of proper (normalizable) priors, with mathematically well-behaved results. If even that does not yield a proper posterior distribution, then the

robot is warning us that the data are too uninformative about either very large s or very small s to justify any definite conclusions, and we need to get more evidence before any useful inferences are possible' (JAYNES, 1996, p. 629).

For the parametrization $\theta^{(3)} = (\nu^2, \eta^2)$ we obtain the likelihood function

$$L_x^{(3)}(\theta_1, \theta_2) = \frac{1}{2\pi\sqrt{\theta_1\theta_2}} \exp\left[-\frac{1}{2}\left(\frac{(x_1 - x_2)^2}{\theta_1} + \frac{(x_1 + x_2)^2}{\theta_2}\right)\right].$$

and the information matrix

$$I_x^{(3)}(\theta_1, \theta_2) = \begin{bmatrix} \frac{1}{2\theta_1^2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{bmatrix}.$$

Hence $|I_x^{(3)}(\theta_1, \theta_2)|^{1/2} = 1/(2\theta_1\theta_2)$ and

$$q^{(3)}(\theta_1, \theta_2) = C\theta_1^{-3/2}\theta_2^{-3/2} \exp\left(-\frac{1}{2}\left(\frac{(x_1 - x_2)^2}{\theta_1} + \frac{(x_1 + x_2)^2}{\theta_2}\right)\right)$$

if $\theta_2/\theta_1 \in [\frac{1}{3}, 3)$, with the proportionality constant C obtained via integrating over $\Theta^{(3)}$, yielding

$$\begin{aligned} C^{-1} &= \int_0^\infty \int_{\theta_1/3}^{3\theta_2} L_x^{(3)}(\theta_1, \theta_2) w^{(3)}(\theta_1, \theta_2)\, \mathrm{d}\theta_2 \mathrm{d}\theta_1 \\ &= \frac{4}{\sqrt{|x_1^2 - x_2^2|}} \left(\arctan\sqrt{\frac{3(x_1 + x_2)^2}{(x_1 - x_2)^2}} - \arctan\sqrt{\frac{(x_1 + x_2)^2}{3(x_1 - x_2)^2}}\right). \end{aligned}$$

The information matrices, and hence Jeffreys likelihood inferences, for other parametrizations can either be obtained directly or by using Jacobian matrices. This yields

$$|I_x^{(1)}(\theta_1, \theta_2)|^{1/2} = \frac{1 - \psi^2}{\sigma^2(1 + \psi^2 + \psi^4)} \quad \text{and} \quad |I_x^{(2)}(\theta_1, \theta_2)|^{1/2} = \frac{1}{(1 - \rho^2)\tau^2}.$$

Although results can be derived, we are not satisfied. Furthermore, JAYNES (1996, p. 630) noted that 'while the Jeffreys prior is theoretically the correct one, it is in practice a small refinement that makes a difference only in the very small sample case'. We do not agree with the qualification as 'the' correct one because integration with respect to Lebesgue measure is only one possibility. Moreover no prior probability mass is assigned to the hypothesis H_0: $\psi = -1$ of interest.

Fisher-Neyman compromise

Our main interest is in constructing distributional inferences for η^2/ν^2, ρ and ψ using a modernization of Fisher's fiducial argument. The idea is very simple: one-sided P-values are used to express the degree of belief in one-sided hypotheses (SALOMÉ, 1998; see also FISHER, 1973, p. 63, where he, in some sense, discredits his unpolished

version this useful idea). Under certain restrictive conditions, 'optimality' of thus defined procedures can be established (they are 'best strongly similar', see KROESE ET AL., 1999). The analogy with 'best unbiased estimates' is obvious. The modernized fiducial argument is seldomly applicable and if it is applicable then the methods of inference generated may still fail to be satisfactory. Much of the work done in asymptotic theory can be regarded as in line with this modernized fiducial argument. One is establishing that methods of inference are 'asymptotically strongly similar' in the sense that 'confidence intervals based on it are asymptotically correct', etcetera. The 'exactness' of the fiducial method is very appealing. Consider the problem of making a distributional inference about η^2/ν^2. We need a procedure $Q : [0,\infty)^2 \to [\frac{1}{3}, 3)^*$ (the space of probability distributions on $[\frac{1}{3}, 3)$) such that $Q(x)$ provides our 'opinion' about the true value η^2/ν^2 of $\theta_2^{(3)}/\theta_1^{(3)}$. Let $\mathcal{G} = \{G_x |\ x \in (0,\infty)^2\}$ denote the family of distribution functions. With $G_x(z)$ we express the degree of belief in H_z: $\eta^2/\nu^2 \leq z$. We have, of course, $G_x(z) = 0$ if $z \leq \frac{1}{3}$ and $G_x(z) = 1$ if $z \geq \frac{1}{3}$, and the Fisher-Neyman compromise prescribes to define $G_x(z) = \alpha_z(x)$ where $\alpha_z(x)$ is the P-value if H_z: $\eta^2/\nu^2 \leq z$ is tested against A_z: $\eta^2/\nu^2 > z$. This, of course, is done with the F-test using Y_2^2/Y_1^2 as test statistic having the $zF_{1,1}$ distribution if $\eta^2 = z\nu^2$. We simply have

$$\alpha_z(x) = \mathrm{P}\left(F_{1,1} \geq \frac{y_2^2}{z y_1^2}\right) = \mathrm{P}\left(F_{1,1} \leq z\frac{y_1^2}{y_2^2}\right),$$

because $\mathcal{L}\left(F_{1,1}\right) = \mathcal{L}\left(F_{1,1}^{-1}\right)$ and have

$$G_x(z) = \begin{cases} 0 & \text{if } z < \frac{1}{3} \\ \mathrm{P}\left(F_{1,1} \leq z\frac{y_1^2}{y_2^2}\right) = \frac{2}{\pi}\arctan\sqrt{z\frac{y_1^2}{y_2^2}} & \text{if } \frac{1}{3} < z < 3 \\ 1 & \text{if } z \geq 3 \end{cases}$$

as our distributional inference[4]. This family is *strongly similar* in the sense that $G_X^*(z) \sim \mathcal{U}(0,1)$ holds if $\eta^2/\nu^2 = z$ (except in the case $z = 3$ where a degeneracy appears). Note that this distributional result implies that

$$\mathrm{P}\left(G_X^{-1}\left(\frac{1}{2}\alpha\right) \leq \frac{\eta^2}{\nu^2} \leq G_X^{-1}\left(1 - \frac{1}{2}\alpha\right)\right) > 1 - \alpha$$

(strict inequality is caused by the degeneracies at $\frac{1}{3}$ and 3). It can be established along the lines of KROESE ET AL. (1999) that any other family $\{G_x; x = (x_1, x_2) \in (0,\infty)^2\}$ satisfying the strong similarity condition is 'uniformly worse' than $\{G_x^*; x = (x_1, x_2) \in (0,\infty)^2\}$ from a variety of perspectives. (The degeneracies at $\frac{1}{3}$ and 3 have the effect that the conditions used in KROESE ET AL. (1999) are not satisfied. The statement

[4]The result for $\frac{1}{3} < z < 3$ is easily seen after observing that

$$f_{1,1}(u) = \frac{1}{\pi(1+u)\sqrt{u}} \quad \to \quad F_{1,1}(u) = \int_0^u \frac{\mathrm{d}x}{\pi(1+v)\sqrt{v}} = \frac{2}{\pi}\arctan\sqrt{u}.$$

is not completely correct. That is why we used ' '). Above 'fiducial inference' can immediately be transformed to obtain optimal distributional inferences for

$$\rho = \frac{\eta^2 - \nu^2}{\eta^2 + \nu^2} \quad \text{and} \quad \psi = \frac{1 - \sqrt{1 - 4\rho^2}}{2\rho}.$$

Conclusion
The case $n = 2$ is, as said before, only of methodological interest. The Neyman-Pearson test for H_0 and distributional inferences for η^2/ν^2, ρ, and ψ are fairly attractive though, of course, nobody should consider such problems seriously if n is only 2. The restrictions to the parameter space cause trouble if Bayesian inferences (with respect to continuous weight functions w) are considered. Maximum likelihood estimators are quite attractive, though solely for the purpose of point estimation.

6.3 Exact inferences in the case $n > 2$

For the parametrization (ν^2, η^2), just obtained the distributional result

$$\frac{Y_2^2}{Y_1^2} \sim \frac{1+\rho}{1-\rho} F_{1,1}$$

as the basis of *exact* inferences. A situation of more interest is that where observable random vectors $(Y_{1,1}, Y_{1,2}), \ldots, (Y_{m,1}, Y_{m,2})$ are independently distributed with the same distribution as (Y_1, Y_2). As a basis of the inference one will use the outcome v of

$$V = \frac{\sum_{j=1}^m Y_{2,j}^2}{\sum_{j=1}^m Y_{1,j}^2} \sim \frac{1+\rho}{1-\rho} F_{m,m}.$$

This solution can be enforced to our actual situation by defining $m = \lfloor \frac{n+1}{3} \rfloor$, and by taking

$$\begin{pmatrix} Y_{1,j} \\ Y_{2,j} \end{pmatrix} = \begin{pmatrix} X_{3j-1} - X_{3j-2} \\ X_{3j-1} + X_{3j-2} \end{pmatrix}$$

($j = 1, \ldots, m$, tacitly skipping the observations $X_3, X_6, \ldots$). One possibility to test $H_0 : \psi = -1$ (or, equivalently, $\rho = -\frac{1}{2}$) against $H_1 : \psi > -1$ at level α is to reject H_0 if and only if $3V \geq F_{m,m;\alpha}$. This test is UMP unbiased size α, UMP similar size α, and UMP invariant size α (see, again, LEHMANN, 1986); but, of course, only after the deletion of $X_3, X_6, \ldots$.

The results thus obtained are exact, yet not satisfactory because one third of the data has been ignored. We can develop alternative approaches with

$$V^* = \frac{\sum_{j=1}^m (X_{3j} + X_{3j-1})^2}{\sum_{j=1}^m (X_{3j} - X_{3j-1})^2} \sim \frac{1+\rho}{1-\rho} F_{m,m}$$

$$V^{**} = \frac{\sum_{j=1}^m (X_{3j+1} + X_{3j})^2}{\sum_{j=1}^m (X_{3j+1} - X_{3j})^2} \sim \frac{1+\rho}{1-\rho} F_{m,m}$$

based on the deletion of $X_1, X_4, \ldots$ and $X_2, X_5, \ldots$ respectively (assuming, for the sake of simplicity, that $n = 3m + 1$). The construction of tests, confidence intervals, and distributional inferences using V^* or V^{**} is identical to that using V, since the three have the same distribution. However, they are not independent.

Bonferroni-inequality based combinations can be obtained, but these are too 'conservative'. We would like to do better. One intuitive approach consists of combining V, V^*, V^{**} by adding the numerators and denominators separately. This provides the statistic

$$W = \frac{\sum_{i=1}^{n-1}(X_{i+1} + X_i)^2}{\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2} = \frac{1+R}{1-R}$$

where

$$R = \frac{2\sum_{i=1}^{n-1} X_i X_{i+1}}{\sum_{i=1}^{n-1}(X_i^2 + X_{i+1}^2)}$$

is a natural estimator of the autocorrelation ρ. Asymptotic theory can be developed. These intuitive manipulations, however, are not completely satisfactory either; why should the combination of V, V^*, V^{**} be done by adding numerators and denominators? In this respect, work on combining expert opinions is of some interest (cf. GENEST AND ZIDEK, 1986, GENEST ET AL., 1990). In the remainder of this chapter various attempts are described to obtain more satisfactory 'asymptotic' results. The discussion will be somewhat confusing. We are not satisfied by it. In our opinion the exact results just described (F test for H_0, strongly similar distributional inferences for $(1+\rho)/(1-\rho)$, ρ, ψ) are attractive in spite of the fact that one third of the observations is discarded from the analysis. (The loss of information is probably less because of the correlation.)

6.4 Focusing on exact tests for the null hypothesis

In his breakthrough paper, KARL PEARSON (1900) focused on the null hypothesis that certain theoretical probabilities were correct. Later generations of statisticians emphasized that the 'set of all alternatives' should be taken into account. From here we restrict the attention to the testing of H_0:$\psi = -1$ i.e. $X_i = E_i - E_{i-1}$ $(i = 1, \ldots, n)$ where the E_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. This has as a consequence that the suggestions in the last part of Section 6.3 are not necessarily useful. A different suggestion is as follows. As taking partial sums is the inverse of differencing, we consider

$$\begin{aligned} S &= \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{pmatrix} = \begin{pmatrix} X_1 \\ X_1 + X_2 \\ \vdots \\ X_1 + \ldots + X_n \end{pmatrix} \overset{\mathrm{H}_0}{=} \begin{pmatrix} E_1 - E_0 \\ E_2 - E_0 \\ \vdots \\ E_n - E_0 \end{pmatrix} \\ &\sim \mathcal{N}_n\left(0_n, \sigma^2(I_n + \iota_n \iota_n')\right). \end{aligned}$$

(ι_n is the vector of size n with all entries one). The largest eigenvalue of $I_n + \iota_n\iota_n'$ is $n+1$ with corresponding eigenvector $n^{-1/2}\iota_n$. All other eigenvalues are equal to 1, the corresponding eigenvectors can be chosen arbitrarily as long as they are perpendicular and have elements that sum up to zero. To fix a choice, we use the matrix

$$U_n = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n^2-n}} & \frac{1}{\sqrt{n^2-n}} & \frac{1}{\sqrt{n^2-n}} & \dots & \frac{-(n-1)}{\sqrt{n^2-n}} \end{bmatrix}$$

(HELMERT, 1876). We now have that $I_n + \iota_n\iota_n' = U_n'\Lambda_n U_n$ where $\Lambda_n = \text{diag}(n+1, 1, \dots, 1)$. Hence $\mathcal{L}\ (U_n S) = \mathcal{N}_n(0_n, \sigma^2\Lambda_n)$ and $T = \Lambda_n^{-1/2}U_n S \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$. Note that there is a $1:1$ correspondence between T and X. Now it is obvious that a large number of F-tests can be constructed. For each $f, g \in \mathbb{N}$, with $f + g \leq n$

$$\frac{(T_1^2 + \dots + T_f^2)/f}{(T_{f+1}^2 + \dots + T_{f+g}^2)/g} \sim F_{f,g}$$

provides a test. At first sight, $f = 1, g = n-1$ might be of particular interest. In that case, elementary computations provide that $T_1^2 = \frac{n}{n+1}\bar{S}^2$ and $T_2^2 + \dots + T_n^2 = \sum_{i=1}^n (S_i - \bar{S})^2$, where $\bar{S}$ is the average of the partial sums. But, eventually, a more even partitioning of $n = f + g$ into f and g may be more appropriate. The message is clear: the alternative hypothesis should play its part.

Testing against specific alternatives

It follows from the above that the null hypothesis allows the complete sufficient statistic $\|T\|_2^2 \sim \sigma^2\chi_n^2$. Hence, given a specific alternative (ψ, σ^2) with $\psi > -1$ (or (ρ, τ^2) with $\rho > -\frac{1}{2}$), we can try to obtain the MP similar size-α test by using the theory of tests with Neyman structure. For $n = 2$ the test suggested before is UMP similar, unbiased size-α. For $n = 3$ the situation is already very complicated because the model does not allow a 2-dimensional exponential family representation: a 3-dimensional family is needed. In case $n = 3$ (later on, results will be generalized to the case of general n) we have that

$$T = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \psi & 1 & 0 & 0 \\ \psi & 1+\psi & 1 & 0 \\ \psi & 1+\psi & 1+\psi & 1 \end{bmatrix} \begin{pmatrix} E_0 \\ E_1 \\ E_2 \\ E_3 \end{pmatrix}$$

is such that $T \sim \mathcal{N}_3(0_3, \sigma^2\Sigma_\psi)$ where

$$\Sigma_\psi = \Lambda_n^{-1/2}U_n \begin{bmatrix} \psi & 1 & 0 & 0 \\ \psi & 1+\psi & 1 & 0 \\ \psi & 1+\psi & 1+\psi & 1 \end{bmatrix} \begin{bmatrix} \psi & \psi & \psi \\ 1 & 1+\psi & 1+\psi \\ 0 & 1 & 1+\psi \\ 0 & 0 & 1 \end{bmatrix} U_n^{-1}\Lambda_n^{-1/2}$$

can, in principle, be computed for any fixed $\psi \in [-1, 1)$. (Note that Σ_{-1} is equal to the identity matrix, as we saw earlier.) Let V be an orthogonal matrix such that $\Sigma_\psi = V\widetilde{\Lambda_n}V'$ where $\widetilde{\Lambda_n} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$. Note that V is a known matrix (if ψ is given). In terms of $\tilde{T} = VT$ the problem of determining the test with Neyman-structure, optimal in (ψ, σ^2) is that where the conditional distribution of $\tilde{T}$, given the value of $\|\tilde{T}\|_2^2$ observed, has to be considered under the alternative hypothesis that the $\tilde{T}_i \sim \mathcal{N}(0, \sigma^2\lambda_i)$ are independently distributed. As

$$\mathcal{L}\left(\tilde{T}\right) = \left(\lambda_1\lambda_2\lambda_3\right)^{-1/2}\left(2\pi\sigma^2\right)^{-3/2}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{t_1^2}{\lambda_1} + \frac{t_2^2}{\lambda_2} + \frac{t_2^3}{\lambda_3}\right)\right]$$

we shall have to reject H_0 for small values of $\lambda_1^{-1}t_1^2 + \lambda_2^{-1}t_2^2 + \lambda_3^{-1}t_3^2$ given the outcome $\tilde{T}_1^2 + \tilde{T}_2^2 + \tilde{T}_3^2$.

The critical level has to be determined by studying the conditional distribution of $\lambda_1^{-1}\tilde{T}_1^2 + \lambda_2^{-1}\tilde{T}_2^2 + \lambda_3^{-1}\tilde{T}_3^2$ given $\|\tilde{T}\|_2^2$ (under H_0). This conditional distribution does not depend on σ^2 because $\|\tilde{T}\|_2^2 \sim \sigma^2\chi_3^2$ is a (complete) sufficient statistic. One can equally well study the distribution of

$$\left(\frac{\tilde{T}_1^2}{\lambda_1} + \frac{\tilde{T}_2^2}{\lambda_2} + \frac{\tilde{T}_3^2}{\lambda_3}\right)/\|\tilde{T}\|_2^2$$

because $\left(\frac{\tilde{T}_1^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_2^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_3^2}{\|\tilde{T}\|_2^2}\right)$ and $\|\tilde{T}\|_2^2$ are independent (under H_0), this follows from Basu's theorem. Moreover,

$$\mathcal{L}\left(\frac{\tilde{T}_1^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_2^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_3^2}{\|\tilde{T}\|_2^2}\right)$$

is the Dirichlet$(\frac{1}{2}, \frac{1}{2}; \frac{1}{2})$ distribution (see, e.g., WILKS, Section 7.7). For general n, this result extends to the Dirichlet$(\frac{1}{2}, \ldots, \frac{1}{2}; \frac{1}{2})$ distribution and corresponding test statistic. Elaboration of this extension is not made because the argumentation is an analogue of that for the case $n = 3$ given above, while the choice of ψ and that of the λ_i remains problematic.

6.5 Survey of the literature for the case $n = \infty$

In this section we will review existing asymptotic theory on the unit root problem.

DAVIS ET AL. (1995) and DAVIS AND DUNSMUIR (1996), amongst others, discuss the MA(1) process $y_t = \epsilon_t - \theta_0\epsilon_{t-1}$ where the $\epsilon \sim$ i.i.d.$(0, \sigma^2)$ are not necessarily Gaussian. Both articles consider the reparametrization $\theta_T = 1 - \beta/T$ where $\beta \geq 0$ and T is the sample size. The true parameter is denoted by $\theta_0 = 1 - \gamma/T$. This idea of a value of $\psi = -\theta_0 = -1 + \gamma/n$ very close to -1 may provide us with a *locally* best

similar size-α test if we incorporate the idea in Section 6.4. This is an interesting subject for further research.

DAVIS ET AL. (1995) discuss four different tests for H: $|\theta_0| = 1$ versus H: $|\theta_0| < 1$. The first one is based on $\hat{\theta}_{\text{MLE}}$, 'the value of θ which maximizes the likelihood over the interval $\theta \in [-1, 1]$', the second one on $\hat{\theta}_{\text{LM}}$, 'defined as the local maximizer of the likelihood closest to $\theta = \pm 1$'. The asymptotic distributions of $\hat{\theta}_{\text{MLE}}$ and $\hat{\theta}_{\text{LM}}$ are *not* the same. The authors prefer the Local Maximizer estimator for slightly better behavior and easier computation. The third and fourth test are the generalized likelihood ratio and Tanaka's score type test (TANAKA, 1990), respectively. DAVIS ET AL. (1995) state that 'since 'the distribution of the MLE of θ under H_0: $\theta = 1$ is unknown even asymptotically,' TANAKA (1990) was 'reluctant to use such tests as likelihood ratio test of Wald tests'.

Simulations suggest the use of the third or fourth test. Other simulations provide critical values $b_{\text{MLE}}(\alpha)$ and $b_{\text{LM}}(\alpha)$ for the Maximum Likelihood and Local Maximum estimators for β, and hence, yield tests for H_0, in case the ϵ are normally distributed. The limiting behavior of $\hat{\theta}_{\text{LM}}$ under a sequence of local values of θ_0 which converge to 1 at rate $1/T$ is considered. 'In particular, [...] $T(1 - \hat{\theta}_{\text{LM}}) \to^d \beta_0$ where β_0 is the minimizer of some stochastic process. The limit distribution has both a discrete component at the value 0, called the pile-up effect, and a continuous component' (DAVIS AND MIKOSCH, 1998, p. 100). This pile-up effect corresponds to the probability that, under the hypothesis, the LM estimator is equal to 1. For example for, again, $T = 50$ and $\alpha = 5\%$, we have $P_{H_0}(\hat{\theta}_{\text{LM}} = 1) = .649$ (exact result, DAVIS AND DUNSMUIR, 1996, p. 24). Furthermore, on basis of simulations, tests for H_0 for various levels α can be constructed. For example, when $T = 50$ and $\alpha = 5\%$, H_0 is rejected iff $\hat{\theta}_{\text{LM}} < .874$. DAVIS ET AL. (lecture notes mentioned in Footnote 1, p. 17) provide $P_{H_0}(\hat{\theta}_{\text{MLE}} = 1) = .955$. Hence, an exact level $\alpha = 0.045$ test for H_0, based on the maximum likelihood is to reject H_0 iff the MLE differs from 1.

DAVIS AND MIKOSCH discuss the MA(1) process $y_t = \epsilon_t - \theta_0 \epsilon_{t-1}$ where $|\theta_0| \leq 1$ and $\{\epsilon_t\}$ is an i.i.d. sequence of symmetric α-stable variables (see, e.g., BILLINGSLEY (1995, p. 378)) which implies that the process has infinite variance. Here the pile-up effect seems to be slightly smaller than in the finite variance case.

6.6 Discussion

The applied statistician is confronted with many challenging problems. In practice independence assumptions are very often not satisfied because time and/or space dependencies are involved. The theory of *exact* distributional inferences (or of exact confidence intervals) is often illustrated in classroom examples with i.i.d. observations. That is one of the reasons why we were challenged by Davis' lecture. One other reason is that his problem belonged to multivariate analysis with structured covariance matrices allowing an 'exact' treatment. The 'exact' results obtained (for small n) can be compared with the asymptotic results of Davis and others.

Chapter 7

Making statistical inferences about a frequency unseen: an application to ornithology

'To count or to think, that is the question'

P.R. Halmos[1]

'Whenever you can, count'

Sir F. Galton[2]

The ornithologist G.Th. de Roos is involved in a study of catching, measuring, ringing and colourringing, counting and indentifying individual *Ruddy Turnstones*, a wader species belonging to the *Charadriiformes* order. Part of his program deals with resight observations. On a particular day he visually inspects a fairly constant population to identify the ringed birds by reading their ring-number. Some ringed birds will be missed. That is why observations are repeated on other days. The issue of interest is whether, after some repetitions, De Roos can be fairly sure that he has identified all ringed birds in his population or, equivalently, that the frequency of unseen birds is zero.

There is, of course, extensive literature about the estimation of abundances, like that of the ringed birds. Most theory is concerned with an asymptotic setting. In our context the emphasis is upon the determination of the 'probability' that the frequency unseen is zero. The methods of inference we develop are based on the assumption of a bird-independent probability p_i of identifying a ringed bird on day i. In Section 7.3 results are derived on the additional assumption of equality of the p_i, an assumption which seems reasonable for the data set submitted. In Section 7.4 different p_i's are allowed. In Section 7.5 we will review the different approaches, and discuss in what cases our theory may be used.

7.1 Introduction

De Roos has collected and is still collecting biometric and moult data on Ruddy Turnstones (*Arenaria interpres*, see the drawing in Figure 7.1[3]). Having their breeding

[1] To count or to think, that is the question, *Nieuw Archief voor Wiskunde*, **13**:1, 1995.

[2] In: J.R. Newman, The World of Mathematics, Simon and Schuster, 1956.

[3] Taken from Naumann, Naturgeschichte der Vögel Mitteleuropas, Band VIII, Table 5, Gera, 1902. Digital rendering by Peter von Sengbusch, Hamburg University.

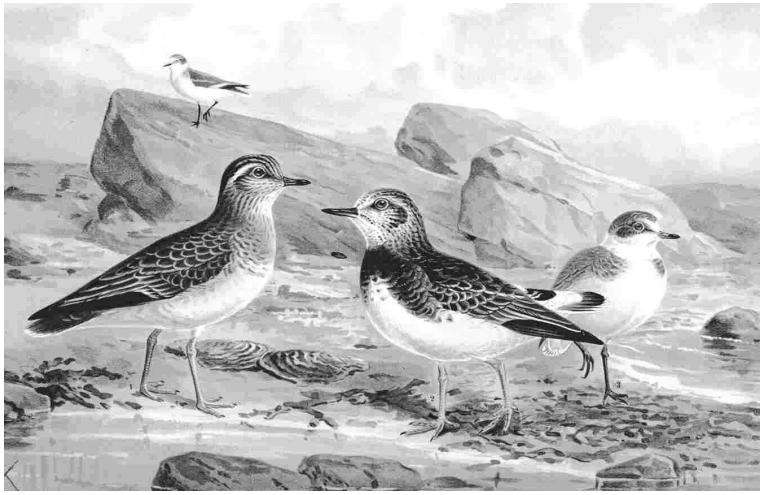


Figure 7.1: Ruddy turnstones

habitat in the high-artic tundras of Siberia and Greenland, these waders may migrate via the Frisian island Vlieland which provides good opportunities for De Roos to study them. The reason is that this place provides optimal opportunities for catching Turnstones and Purple Sandpipers because these species have their feeding areas and high tide roosts on breakwaters, stretching into the North Sea, because these act like artificial rocky shores for these birds. There the birds can be caught by torch at night during dark moon, cloudy sky and favourable tide conditions depending on the wind direction and wind force.

The birds that have been caught, were ringed and colourringed (in case they have not been ringed before), sexed, classified according to age, weighted and measured by determining various aspects of size and shape.

During the day, at high tide, the birds can be counted from a distance and, individually, identified by reading the ring-number with a telescope. Measurements were performed during the period August 1985 until April 1999. In total 1132 different Turnstones were caught and measured at least once. Also, rich spotting data are available. An analysis of the morphology (size and shape) of the birds is performed by DE BRUIN (2003). In this paper we shall restrict ourselves to the following problem: *after how many days of identifying ringed birds can one safely assume that all ringed birds being present in the population have been actually seen?*

We are fascinated by this seemingly simple problem. On a number of days, mostly three days apart, De Roos inspected a stable group of about seventy birds, some of which were ringed. Each day not all ringed birds will be identified. But by continuing the procedure regularly, the observer may feel certain that, after a certain

Date	a	b	c	d	e	f	g	h	i	j	k	l	m	n	Total
20-8		1		1								1			$s_1 = 3$
23-8		1										1		1	$s_2 = 3$
26-8		1		1					1			1	1		$s_3 = 5$
29-8									1	1	1	1	1		$s_4 = 5$
2-9			1				1		1	1	1	1			$s_5 = 6$
5-9		1	1	1					1		1	1			$s_6 = 6$
8-9		1	1	1	1	1	1				1	1			$s_7 = 8$
11-9		1	1	1			1		1			1	1	1	$s_8 = 8$
14-9			1	1	1	1	1		1	1	1		1		$s_9 = 9$
17-9	1	1				1		1	1				1		$s_{10} = 6$
20-9		1	1	1	1	1	1			1		1		1	$s_{11} = 9$
Total	1	8	6	7	3	4	5	1	7	4	5	9	5	3	68

Table 7.1: A one denotes that the bird indicated is spotted on the day indicated. All dates are in 1992. Note that $r_1 = 2$, $r_2 = 0$, $r_3 = 2$, $r_4 = 2$, $r_5 = 3$, $r_6 = 1$, $r_7 = 2$, $r_8 = 1$, *and* $r_9 = 1$ *(see the text). Note also that* $n = 68$ *identifications have been made involving* $m = 14$ *birds on* $k = 11$ *days.*

day (day 11 in our case), all ringed birds have been identified. A statistical approach to characterize the degree of (un)certainty requires a mathematical model.

Using the letters a, b, ..., n to indicate birds identified, the data appear in the form presented in Table 7.1. Apart from s_i, the number of birds seen on day i ($i = 1, \ldots, k = 11$) and reported in the table, we use the notation r_h for the number of birds identified on (exactly) h different days. Note that r_0 is the unknown value of interest: the number of ringed birds unseen. Finally we introduce $m = r_1 + \ldots + r_9 = 14$ as the total number of birds seen, $r = r_0 + m$ as the total number of ringed birds, and $n = \sum_{i=1}^{11} s_i = \sum_{h=1}^{11} h r_h = 68$ as the number of identifications made. Note that it takes until day 10 before each one of these 14 birds has been seen.

The mathematical-statistical problem

Given are the data presented in Table 7.1, the outcomes $r_1, \ldots, r_9$ in particular. Required is a distributional inference about the number $y = r_0$ of ringed birds, available in the group but not identified by the ornithologist. In particular an assessment of the probability that $r_0 = 0$, i.e. all ringed birds have been seen, is requested.

Before starting the analysis, we first note that it is *not* necessarily most appropriate to solve the second problem by identifying the epistemic probability of $r_0 = 0$ with the value $\{Q(x)\}(0)$ assigned by the distributional inference $Q(x)$ to $\{R_0 = 0\}$. The reason is that the principles to be involved in deriving $Q(x)$ are not necessarily in complete agreement with those involved in deriving the probability that the hypothesis H_0: $r_0 = 0$ is true. (Some 'fuzziness' is involved in any statistical analysis.)

Secondly we note that our problem is closely related to the proofreaders problem studied extensively in the literature. POLYA (1975) studied this problem for two

proofreaders ('days' in our context) who read, independently of each other, the same manuscript. Let $A+C$ and $B+C$ denote the number of misprints found by each reader, where C is the number of commonly found misprints and let M be the (unknown) total number of misprints, then $M-A-B-C$ is the number of undiscovered misprints. Polya's estimate for this number is AB/C, the statistical uncertainties involved are derived using the δ-method. In YANG ET AL. (1982) an 'optimal stopping rule' for rereading the manuscripts is discussed. In comparison with the proofreading problem, our problem has the advantage that probabilistic independence assumptions are less awkward.

At the time we developed our theory, we did not yet have access to the data reported in Table 7.1. The examples suggested to us were such that the number k of days is so small (yet larger than 2) that it is practically impossible to falsify the hypothesis of bird-independent 'experimental' probabilities. The capture-mark-recapture literature (e.g. OTIS ET AL., 1978, and WHITE ET AL., 1982) and the software package `MARK`[4] (suggested to us by J.B. Hulscher, personal communication) emphasize that 'tests for equal catchability' or 'equal identifiability' should be performed. While we worried about day-effects we somewhat overlooked the possibility (evidently present in the proofreaders problem) of bird-effects. We shall return to this issue in Section 7.5.

7.2 Primary approaches

We have the outcomes $(r_1, \ldots, r_{11}) = (2, 0, 2, 2, 3, 1, 2, 1, 1, 0, 0)$ and need the latent outcome r_0 or, more precisely, the 'probabilities' that $r_0 = 0$, 1, 2, ... respectively. There are, of course, many approaches to perform such extrapolation. One such approach is to assume that $r_0, r_1, \ldots, r_k$ are outcomes of independent Poisson variables with parameters λ_h satisfying some model $\lambda_h = \lambda_\theta(h)$, e.g. with $\log \lambda_h = \alpha + \beta h$. The extensive analysis to be made in the sequel will contain also some ad hoc modelling as well.

The probabilistic context

In Table 7.1 the letters a, b, ..., n are used to indicate the 14 birds. It could be, for instance, that the older birds have a central position in the group of birds on the peer, and are therefore less easily identifiable. Other explanations for deviant stochastic behaviour are also available. Nevertheless we will regard the birds as probabilistically homogeneous, and the ringed ones as 'red balls in an urn'. Also time effects are, at least for the moment, ignored, and the assumption is made that the population on each of the 11 days is the same (actually, the total number of birds, ringed and unringed, varies between 60 (on day 1) and 72 (on day 10)). We would like to assume that there is an unknown ('experimental') probability p, independent of i and j, that the j-th one of the ringed birds is seen on day i. Here $j = 1, \ldots, r = \sum_{h=0}^{k} r_h$ and $i = 1, \ldots, k(= 11)$. Note that $r = r_0 + m = r_0 + 14$. Making some independence

[4]Developed by G.C. White, Department of Fishery and Wildlife Biology, Colorado State University. Available at `http://www.cnr.colostate.edu/~gwhite/mark/mark.htm`. Documentation can be found at `http://www.cnr.colostate.edu/class_info/fw663/Mark.html`.

assumption in addition, the essence of Table 7.1 is captured in the Kolmogorovian setting $(\Omega, \mathcal{F}, \mathrm{P})$ where $\mathcal{F}$ is the space of all $k \times r$ matrices ω where

$$\omega_{i,j} = \begin{cases} 1 & \text{if bird } j \text{ is seen on day } i \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathrm{P}(\{\omega\}) = p^n (1-p)^{kr-n}$$

where, as indicated before, $n = \sum_{h=1}^{k} h r_h = 68$ is the total number of identifications made from the theoretical maximum kr which appears if all r ($r \geq 14$) birds are seen on all $k = 11$ days. The random variables (some are 'statistics' in the sense that their outcome is available, some are not) we are interested in are

1. $S_i(\omega) = \sum_{j=1}^{r} \omega_{i,j}$ the number of birds identified on day i;
2. $T_j(\omega) = \sum_{i=1}^{k} \omega_{i,j}$ the number of times bird j is seen;
3. $R_h = \{j | T_j = h\}$ the number of birds seen on h of the k days;
4. $M = \sum_{h=1}^{k} R_h$ the number of birds seen at least once;
5. $N = \sum_{h=1}^{k} h R_h = \sum_{j=1}^{k} S_j$ the total number of identifications made.

For this probabilistic model, the following statements are trivial:

1. $S_1, \ldots, S_k$ are i.i.d., $S_i \sim B(r, p)$;
2. $T_1, \ldots, T_r$ are i.i.d., $T_j \sim B(k, p)$;
3. $(R_0, \ldots, R_k) \sim \text{Multinomial}(r; (1-p)^k, \binom{k}{1} p (1-p)^{k-1}, p^k)$;
4. $M \sim B(r, 1-(1-p)^k)$;
5. $N \sim B(rk, p)$;
6. Given $\{T_k \geq 1\}$, the conditional distribution of T_j is defined by the probabilities

$$\mathrm{P}(T_j = c | T_j \geq 1) = \frac{\binom{k}{c} p^c (1-p)^{k-c}}{1-(1-p)^k} \quad (c = 1, \ldots, k);$$

7. The conditional distribution of N, given $\{M = m\}$ corresponds to that of the sum $\sum_{g=1}^{m} N_g$ of independent random variables $N_1, \ldots, N_m$, all having the distribution specified under 6.

7.3 Obtaining a solution by ignoring day effects

We have already stated that we like to assume the existence of a *constant* experimental probability p. Is this realistic? It would apply if the numbers $s_1, s_2, \ldots, s_{11}$ are outcomes of independent $B(r, p)$ variables, with both r and p unknown. Our observations $s_1, \ldots, s_{11}$ provide no convincing evidence to reject the null hypothesis

that they came from a $B(14, \frac{\bar{s}}{14} = .44)$ (or, possibly, $B(15, .44)$) distribution. In the current context the assumption of constant probability seems reasonable. However, there are also various arguments against the hypothesis. Even if the ornithologist proceeds in equivalent manners on different days, weather conditions, for instance, are likely to play a role, providing the possibility that the birds are on one day less likely to present themselves to the telescope in the appropriate position than on the other. This suggests that the assumption of fixed p might be unrealistic. In this section, however, we will use the assumption of a constant probability and it seems reasonable to concentrate the attention on the outcome $x = (m, n)$ of (M, N) and on the number of days k taken into account. From Table 7.1 we have $k = 11$ and $x = (14, 68)$.

For these outcomes we shall estimate p by equating

$$\mathbf{E}\,(N|M = m) = m\mathbf{E}\,(T_j|T_j \geq 1) = mkp/(1 - (1 - p)^k)$$

to the outcome n or, equivalently, by computing the estimate $\hat{p}$ as the solution of

$$p = \frac{n}{mk}(1 - (1 - p)^k).$$

Finally, ignoring the uncertainties involved in the estimation of p and concentrating the attention on the outcome m of

$$M \sim B(r, 1 - (1 - \hat{p})^k),$$

the distributional inference $Q(m)$ about $r_0 = r - m$ is composed by taking

$$Q(m) = \text{NegBin}(m, q).$$

where

$$q = 1 - (1 - \hat{p})^k$$

The steps behind this construction are as follows. We concentrate the attention on constructing $Q(m)$ such that the corresponding value $G_m(\theta)$ of its distribution function equals the degree of belief in H_θ: $r \leq \theta$. This fiducial approach (see SALOMÉ, 1998 and KROESE ET AL., 1999) yields the $Q(m)$ mentioned above. KARDAUN AND SCHAAFSMA (2003) show that the distributional inference

$$\tilde{Q}(m) = (\tfrac{1}{2} + \tfrac{1}{2}q)\text{NegBin}(m, q) + (\tfrac{1}{2} - \tfrac{1}{2}q)\text{NegBin}(m + 1, q)$$

provides a slight refinement (in the sense that $\tilde{Q}$ is weakly similar, whilst Q is not). Above methods provide $\hat{p} = .441$ as the solution of $14 \cdot 11p = 68(1 - (1 - p)^{11})$ and $Q(m) = \text{NegBin}(14, .998)$ which assigns probability

$$\{Q(m)\}(0) = .998^{14} = .977$$

k_k	2	3	4	5	6	7	8	9	10	11
m_k	4	6	8	10	10	12	12	12	14	14
n_k	6	11	16	22	28	36	44	53	59	68
$\hat{p}$	.667	.559	.456	.408	.454	.419	.455	.490	.420	.441
$\{Q\}$	.624	.584	.480	.470	.765	.762	.910	.972	.941	.977
$\{\tilde{Q}\}$	.613	.573	.469	.460	.759	.758	.909	.972	.940	.977

Table 7.2: First three rows: number of birds spotted at least once (m_k) and total number of observations (n_k); both after days k_k. Last three rows: estimate $\hat{p}$ and the epistemic probabilities $\{Q(x)\}(0)$ and $\{\tilde{Q}(x)\}(0)$.

to the hypothesis H_0: $r_0 = 0$ (or, equivalently, $r = 14$). Note that $\tilde{Q}(m)$ assigns the same probability up to three digits (it is trivial that Q and $\tilde{Q}$ converge to each other when $m \to \infty$ or $p \to 1$).

Of course, also after the first, second, ..., etcetera, day, De Roos could have tried to make a distributional inference about the number of unseen birds. The outcomes (k_k, m_k, n_k) $(k_k = 1, \ldots, k = 11)$ of interest as well as the epistemic probabilities $\{Q(x)\}(0) = (1 - (1 - \hat{p}^k)^m)$ are displayed in Table 7.2. The last two rows of this table contains the epistemic probabilities $\{Q(x)\}(0)$ and $\{\tilde{Q}(x)\}(0)$assigned in favor of H_0: $r_0 = 0$, i.e. the degree of belief in the statement that all ringed birds have been identified at least once. It is not until after the ninth day that we are sufficiently certain (at $\alpha = 5\%$) that De Roos has seen all ringed birds available. Nevertheless the next day (day 10), two new birds appeared (birds a and h).

Though we are completely satisfied by this approach in the present context, we are not in general, because the existence of a fixed probability p, independent of day and bird, is often questionable. Another drawback is that the statistical uncertainties involved in the construction of p have been ignored. In the next section theory will be presented to deal with the case that day effects are not ignorable.

7.4 Obtaining a solution by taking day effects into account

Concerning the question whether the s_i (see Table 7.1) are sufficiently alike to accept H_0: $p_1 = \ldots = p_{11}$, we decided to accept H_0 at the beginning of Section 7.3. However, as stated earlier, there are reasons to question the absence of day effects. It would be nice if we could adapt the theory of the previous section to the situation where on day i a probability p_i is involved, and no assumptions are made about the equality of these p_i.

To cope with this situation we shall apply a 'dirty trick': by conditioning to $S_1 = s_1, \ldots, S_k = s_k$ we can get rid of $p_1, \ldots, p_k$. This is very convenient and natural, but not compelling because $s_1, \ldots, s_k$ contains some information about r, e.g. the logical fact that $r \geq s = \max(s_1, \ldots, s_k)$.

Given the number s_i of ringed birds on day i, all $\binom{r}{s_i}$ combinations of s_i birds from the r ringed birds available have the same probability $1/\binom{r}{s_i}$ to be the s_i ones seen. Following a recommendation by A.J. Stam (personal communication), Ω consists of the sets

$$\{v_{i,1}, \ldots, v_{i,11}\} \subset \{1, \ldots, r\}$$

of the index numbers of the s_i birds seen on day i. Assuming independence between days (no bird-effect), the conditional probability distribution P on $(\Omega, \mathcal{F})$ is determined by

$$\mathrm{P}(\{\omega\}) = \begin{cases} 1/\left(\binom{r}{s_1}\binom{r}{s_2}\ldots\binom{r}{s_k}\right) & \text{if } S_i(\omega) = s_i \ (i = 1, \ldots, k) \\ 0 & \text{otherwise.} \end{cases}$$

The distribution of the relevant observable $M = \sum_{j=1}^{r} \mathbf{1}_{\{T_j \geq 1\}} = r - R_0$ can now be studied for any a priori possible value $\theta \in \Theta = \{s, s+1, \ldots\}$ of r (remember that $s = \max(s_1, s_2, \ldots)$). Standard theory developed by LOÈVE (1955, see also PARZEN, 1960) provides that

$$\mathrm{P}(M = m) = \sum_{j=m}^{r} (-1)^{j-m} \binom{j}{m} Q_j$$

and that

$$\mathrm{P}(M \geq m) = \sum_{j=m}^{r} (-1)^{j-m} \binom{j-1}{m-1} Q_j$$

where

$$\begin{aligned} Q_0 &= 1 \\ Q_1 &= \sum_{j=1}^{r} \mathrm{P}(A_j) = r \prod_{i=1}^{k} (1 - s_i/r) \\ Q_2 &= \sum_{j_1=1}^{r} \sum_{j_2=j_1+1}^{r} \mathrm{P}(A_{j_1} A_{j_2}) \\ &\vdots \\ Q_r &= \mathrm{P}(A_1 A_2 \ldots A_r) \end{aligned}$$

and $A_j = \{T_j = 0\}$ $(j = 1, \ldots, r)$.

As r is unknown, we introduce the auxiliary random variable M_θ such that M_θ has the distribution which M would have had if $r = \theta$. The result just mentioned provides the relevant 'physical' probabilities $p_\theta(\mu) = \mathrm{P}(M_\theta = \mu)$ $(\mu = s, s+1, \ldots, \theta)$ whether

or not a priori probabilities are specified, we are interested in the construction of (epistemic) posterior probabilities

$$q_m(\theta) \quad (\theta = m, m+1, \ldots)$$

specifying the opinion we should have about r after observing the outcome m of M. The posterior probability $q_m(m)$ is of particular interest because it refers to the (epistemic) probability that De Roos has identified all ringed birds in the population. There, obviously, are two issues involved:

1. The determination of $p_\theta(\mu)$;
2. How to convert the $p_\theta(\mu)$, given the outcome m of M, into the $q_m(\theta)$.

Finally, there is a minor issue with respect to the determination of $q_m(m)$ or, equivalently, of the 'probability' that $r_0 = 0$ (and $r = m$). These issues are now successively dealt with. (The last mentioned minor issue does not seem of much interest in this application.) The distribution of M has been derived in the above. Its expectation is given by

$$\begin{aligned} \mathbf{E}\,(M) &= \sum_{j=1}^{r} \mathrm{P}(T_j \geq 1) \\ &= \sum_{j=1}^{r} (1 - \mathrm{P}(T_j = 0)) \\ &= r - r \prod_{i=1}^{k} \left(1 - \frac{s_i}{r}\right) \end{aligned}$$

because $\{T_j = 0\} = \{\omega | \omega_{ij} = 0 \ i = 1, \ldots, k\}$ is the event which occurs if bird j is never seen. The (conditional) probability that bird j is not seen on day i is equal to $1 - s_i/r$.

Thus having obtained

$$\mathbf{E}\,(M_\theta) = \theta \left(1 - \prod_{i=1}^{k} (1 - \frac{s_i}{\theta})\right)$$

we shall content ourselves by providing approximate $p_\theta(\mu)$'s by equating $\mathcal{L}\,(M_\theta)$ to the distribution on $\{s, s+1, \ldots, \theta\}$ which maximizes the entropy

$$-\sum_{\mu=s}^{\theta} p(\mu) \log p(\mu)$$

under the restrictions

$$p(\mu) \geq 0, \quad \sum_{\mu=s}^{\theta} p(\mu) = 1, \quad \sum_{\mu=s}^{\theta} \mu p(\mu) = \mathbf{E}\,(M_\theta)$$

having $\tilde{p_\theta}(\mu) = \mathrm{e}^{c_\theta \mu - \psi(\theta)} \quad (\mu = s, \ldots, \theta)$ as its solution with $\psi(\theta) = \log \sum_{\mu=s}^{\theta} \mathrm{e}^{c_\theta \mu}$ and with c_θ such that $\sum \mu \tilde{p_\theta}(\mu) = \mathbf{E}\, M_\theta$. We believe that it is not reasonable, in the context of Table 7.1 with $k = 11$, $s = 9$, $m = 14$, $n = 68$, to convert the $\tilde{p_\theta}(\mu)$ into posterior probabilities $q_m(\theta)$ by normalizing the likelihood function $l_m(\theta) = \tilde{p_\theta}(m)$ or, equivalently, by using the formal Bayes approach with improper prior $w(\theta) = 1$ ($\theta = s, s+1, \ldots$). It seems more reasonable to try to comply with some requirement of weak similarity by using some form of Fisher's fiducial argument, e.g. that where the distribution function

$$G_m(z) = \sum_{\theta=m}^{z} q_m(\theta)$$

of the distributional inference about r is equated to the P-value $\alpha_z(m) = \mathrm{P}(M_z \geq m)$ or to the symmetrized P-value $\tilde{\alpha_z}(m) = \frac{1}{2}\mathrm{P}(M_z \geq m) + \frac{1}{2}\mathrm{P}(M_z \geq m+1)$where, of course, the approximate values

$$\begin{aligned} \mathrm{P}(M_z \geq m) &\approx \sum_{\mu=m}^{z} \tilde{p_z}(m) \\ &= \mathrm{e}^{-\psi(z)} \sum_{\mu=m}^{z} \mathrm{e}^{c_z \mu} \\ &= \mathrm{e}^{-\psi(z)} \left(\mathrm{e}^{c_z(z+1)} - \mathrm{e}^{c_z(m)} \right) / (1 - \mathrm{e}^{c_z}) \end{aligned}$$

are used. With respect to the determination of $q_m(m)$, we believe that it is appropriate, in the present context, to use a Bayesian approach where the data-dependent prior $w(\theta) = \frac{1}{2} \quad (\theta = m, m+1)$ is used. It provides us with

$$\begin{aligned} q_m(m) &= \frac{p_m(m)}{p_m(m) + p_{m+1}(m)} \\ &\approx 1/(1 + \mathrm{e}^{(c_{m+1} - c_m)m - \psi(m+1) + \psi(m)}) \end{aligned}$$

Application of above theory to our data yields the following. Remember that $s = \max(s_i) = 9$, $k = 11$, and $m = 14$. Taking $\theta = 14$ provides $\mathbf{E}\, M_{14} = 14(1 - \prod(1 - s_i/14)) = 13.984$, and the maximum entropy solution $\tilde{p}_{14}(\mu) \propto \exp(c_{14}\mu)$ with $c_{14} = 4.175$, hence

$$(\tilde{p}_{14}(9), \ldots, \tilde{p}_{14}(14)) = (0, 0, 0, 0, .015, .985).$$

For the denominator of $q_m(m)$, given above, it is also necessary to look at the situation $\theta = m + 1$. This provides $\mathbf{E}\, M_{15} = 14.968$, and $\tilde{p}_{15}(\mu) \propto \exp(c_{15}\mu)$ with $c_{15} = 3.470$, hence

$$(\tilde{p}_{15}(9), \ldots, \tilde{p}_{15}(15)) = (0, 0, 0, 0, .001, .030, .969).$$

k	2	3	4	5	6	7	8	9	10	11
m	4	6	8	10	10	12	12	12	14	14
s	3	5	5	6	6	8	8	9	9	9
$\tilde{p}_m(m)$	.750	.750	.681	.664	.836	.839	.940	.984	.958	.985
$\tilde{p}_{m+1}(m)$	.323	.302	.256	.250	.195	.186	.099	.036	.069	.030
$q_m(m)$	.700	.713	.727	.726	.811	.819	.905	.965	.933	.970

Table 7.3: First three rows: number of birds spotted at least once (m) at day k and maximum number of observations at a day s until day k. Last rows: the derivation of the epistemic probabilities $\{Q(x)\}(0)$.

This provides us with *a* probability that $r = 14$ (which implies $r_0 = 0$), namely

$$q_{14}(14) = \frac{\tilde{p}_{14}(14)}{\tilde{p}_{14}(14) + \tilde{p}_{15}(14)} = \frac{.985}{.985 + .030} = .970,$$

so, according to this model, the hypothesis that all ringed birds available have been seen is not rejected at $\alpha = 5\%$. (The minor issue of whether $q_m(m)$ is appropriate is skipped over here.)

Note that, similar to the previous section, also after the first, second, etcetera, day, we could have made a statement about the number of unseen birds. The 'probabilities' $q_m(m) =$ expressed in Table 7.3 denote the probabilities assigned in favor of H_0: $r_0 = 0$, i.e. the degree of belief in the statement that are ringed birds are spotted at least once, after the certain day. The agreement between the statements $Q(x)(0)$ (or $\{\tilde{Q}(x)\}$) made in Section 7.3, Table 7.2 and $q_m(m)$ made in Section 7.4, Table 7.3 is very satisfying.

The theory in this section is not completely compelling, since much more information is needed about the accuracy of the maximum-entropy approximation to the true distribution $\mathcal{L}\,(M_\theta)$.

7.5 Discussion

In 1994 P.R. HALMOS gave a lecture in Groningen in honor of Johann Bernoulli. The lecture was entitled *'To count or to think, that is the question'*. He made it very clear that he was a rationalist and an idealist. He wanted to think: he was a mathematician. Empiricists like Francis Galton, prefer to count. It is easy to conclude from everything in this chapter that the answer is that one has to count *and* to think. This combination of activities, however, is not as easy as it sounds.

While De Roos was involved in his experiments, we were developing the theory for the problem indicated without having access to the real data. We were thinking about a small number of consecutive days (say, $k = 6$) such that falsification of the hypothesis of bird-independent experimental probabilities would not be feasible. In

such situations the research worker may decide to make the assumption of 'no bird effects'. If the hypothesis $p_1 = \ldots = p_k$ is acceptable (like in Table 7.1), the theory of Section 7.3 is applicable. If this hypothesis is not reasonable, then one may use the theory of Section 7.4. Note that the results reported in Tables 7.2 and 7.3 are not much different. This could be expected on the basis of the acceptability of the hypothesis $p_1 = \ldots = p_k$. In practice, it may very well happen that day-effects are present. Ornithologist J.B. Hulscher was dealing with counting all ringed Oystercatchers (*Haematopus ostralegus*) on Schiermonnikoog (another Frisian island) in winter, the birds being ringed in spring. In his experience (personal communication) the frequencies $s_1, \ldots, s_k$ of birds counted on k consecutive days were too much different to assume a common p. This implies that the theory in Section 7.4 is of practical interest as well.

However, if we study the frequencies $(r_1, \ldots, r_{11}) = (2, 0, 2, \ldots, 0)$ of Turnstones with 1, 2, ..., 11 identifications then these frequencies are too much dispersed to satisfy the probabilistic assumption of no bird-effect. This leaves us with an awkward issue. We have developed and applied theory (more exact than that of, e.g. OTIS ET AL. 1978, and WHITE ET AL., 1982) to a table which is in conflict with the necessary assumption of 'equal watchability'. In less extensive applications this assumption will not be rejected on the basis of available data. If one does not accept the existence of 'experimental' probabilities p_i (possibly day-dependent) then only ad-hoc methods are possible. E.g. those applied by STAM in his article about numismatics (1987). He fitted a truncated negative binomial distribution instead of a Poisson one to observations of the kind $(r_1, r_2, \ldots)$. If we have agreed that we have to count *and* to think, then the next question is how to combine these two activities. The conclusion from this chapter is that this may be awkward.

Appendices

Appendix A

Proof of Theorem 4.1

The bivariate normality follows from general U-statistic theory, and the central limit theorem. Theorem 4.1 is essentially equal to the following three lemmata.

Lemma A.1 *Under* $H_0 : f = \psi$ *we have* Var $\varepsilon = \frac{1}{n}\sigma^2 = \frac{1}{12n}$.

Lemma A.2 *Under* $H_0 : f = \psi$ *we have*

$$\text{Var } \delta = \frac{1}{45}\frac{n+3}{n(n-1)} \to \tau^2 = \frac{1}{45}.$$

Lemma A.3 *Under* $H_0 : f = \psi$ *we have* Cov $(\varepsilon, \delta) = 0$.

Proof of Lemma A.1
If $Y_i \sim$U$(0,1)$ i.i.d. then Var $(Y_i - \frac{1}{2}) = \frac{1}{12}$ and Var $(\frac{1}{n}\sum(Y_i - \frac{1}{2})) = \frac{1}{12n}$. □

Proof of Lemma A.2
This result can be found in Nair (1936)[1] Alternatively, it can be derived along the lines of Hoeffding (1948, p. 308) and Fraser (1959, p.230), which will be repeated here. The variance of Gini's mean difference can be written as

$$\text{Var } g = \frac{2}{n(n-1)}[2\zeta_1(n-2) + \zeta_2]$$

with[2]

$$\begin{aligned}
\Delta &= \int_0^1 \int_0^1 |y_1 - y_2| \, dy_1 \, dy_2 = \frac{1}{3} \\
\zeta_1 &= \int_0^1 [\int_0^1 |y_1 - y_2| \, dy_2]^2 \, dy_1 - \Delta^2 \\
&= \int_0^1 (y_1^2 - y_1 + \frac{1}{2})^2 \, dy_1 - \Delta^2 = \frac{7}{60} - \frac{1}{9} = \frac{1}{180} \\
\zeta_2 &= \int_0^1 \int_0^1 (y_1 - y_2)^2 \, dy_1 \, dy_2 - \Delta^2 \\
&= 2\text{Var } y_i - \left(\tfrac{1}{3}\right)^2 = \tfrac{1}{18}.
\end{aligned}$$

[1] Lomnicki (1952) showed that there was an error in Nair's formulations, though in the uniform case the results were true (cf. Kendall and Stuart (1958, p. 241–242)).

[2] Note that the formula for ζ_1 in Fraser (1959, p.230) is wrong.

So,

$$\begin{aligned} \text{Var } g &= \frac{2}{n(n-1)}\left(\frac{2(n-2)}{180}+\frac{1}{18}\right) = \frac{1}{45}\frac{n+3}{n(n-1)} \\ \tau^2 &= \lim_{n\to\infty} n \text{ Var } g = \frac{1}{45}. \end{aligned}$$

□

Proof of Lemma A.3
Approach 1
Of course, Cov (ε, δ) = Cov $(\bar{u}, g)$. That Cov $(\bar{u}, g) = 0$ is easily seen with this symmetry-argument: define $y_i^* = 1 - y_i$. Then, of course, $\bar{u^*} = 1 - \bar{u}$ is also transformed, yet $g^* = g$ remains the same. By definition

$$\text{Cov } (\bar{u}, g) = \text{Cov } (1 - \bar{u}^*, g^*) = -\text{Cov } (\bar{u}^*, g^*).$$

Therefore, the covariance is equal to zero.

Approach 2
The covariance between $\bar{u}$ and g can be written as a function of the covariances of the order statistics,

$$\begin{aligned} \text{Cov } (\bar{u}, g) &= \text{Cov } \left(\sum \frac{u_{[i]}}{n}, \sum \frac{2(2i-n-1)}{n(n-1)} u_{[i]}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{2(2j-n-1)}{n^2(n-1)} \text{Cov } (u_{[i]}, u_{[j]}) \\ &= \frac{2(2j-n-1)}{n^2(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} (2j-n-1) \text{Cov } (u_{[i]}, u_{[j]}) \end{aligned}$$

To compute the Cov $(u_{[i]}, u_{[j]})$, we need the joint distribution function of $U_{[i]}$ and $U_{[j]}$, which is, according to WILKS (1962, eq. 7.7.6), for $i < j$

$$\begin{aligned} f_{U_{[i]},U_{[j]}}(u,v) &= c_1 u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j} \\ \text{with } c_1 &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(j-i)\Gamma(n+1-j)} \end{aligned}$$

and $0 < u < v < 1$, $1 \le i < j \le n$. This joint probability density function is that of the ordered bivariate Dirichlet$(i, j-i, n+1-j)$ distribution. (This result can also be found using KENDALL AND STUART, 1958, p. 325, who provides the formula for the joint distribution for general F.) Hence

$$\begin{aligned} \text{Cov } (u_{[i]}, u_{[j]}) &= \mathbf{E}\, u_{[i]}u_{[j]} - \mathbf{E}\, u_{[i]}\mathbf{E}\, u_{[j]} \\ &= \mathbf{E}\, u_{[i]}u_{[j]} - \frac{ij}{(n+1)^2}. \end{aligned}$$

The expectation of the product being

$$\begin{aligned}
\mathbf{E}\, U_{[i]}U_{[j]} &= \int_0^1 \int_0^v uv f_{U_{[i]},U_{[j]}}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\
&= c_1 \int_0^1 v(1-v)^{n-j}\left[\int_0^v u^i (v-u)^{j-i-1}\,\mathrm{d}u\right]\,\mathrm{d}v \\
&= c_1 \int_0^1 v(1-v)^{n-j}\left[v^s \mathrm{Beta}(r+1,s-r)\right]\,\mathrm{d}v \\
&= c_1 c_2 \int_0^1 v^{j+1}(1-v)^{n-j}\,\mathrm{d}v \\
&= \tfrac{\Gamma(n+1)}{\Gamma(i)\Gamma(j-i)\Gamma(n-j+1)} \tfrac{\Gamma(i+1)\Gamma(j-i)}{\Gamma(j+1)} \tfrac{\Gamma(j+2)\Gamma(n-j+1)}{\Gamma(n+3)} \\
&= \tfrac{i(j+1)}{(n+1)(n+2)}.
\end{aligned}$$

That $c_2 = \mathrm{Beta}(r+1, s-r)$ can be seen through

$$\int_0^v u^i (v-u)^{j-i-1}\,\mathrm{d}u = v^j \int_0^1 w^i (1-w)^{j-i-1}\,\mathrm{d}w = \mathrm{B}(i+1, j-i)v^j.$$

Using $\mathbf{E}\, U_{[i]}U_{[j]}$, we can calculate the covariance

$$\mathrm{Cov}\,(U_{[i]}U_{[j]}) = \begin{cases} \tfrac{i(j+1)}{(n+1)(n+2)} - \tfrac{ij}{(n+1)^2} = \tfrac{i(n+1-j)}{(n+1)^2(n+2)} & \text{if } i < j \\ \tfrac{i(n+1-i)}{(n+1)^2(n+2)} & \text{if } i = j \\ \tfrac{j(n+1-i)}{(n+1)^2(n+2)} & \text{if } i > j \end{cases}$$

Note that the formula for $i > j$ immediately follows from that for $i < j$ by swapping the i and j. The formula for $i = j$ is easily derived by having that $\mathrm{Cov}\,(U_{[i]}U_{[j]}) = \mathrm{Var}\,(U_{[i]})$ and $U_{[i]} \sim \mathrm{Beta}(i, n+1-i)$. Note that this formula coincides with the formula for $i < j$ as well as the formula for $i > j$. The final step is to use all these covariances to derive that $\mathrm{Cov}\,(\bar{u}, g) = 0$, and, hence, $\rho = 0$. $\mathrm{Cov}\,(\bar{u}, g) =$

$$\begin{aligned}
&= \frac{2}{n^2(n-1)} \sum_{j=1}^{n}\sum_{i=1}^{n}(2j-n-1)\mathrm{Cov}\,(u_{[i]}, u_{[j]}) \\
&= \tfrac{2}{(n-1)n^2(n+1)^2(n+2)} \left(\sum_{j=1}^{n}\sum_{i=1}^{j-1}(2j-n-1)i(n+1-j) + \right. \\
&\quad \left. + \sum_{j=1}^{n}(2j-n-1)j(n+1-j) + \sum_{j=1}^{n}\sum_{i=j+1}^{n}(2j-n-1)j(n+1-i) \right) \\
&= \tfrac{2}{(n-1)n^2(n+1)^2(n+2)} \left(\tfrac{(n-1)n(n+1)(n+2)(n+3)}{120} + 0 - \tfrac{(n-1)n(n+1)(n+2)(n+3)}{120} \right) \\
&= 0 \qquad \square
\end{aligned}$$

Appendix B

Critical values for the $||f_n^{(m)} - \psi||_1$-test statistic

The Tables B.1, B.2, and B.3 are each constructed on the basis of the following simulation experiment. For $m = 2$, 3, and 4, and for each value of n, a sample of size n was drawn from the standard uniform distribution providing an outcome $t_n^{(m)}$ of the test statistic $T_n^{(m)}$. This process is repeated 100 000 times. (Simulation studies showed that this number of replications suffices to obtain reliable critical values.) The percentiles are taken from the empirical distribution of $T_n^{(m)}$. (Note: these values are computed using the true $t_n^{(m)}$, not the approximation mentioned in Section 4.4 for $m = 2$.)

n	.90	.75	.50	.25	.10	.05	.025	.01
5	.0499	.0838	.1422	.2311	.3241	.3801	.4302	.4865
6	.0439	.0746	.1283	.2093	.2949	.3482	.3954	.4466
7	.0394	.0675	.1177	.1934	.2717	.3215	.3641	.4133
8	.0365	.0627	.1102	.1807	.2539	.3007	.3436	.3906
9	.0336	.0582	.1029	.1700	.2401	.2842	.3224	.3664
10	.0320	.0552	.0979	.1613	.2276	.2684	.3055	.3473
12	.0286	.0494	.0889	.1468	.2075	.2463	.2802	.3207
14	.0259	.0455	.0818	.1357	.1914	.2268	.2574	.2950
16	.0241	.0424	.0765	.1274	.1798	.2125	.2415	.2769
18	.0225	.0395	.0718	.1198	.1700	.2013	.2286	.2620
20	.0212	.0374	.0685	.1132	.1604	.1899	.2164	.2486
22	.0201	.0354	.0651	.1085	.1531	.1816	.2065	.2379
24	.0191	.0339	.0623	.1037	.1471	.1741	.1990	.2283
26	.0185	.0326	.0596	.0995	.1406	.1667	.1908	.2178
28	.0176	.0312	.0576	.0962	.1361	.1613	.1838	.2118
30	.0171	.0303	.0555	.0922	.1310	.1562	.1788	.2050
35	.0155	.0278	.0512	.0849	.1209	.1434	.1633	.1890
40	.0147	.0262	.0482	.0798	.1133	.1345	.1538	.1759
45	.0137	.0244	.0454	.0754	.1066	.1266	.1452	.1677
50	.0130	.0232	.0428	.0715	.1014	.1211	.1379	.1580
60	.0118	.0211	.0391	.0651	.0925	.1101	.1259	.1437
70	.0109	.0195	.0362	.0607	.0856	.1015	.1160	.1331
80	.0103	.0183	.0339	.0564	.0802	.0954	.1084	.1248
90	.0096	.0172	.0319	.0532	.0756	.0898	.1024	.1175
100	.0090	.0162	.0302	.0506	.0718	.0852	.0971	.1115
120	.0082	.0148	.0276	.0459	.0653	.0777	.0889	.1024
140	.0077	.0138	.0257	.0425	.0606	.0722	.0823	.0940
160	.0072	.0129	.0238	.0399	.0567	.0675	.0770	.0881
180	.0067	.0121	.0225	.0378	.0535	.0637	.0729	.0843
200	.0064	.0115	.0215	.0355	.0504	.0600	.0688	.0788
250	.0057	.0103	.0190	.0319	.0454	.0538	.0614	.0708
300	.0052	.0093	.0174	.0291	.0414	.0493	.0562	.0646
400	.0045	.0081	.0151	.0254	.0360	.0429	.0490	.0560
500	.0040	.0072	.0135	.0226	.0321	.0382	.0436	.0498

Table B.1: Critical values for $m = 2$

n	.90	.75	.50	.25	.10	.05	.025	.01
5	.0784	.1267	.1957	.2899	.3976	.4656	.5253	.5930
6	.0684	.1119	.1747	.2611	.3595	.4228	.4776	.5393
7	.0618	.1007	.1590	.2411	.3331	.3915	.4422	.5001
8	.0565	.0924	.1469	.2246	.3101	.3641	.4131	.4730
9	.0517	.0860	.1376	.2114	.2932	.3456	.3918	.4444
10	.0491	.0811	.1302	.1992	.2767	.3265	.3712	.4234
12	.0440	.0728	.1176	.1817	.2526	.2980	.3398	.3879
14	.0397	.0665	.1081	.1679	.2336	.2760	.3137	.3592
16	.0366	.0616	.1003	.1567	.2188	.2584	.2938	.3360
18	.0344	.0576	.0947	.1477	.2061	.2440	.2779	.3180
20	.0327	.0546	.0892	.1397	.1952	.2310	.2625	.3015
22	.0310	.0519	.0851	.1327	.1863	.2204	.2512	.2869
24	.0293	.0494	.0812	.1272	.1777	.2103	.2403	.2735
26	.0284	.0474	.0778	.1223	.1714	.2027	.2302	.2645
28	.0270	.0454	.0747	.1182	.1642	.1949	.2228	.2554
30	.0259	.0437	.0723	.1137	.1590	.1880	.2147	.2465
35	.0240	.0403	.0668	.1051	.1475	.1748	.1992	.2287
40	.0223	.0375	.0623	.0985	.1380	.1640	.1870	.2144
45	.0210	.0353	.0585	.0923	.1287	.1532	.1749	.2009
50	.0198	.0336	.0557	.0877	.1230	.1457	.1657	.1889
60	.0181	.0306	.0508	.0799	.1119	.1329	.1516	.1733
70	.0165	.0282	.0471	.0743	.1043	.1237	.1405	.1609
80	.0155	.0263	.0440	.0694	.0972	.1155	.1318	.1519
90	.0147	.0248	.0414	.0655	.0917	.1085	.1232	.1411
100	.0138	.0232	.0390	.0619	.0868	.1032	.1172	.1343
120	.0126	.0214	.0357	.0565	.0794	.0940	.1073	.1234
140	.0117	.0197	.0330	.0523	.0735	.0870	.0992	.1142
160	.0108	.0185	.0310	.0488	.0687	.0814	.0924	.1058
180	.0102	.0173	.0291	.0461	.0649	.0767	.0871	.0996
200	.0097	.0165	.0276	.0439	.0616	.0730	.0832	.0954
250	.0087	.0148	.0248	.0394	.0552	.0652	.0744	.0854
300	.0079	.0133	.0225	.0357	.0502	.0594	.0677	.0774
400	.0068	.0116	.0195	.0310	.0435	.0516	.0587	.0673
500	.0061	.0104	.0174	.0278	.0389	.0461	.0525	.0600

Table B.2: Critical values for $m = 3$

n	.90	.75	.50	.25	.10	.05	.025	.01
5	.1111	.1650	.2464	.3447	.4530	.5264	.5909	.6674
6	.0943	.1439	.2171	.3088	.4118	.4834	.5465	.6163
7	.0833	.1285	.1966	.2810	.3768	.4428	.5023	.5698
8	.0758	.1173	.1806	.2608	.3524	.4142	.4691	.5328
9	.0699	.1094	.1688	.2453	.3321	.3911	.4415	.5015
10	.0648	.1025	.1589	.2312	.3142	.3689	.4182	.4755
12	.0576	.0916	.1424	.2098	.2857	.3347	.3800	.4334
14	.0526	.0839	.1306	.1931	.2641	.3118	.3529	.4057
16	.0487	.0782	.1219	.1806	.2469	.2907	.3312	.3761
18	.0448	.0726	.1139	.1692	.2319	.2733	.3095	.3529
20	.0426	.0682	.1073	.1609	.2216	.2618	.2961	.3392
22	.0401	.0647	.1023	.1527	.2093	.2470	.2804	.3202
24	.0381	.0619	.0978	.1460	.2009	.2368	.2697	.3093
26	.0364	.0591	.0937	.1404	.1937	.2283	.2587	.2952
28	.0349	.0567	.0902	.1350	.1857	.2197	.2494	.2836
30	.0333	.0543	.0865	.1301	.1792	.2115	.2399	.2744
35	.0309	.0504	.0800	.1202	.1653	.1949	.2230	.2537
40	.0286	.0465	.0745	.1121	.1555	.1834	.2091	.2397
45	.0271	.0441	.0703	.1056	.1460	.1722	.1959	.2239
50	.0255	.0416	.0666	.1007	.1393	.1640	.1867	.2135
60	.0234	.0381	.0611	.0918	.1266	.1494	.1697	.1946
70	.0215	.0349	.0561	.0846	.1165	.1378	.1566	.1795
80	.0198	.0327	.0523	.0795	.1098	.1294	.1480	.1696
90	.0187	.0308	.0495	.0749	.1036	.1220	.1381	.1574
100	.0177	.0290	.0467	.0709	.0979	.1156	.1314	.1503
120	.0164	.0265	.0426	.0646	.0893	.1054	.1202	.1377
140	.0148	.0246	.0396	.0599	.0825	.0976	.1111	.1267
160	.0138	.0228	.0368	.0561	.0775	.0912	.1039	.1185
180	.0132	.0215	.0348	.0525	.0729	.0859	.0980	.1120
200	.0125	.0203	.0329	.0501	.0690	.0815	.0926	.1059
250	.0112	.0183	.0294	.0447	.0618	.0730	.0828	.0949
300	.0102	.0167	.0269	.0407	.0564	.0664	.0755	.0861
400	.0088	.0145	.0234	.0353	.0489	.0581	.0657	.0753
500	.0079	.0129	.0209	.0315	.0437	.0515	.0586	.0669

Table B.3: Critical values for $m = 4$

Bibliography

R. op den Akker (1998). What is probability theory about? Preprint, CTIT, University of Twente.

C.J. Albers (1998). Estimating bivariate distributions assuming some form of dependence. Master's thesis, University of Groningen.

C.J. Albers (2000). How to assign probabilities if you must — additions on article. Technical Report IWI-2000-5-04, University of Groningen. Also available from `http://www.math.rug.nl/~casper`.

C.J. Albers (2001). Presentatie van MWTL-data m.b.v. een speciale dichtheidsschatter. Technical Report IWI 2001-5-03, Rijksuniversiteit Groningen. Available from `http://www.math.rug.nl/~casper`.

C.J. Albers and B.P. Kooi and W. Schaafsma (2003). Trying to resolve the two-envelope problem. *To be published*. Available from `http://www.math.rug.nl/~casper`.

C.J. Albers and W. Schaafsma (2001a). Details on the standard error of a special density estimate. Technical Report IWI 2001-5-04, University of Groningen.

C.J. Albers and W. Schaafsma (2001b). How to assign probabilities if you must. *Statistica Neerlandica*, 55(3):346–357.

C.J. Albers and W. Schaafsma (2003a). Estimating a density by adapting an initial guess. *Computational Statistics & Data Analysis*, 42(1-2):27–36. Until published, available from `http://www.sciencedirect.com/science/journal/01679473`.

C.J. Albers and W. Schaafsma (2003b). A goodness-of-fit test smoother than smooth. *To be published*. Available from `http://www.math.rug.nl/~casper`.

W. Albers and P.C. Boon and W.C.M. Kallenberg (2000). Size and power of pretest procedures. *Annals of Statistics*, 28(1):195–214.

W. Albers and P.C. Boon and W.C.M. Kallenberg (2001). Power gain by pre-testing? *Statistics & Decisions*, 19(3):253–276.

J. Bernoulli (1713). *Ars Conjectandi, opus posthumum*. Basileae: Impensis Thurnisiorum.

D.J. Best and J.C.W. Rayner (1981). Are two classes enough for the χ^2 goodness of fit tests? *Statistica Neerlandica*, 35(3):157–164.

P. Billingsley (1995). *Probability and measure*. John Wiley & Sons, Inc., third edition.

P.C. Boon (1999). *Asymptotic behavior of pre-test procedures*. PhD thesis, Faculty of Mathematical Sciences, University of Twente.

G.W. Brier (1950). Verification of weather forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78:1–3.

J. Broome (1995). The two-envelope paradox. *Analysis*, 55(1):6–11.

M. Brown (1995). The distribution of total variation distance, with applications to simultaneous confidence intervals. *Computers & Operations Research and their Application to Problems of World Concern*, 22(4):373–381.

R. de Bruin (2003). *Size and shape of man and ape*. PhD thesis, University of Groningen.

R. de Bruin and D. Salomé and W. Schaafsma (1999). A semi-Bayesian method for nonparametric density estimation. *Computational Statistics & Data Analysis*, 30:19–30.

J. Chase (2002). The non-probabilistic two envelope paradox. *Analysis*, 62(2):157–160.

M. Clark and N. Shackel (2000). The two-envelope paradox. *Mind*, 109(435):415–442.

J.P. Cleave (1991). *Study of logics*. Oxford Science Publications.

G.S. Datta and R. Mukerjee and M. Ghosh and T.J. Sweeting (2000). Bayesian prediction with approximate frequentist validity. *Annals of statistics*, 28(5):1414–1426.

R.A. Davis and M.J. Chen and W.T.M. Dunsmuir (1995). Inference for MA(1) processes with a root on or near the unit circle. *Probability and Mathematical Statistics*, 15:227–242.

R.A. Davis and W.T.M. Dunsmuir (1996). Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. *Econometric Theory*, 12(1):1–29.

R.A. Davis and T. Mikosch (1998). Gaussian likelihood-based inference for non-invertible MA(1) processes with SαS noise. *Stochastic processes and their applications*, 77:99–122.

H.G. Dehling (1997). Daniel Bernoulli and the St. Petersburg paradox. *Nieuw Archief voor Wiskunde*, 15(3):223–227.

H.G. Dehling and J.N. Kalma and C. Moes and W. Schaafsma (1991). The Islamic mean: A peculiar L-statistic. *Studia Scientiarum Mathematicarum Hungarica*, 26:297–308.

L. Devroye (1987). *A course in density estimation*, volume 14 of *Progress in probability and statistics*. Birkhäuser Boston.

L. Devroye and L. Győrfi (1985). *Nonparametric density estimation: the L_1 view*. John Wiley & Sons, Inc.

P. Diaconis and S.L. Zabell (1986). Some alternatives to Bayes' rule. In B. Grofman and G. Owen (eds.), *Information pooling and group decision making. Proceedings of 2nd Irvine conference on political economy*. JAI Press, London.

L. Dümbgen and V.G. Spokoiny (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, 29(1):124–152.

N. Dunford and J.T. Schwartz (1957). *Linear Operators, Part I.* Interscience Inc.

M.L. Eaton (1982). A method for evaluating improper prior distributions. In S.S. Gupta and J.O. Berger (eds.), *Statistical decision theory and related topics III*, pages 329–352. Academic Press, New York.

M.L. Eaton (1992). A statistical diptych: admissible inferences — recurrence of symmetric Markov chains. *Annals of Statistics*, 20(3):1147–1179.

M.L. Eaton (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli*, 5(5):833–854.

E.S. Epstein (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987.

A.J. van Es (1988). *Aspects of nonparametric density estimation.* PhD thesis, University of Amsterdam.

T.S. Ferguson (1967). *Mathematical statistics, a decision-theoretic approach.* Academic Press, New York.

T.S. Ferguson (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

R.A. Fisher (1930). Inverse probability. In *Proceedings of the Cambridge Philosophical Society*, volume 26, pages 528–535.

R.A. Fisher (1973). *Statistical Methods and Scientific Inference.* Macmillan, New York, third edition. Reprinted in Fisher (1990).

R.A. Fisher (1990). *Statistical Methods, Experimental Design, and Scientific Inference.* Oxford University Press. Collection edited by J.H. Bennett and F. Yates.

D.A.S. Fraser (1959). *Nonparametric methods in statistics.* John Wiley & Sons, Inc., second edition.

M. Gardner (1982). *Aha! Gotcha: paradoxes to puzzle and delight.* W.H. Freeman and Company, New York.

A. Gellius (1946). *The Attic Nights of Aulus Gellius (Noctes Atticae).* Harvard University Press. Translated by J.C. Rolfe.

C. Genest and S. Weerahandi and J.V. Zidek (1990). Conditionalization and likelihood dominance in group belief formation. *Statistical Science*, 8(2):183–198.

C. Genest and J.V. Zidek (1986). Combining probability distributions: a critique and an annotated bibliography. *Statistical Science*, 1(1):114–148.

S. Ghosal (1997). A review of consistency and convergence of posterior distribution. In *National seminar on Bayesian statistics and its applications.* Banaras Hindu University, Varanashi, India.

S. Ghosal (2001). Convergence rates for density estimation with bernstein polynomials. *Annals of Statistics*, 29(5):1264–1280.

S. Ghosal and J.K. Ghosh and R.V. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143 – 158.

S. Ghosal and J.K. Ghosh and A.W. van der Vaart (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.

W.G. Gilchrist (2000). *Statistical modelling with quantile functions*. Chapman & Hall, Inc.

A. Golan and G. Judge and D. Miller (1996). *Maximum entropy econometrics: robust estimation with limited data*. John Wiley & Sons, Inc.

I.J. Good (1952). Rational decisions. *Journal of the Royal Statistical Society*, 14:107–114.

P.R. Halmos (1995). To count or to think, that is the question. *Nieuw Archief voor Wiskunde*, 13(1):61–76.

J. van Heijenoort (1967). Logical paradoxes. In P. Edwards (ed.), *The Encyclopedia of Philosophy*, volume 5, pages 45–51. The Macmillan Company & The Free Press.

F.R. Helmert (1876). Die Berechnung des wahrscheinlichen Beobachtungsfehlers aus den ersten Potenzen der Differenzen gleichgenauer directer Beobachtungen. *Astronomische Nachrichten*, 88:113–132.

J. Hemelrijk (1978). Rules for building statistical models. *Statistica Neerlandica*, 32(3):123–134.

W. Hoeffding (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325.

T. Inglot and Y. Jurlewitz and T. Ledwina (1990). On Neyman-type smooth tests of fit. *Statistics*, 21:549–568.

T. Inglot and W.C.M. Kallenberg and T. Ledwina (1994). Power approximations to and power comparison of smooth goodness-of-fit tests. *Scandinavian Journal of Statistics*, 21(2):131–145.

F. Jackson and P. Menzies and G. Oppy (1994). The two envelope 'paradox'. *Analysis*, 54(1):43–45.

E.T. Jaynes (1996). Probability theory: the logic of science. To be published by Cambridge. Available at `http://omega.math.albany.edu:8008/JaynesBook.html`.

H. Jeffreys (1932). On the theory of errors and least squares. *Proceedings of the Royal Society*, 138:48–55.

H. Jeffreys (1939). *Theory of probability*. Oxford University Press.

M. Kac and G. Rota and J.T. Schwartz (1986). *Discrete thoughts — Essays on Mathematics, Science, and Philosophy*. Birkhäuser, Boston.

W.C.M. Kallenberg and J. Oosterhoff and B.F. Schriever (1985). The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association*, 80(392):959–968.

O.J.W.F. Kardaun and W. Schaafsma (2003). *Distributional Inference, towards a Bayes-Fisher-Neyman compromise*. (Available on request).

M.G. Kendall and A. Stuart (1958). *The advanced theory of statistics*, volume I: Distribution theory. Griffin & Company, London, first edition.

M.G. Kendall and A. Stuart (1973). *The advanced theory of statistics*, volume II: Classical inference and relationship. Griffin & Company, London, third edition.

A. Kolmogorov (1933). Sulla determinazione emprica di una legge di distribuzione. *Giornalle dell Instituto Italiano degli Attuari*, 4:1–11.

B.P. Kooi (1999). The Monty Hall dilemma. Master's thesis, University of Groningen.

M. Kraitchik (1943). *Mathematical recreations*. George Allen & Unwin ltd, London.

A.H. Kroese (1994). *Distributional inference: a loss function approach*. PhD thesis, University of Groningen.

A.H. Kroese and E.A. van der Meulen and K. Poortema and W. Schaafsma (1995). Distributional inference. *Statistica Neerlandica*, 49(1):63–82.

A.H. Kroese and D. Salomé and W. Schaafsma (1999). An optimum property for fiducial inference. *Statistics & decisions*, 17:285–291.

E.L. Lehmann (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, 37:1137–1153.

E.L. Lehmann (1986). *Testing statistical hypotheses*. Springer-Verlag, New York, Inc., second edition.

E. Linzer (1994). The two envelope paradox. *The American Mathematical Monthly*, 101(5):417–419.

M. Loève (1955). *Probability theory*. Princeton University Press.

Z.A. Lomnicki (1952). The standard error of Gini's mean difference. *Annals of Mathematical Statistics*, 23:635–637.

H.B. Mann and A. Wald (1942). On the choice of the number of class intervals in the application of the chi square test. *Annals of Mathematical Statistics*, 13:306–317.

A.W. Marshall (1996). Copulas, marginals, and joint distribution functions. In L. Ruschendorf and B. Schweizer and M.D. Taylor (eds.), *Distributions with fixed marginals and related topics*, volume 28, pages 213–222. IMS Lecture Notes, Monograph Series.

T.J. McGrew and D. Shier and H.S. Silverstein (1997). The two-envelope paradox resolved. *Analysis*, 57(1):28–33.

F. Mosteller (1965). *Fifty challenging problems in probability with solutions*. Addison–Wesley, Reading, Massachusetts.

J. Muñoz Perez and A. Fernández Palacín (1987). Estimating the quantile function by Bernstein polynomials. *Computational Statistics & Data Analysis*, 5:391–397.

U.S. Nair (1936). The standard error of Gini's mean difference. *Biometrika*, 28:428–436.

B. Nalebuff (1988). Cider in your ear, continuing dilemma, the last shall be the first, and more. *Journal of Economic Perspectives*, 2(2):149–156.

B. Nalebuff (1989). The other person's envelope is always greener. *Journal of Economic Perspectives*, 3(1):171–181.

J. von Neumann and O. Morgenstern (1944). *Theory of games and economic behaviour*. Princeton University Press.

J. Neyman (1937). 'Smooth' test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20:149–199.

H. Nikaidô (1953). On a minimax theorem and its applications to functional analysis. *Journal of the Mathematical Society of Japan*, 5:86–94.

H. Nikaidô (1959). On a method of proof for the minimax theorem. *Proceedings of the American Mathematical Association*, 10:205–212.

D.L. Otis and K.P. Burnham and G.C. White and D.R. Anderson (1978). *Statistical inference from capture data on closed animal populations*, volume 62. Wildlife Monographs.

E. Parzen (1960). *Modern probability theory and its applications*. John Wiley & Sons, Inc.

K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175.

K. Pearson (1920). The fundamental problem of practical statistics. *Biometrika*, 13:1–16.

S. Petrone (1999). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 26:373–393.

G. Polya (1975). Probabilities in proofreading. *American Mathematical Monthly*, 83(1):45.

W. Poundstone (1992). *Prisoner's dilemma*. Doubleday, New York.

A. Rapoport (ed.) (1974). *Game theory as a theory of conflict reasoning*. Reidel, Dordrecht, Holland.

A. Rapoport and A.M. Chammah and C.J. Orwant (1965). *Prisoner's dilemma: a study in conflict and cooperation*. Ann Arbor, Michigan.

J.C.W. Rayner and D.J. Best (1989). *Smooth tests of goodness of fit*. Oxford University Press.

S. Russel and P. Norvig (1995). *Artificial intelligence; a modern approach*. Prentice Hall.

D. Salomé (1998). *Statistical inference via fiducial methods*. PhD thesis, University of Groningen.

D. Salomé and R. de Bruin and W. Schaafsma (1999). Q-values for χ^2 problems. *Statistics & Decisions*, 17.

L.J. Savage (1951). The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67.

W. Schaafsma (1969). Minimax risk and unbiasedness for multiple decision problems of type I. *Annals of Statistics*, 5:1684–1720.

W. Schaafsma (1971). The Neyman-Pearson theory for testing statistical hypotheses. *Statistica Neerlandica*, 25(1):1–27.

W. Schaafsma (1997). Laudatio Sir David Cox. *Nieuw Archief voor Wiskunde*, 15(3):229–231.

W. Schaafsma and A.G.M. Steerneman (1981). Discriminant analysis when the number of features is unbound. *IEEE transactions on systems, man, and cybernetics*, SMC-11(2):144–151.

E.F. Schuster (1969). Estimation of a probability density and its derivatives. *Annals of Mathematical Statistics*, 40:1187–1196.

B. Schweizer and A. Sklar (1983). *Probabilistic metric spaces*. North Holland, New York.

S. Selvin (1975). Letter to the editor: a problem in probability. *The American Statistician*, 29(1):67.

B.W. Silverman (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Annals of Statistics*, 6(1):177–184.

B.W. Silverman (1986). *Denstiy estimation for statistics and data analysis*. Chapman and Hall, London.

A. Sklar (1996). Random variables, distribution functions and copulas – a personal look backward and forward. In L. Ruschendorf and B. Schweizer and M.D. Taylor (eds.), *Distributions with fixed marginals and related topics*, volume 28, pages 1–14. IMS Lecture Notes, Monograph Series.

N.V. Smirnov (1939). Sur les écarts de la courbe de distribution empirique. *Recueil Mathématique N.S.*, 6:3–26.

R. Smullyan (1997). *The riddle of Scheherazade, and other amazing puzzles, ancient and modern*. Knopf, New York.

A.J. Stam (1987). Statistical problems in ancient numismatics. *Statistica Neerlandica*, 41(3):151–173.

I. Stewart (2000). Mathematical recreations: paradox lost. *Scientific American*, 6:88–89.

M. Swaving and L. de Vries (2000). Omgaan met waarden onder de detectiegrens. Technical Report E1680-01, CQM BV, Eindhoven.

K. Tanaka (1990). Testing for a moving average unit root. *Econometric Theory*, 6:433–444.

A. Tenbusch (1994). Two-dimensional Bernstein polynomial estimators. *Metrika*, 41:233–253.

W.R. Thompson (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *Annals of Mathematical Statistics*, 7:122–128.

R.A. Vitale (1975). A Bernstein polynomial approach to density estimation. In M.L. Puri (ed.), *Statistical Inference and Related Topics.*, pages 87–100. Academic Press, New York.

A. Wald (1947). Book review of Neumann and Morgenstern, 1944. *Reprinted from The Review of Economic Statistics*, 24(1):47–52.

A. Wald (1964). *Statistical decision functions*. John Wiley & Sons, Inc.

M.P. Wand and L. Devroye (1993). How easy is a given density to estimate? *Computational Statistics & Data Analysis*, 16(3):311–323.

G.C. White and D.R. Anderson and K.P. Burnham and D.L. Otis (1982). Capture-recapture and removal methods for sampling closed populations. Technical Report LA-8787-NERP, Los Alamos National Laboratory.

S.S. Wilks (1962). *Mathematical statistics*. John Wiley & Sons, Inc.

W.L. Winston (1994). *Operations research — applications and algorithms*. Wadsworth, Belmont, California, third edition.

M.C.K. Yang and D.D. Wackerly and A. Rosalsky (1982). Optimal stopping rules in proofreading. *Journal of applied probability*, 19(3):723–729.

S.L. Zabell (1988a). Loss and gain: the exchange paradox. In J. M. Bernardo and M. H. DeGroot and D. V. Lindley and A. F. M. Smith (eds.), *Bayesian statistics 3. Proceedings of the third Valencia international meeting*, pages 233–236. Clarendon Press, Oxford.

S.L. Zabell (1988b). Symmetry and its discontents. In B. Skyrms and W. L. Harper (eds.), *Causation, chance and crecedence*. Kluwer Academic Publishers.

A. Zellner and R.A. Highfield (1988). Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *Journal of Econometrics*, 37:195–209.

Author index

Akker, R. op den, 27
Albers, C.J., 3, 29, 47, 72, 75, 98, 100
Albers, W., 76
Anderson, D.R., 118, 126

Bernoulli, J., 103
Best, D.J., 66, 78, 88, 89
Billingsley, P., 48, 114
Boon, P.C., 76
Brier, G.W., 15
Broome, J., 33, 34, 35
Brown, M., 71
Bruin, R. de, 47, 49, 60, 62, 72, 73, 77, 90, 92, 116
Burnham, K.P., 118, 126

Chammah, A.M., 27
Chase, J., 31, 32
Chen, M.J., 103, 104, 113, 114
Clark, M., 33
Cleave, J.P., 5

Datta, G.S., 107
Davis, R.A., 103, 104, 113, 114
Dehling, H.G., 35, 52
Devroye, L., 49, 79
Diaconis, P., 27
Dümbgen, L., 76
Dunford, N., 49, 79
Dunsmuir, W.T.M., 103, 104, 113, 114

Eaton, M.L., 7
Epstein, E.S., 15
Es, A.J. van, 50

Ferguson, T.S., 38, 40, 47, 48, 49
Fernández Palacín, A., 52
Fisher, R.A., iii, 51, 108, 139, 147, 149
Fraser, D.A.S., 129

Gardner, M., 30
Gellius, A., 32
Genest, C., 111
Ghosal, S., 49, 50, 61
Ghosh, J.K., 49
Ghosh, M., 107
Gilchrist, W.G., 50
Golan, A., 36, 69, 70
Good, I.J., 7
Győrfi, L., 49

Halmos, P.R., 115
Heijenoort, J. van, 31
Helmert, F.R., 112
Hemelrijk, J., 3, 4, 22
Highfield, R.A., 69, 70
Hoeffding, W., 63, 129

Inglot, T., 89, 92

Jackson, F., 33
Jaynes, E.T., 36, 69, 70, 91, 108
Jeffreys, H., 107
Judge, G., 36, 69, 70
Jurlewitz, Y., 89

Kac, M., iv
Kallenberg, W.C.M., 76, 78, 89, 92
Kalma, J.N., 52
Kardaun, O.J.W.F., 120
Kendall, M.G., 78, 129, 130
Kolmogorov, A., 76, 92, 101
Kooi, B.P., 26, 29
Kraitchik, M., 29
Kroese, A.H., iii, 109, 120

Ledwina, T., 89, 92
Lehmann, E.L., 72, 78, 106, 110
Linzer, E., 33
Loève, M., 92, 122
Lomnicki, Z.A., 129

Mann, H.B., 78
Marshall, A.W., 72, 73
McGrew, T.J., 33
Menzies, P., 33
Meulen, E.A. van der, iii
Mikosch, T., 103, 104, 114
Miller, D., 36, 69, 70
Moes, C., 52
Morgenstern, O., 11, 39, 144
Mosteller, F., 26
Muñoz Perez, J., 52
Mukerjee, R., 107

Nair, U.S., 129
Nalebuff, B., 30, 33, 35
Neumann, J. von, 11, 39, 144
Neyman, J., 66, 75, 76, 81, 88, 89
Nikaidô, H., 41
Norvig, P., 27

Oosterhoff, J., 78, 89
Oppy, G., 33
Orwant, C.J., 27
Otis, D.L., 118, 126

Parzen, E., 23, 122
Pearson, K., 43, 75, 76, 91, 111
Petrone, S., 61
Polya, G., 117
Poortema, K., iii
Poundstone, W., 27

Ramamoorthi, R.V., 49
Rapoport, A., 27
Rayner, J.C.W., 66, 78, 88, 89
Rosalsky, A., 118
Rota, G., iv
Russel, S., 27

Salomé, D., 47, 49, 51, 60, 62, 72, 73, 77, 90, 92, 108, 109, 120
Savage, L.J., 38
Schaafsma, W., iii, 3, 5, 29, 41, 47, 49, 52, 60, 62, 72, 73, 75, 77, 90, 91, 92, 103, 109, 120
Schriever, B.F., 78, 89
Schuster, E.F., 60
Schwartz, J.T., iv, 49, 79
Schweizer, B., 72
Selvin, S., 26
Shackel, N., 33
Shier, D., 33
Silverman, B.W., 47, 50, 60
Silverstein, H.S., 33
Sklar, A., 72
Smirnov, N.V., 76
Smullyan, R., 31
Spokoiny, V.G., 76
Stam, A.J., 126
Steerneman, A.G.M., 90, 91
Stewart, I., 33
Stuart, A., 78, 129, 130
Swaving, M., 96, 97, 99, 102
Sweeting, T.J., 107

Tanaka, K., 114
Tenbusch, A., 73
Thompson, W.R., 51

Vaart, A.W. van der, 49
Vitale, R.A., 50, 52, 61, 73
Vries, L. de, 96, 97, 99, 102

Wackerly, D.D., 118
Wald, A., 11, 38, 39, 78
Wand, M.P., 49
Weerahandi, S., 111
White, G.C., 118, 126
Wilks, S.S., 113, 130
Winston, W.L., 27

Yang, M.C.K., 118

Zabell, S.L., 27, 29, 30, 33, 70
Zellner, A., 69, 70
Zidek, J.V., 111

Summary

Distributional inference: the limits of reason

Science advances by combining rational arguments and empirical information. In fields like philosophy and pure mathematics, emphasis is laid on the rational arguments, whilst in the applied sciences the collection and interpretation of data are the field of interest. In mathematical statistics one tries to combine these aspects. The primary goal is to make statistical inferences about something unknown. Such inferences can be of help in further discussion, e.g. in selecting a decision. The methods should not depend on 'the intentions that might be furthered by utilizing the knowledge inferred'[1]. When the available data are too limited, then different procedures may yield different inferences. The statistician should refrain from providing a specific inference in case the differences are 'too large'. When such an inference can be given, this inference should be accompanied by a statement about the uncertainty of the inference. This could be done by providing a distributional inference, or by providing the results of different approaches.

An example is as follows. Ornithologist G.Th. de Roos is observing a population of Ruddy Turnstones (*Arenaria Interpres*) on the Frisian island Vlieland. Some of these birds are ringed, however the ring-number is not always observable, e.g. because another bird is blocking the view. After how many days of observing is it safe to assume that all ringed birds in the population have been observed at least once? This question can be answered by constructing a distributional inference about the number of present, yet unseen, ringed birds, including a probability statement about the hypothesis that all ringed birds have been seen. Of course, the results depend in some way on the probabilistic assumptions one makes, and on the statistical principles one follows.

The first part of this thesis consists of 'finger exercises' illustrating that information about the unknown can only be of value if the mechanism generating the information is (sufficiently well) known. In probability theory, information is incorporated by conditioning to it. This generates difficulties in statistical practice, because unknown aspects are involved in the joint distribution of the random variables X and Y that are behind the observations x and the unknown y. Firstly, this is extensively exemplified by a die-rolling game. From the information 'the number of eyes is even' one cannot conclude automatically: 'the probability that a six has been thrown, equals one third'. The way in which the source of information operates, should be incorporated in

[1] R.A. Fisher, *Statistical Methods and Scientific Inference*, third edition, Macmillan, New York, 1973, p. 107

the statistical model. Secondly, a similar example, the two-envelopes problem, is considered. Again, the difficulties involving the numerical specification of conditional probabilities are in the forefront.

The second and most important part deals with the situation where one has a random sample $x_1, \ldots, x_n$ from a distribution with density f. The goal is to use the sample to form an estimate of f or, almost equivalently, to generate a distributional inference about $y(= x_{n+1})$. A new method is discussed to estimate the density f, where 'initial knowledge' of f is incorporated in the model. This is done by specifying a probability density ψ as the 'initial guess' for f. Also the degree of confidence in this ψ is quantified and incorporated in the method. By means of a multi-modal approach, incorporating aspects from both Classical and Bayesian statistics, and on basis of the sample x, 'initial guess' ψ (and the degree of confidence in ψ), an estimate $\hat{f}$ of f is generated. When the initial guess ψ is not unreasonable, this density estimate performs better, in general, than the generally used kernel methods. This is no surprise, since the kernel method makes no use of ψ. It is at this point unclear how the comparison will turn out when ψ is incorporated in the kernel method.

To study the applicability of the developed method, an extensive data set about the pollution of Dutch waters is considered. Previous investigations showed that the different concentrations of pollutants can reasonably well be described by lognormal distributions. A complication is that the concentrations can only be measured when they are above a certain detection threshold. The density estimation theory of this thesis, adapted to mentioned complication, is used to 'fine-tune' the 'initial guess' of lognormality to the data. The resulting density estimates are better than the density estimates obtained previously by fitting lognormal densities.

The density estimation theory of this thesis can usefully be applied to the goodness of fit context where a statement is required about the truth or falsity of the hypothesis H_0: $f = \psi$. The resulting goodness of fit tests have interesting relations with the well-known χ^2-test, Kolmogorovs test, and Neymans 'smooth tests'.

To emphasize the usefulness of distributional inference, an example from the interface of multivariate analysis and time-series analysis is discussed.

Samenvatting

Het doen van verdelingsuitspraken om de grenzen van het weten te preciseren

De vooruitgang in de wetenschap vindt grotendeels plaats door rationele argumenten te combineren met empirische gegevens. In de filosofie en de zuivere wiskunde spelen rationele argumenten de hoofdrol, terwijl men zich in de toegepaste wetenschap bezighoudt met het verzamelen en interpreteren van data. In de wiskundige statistiek tracht men deze aspecten te combineren. Primair gaat het er dan om statistische uitspraken te doen omtrent iets onbekends. Die uitspraken kunnen dan dienen als basis bij verdere beschouwingen, bijvoorbeeld gericht op het nemen van een beslissing. De wiskundig-statisticus tracht dus methoden voor het doen van zulke uitspraken te leveren. Deze methoden moeten niet afhangen van 'the intentions that might be furthered by utilizing the knowledge inferred' [2]. Indien de beschikbare data te beperkt zijn, dan kunnen verschillende uitwerkingen tot verschillende uitspraken leiden. De statisticus zal zich van een specifiek antwoord onthouden als de verschillen 'te groot' zijn. Wanneer zo'n uitspraak wel gegeven wordt, dient hierbij de onzekerheid betrokken te worden, bijvoorbeeld door de uitspraak in de vorm van een kansverdeling te gieten, of door resultaten van diverse uitwerkingen te vermelden.

Een typerend voorbeeld is als volgt. De ornitholoog G.Th. de Roos observeert een populatie Steenlopers (*Arenaria interpres*) op Vlieland. Een deel van deze vogels is geringd, maar dit ring-nummer is niet altijd te lezen, bijvoorbeeld omdat een andere vogel het zicht belemmert. Na hoeveel dagen van observeren mag De Roos er van uit gaan dat alle aanwezige geringde vogels minstens eenmaal geobserveerd zijn? Dit kan beantwoord worden door een verdelingsuitspraak te maken over het aantal nog niet geziene maar wel aanwezige geringde vogels, inclusief een kansuitspraak over de hypothese dat alle geringde vogels gezien zijn. Vanzelfsprekend hangen de resultaten enigszins af van de kanstheoretische veronderstellingen die men maakt en de statistische principes die men hanteert.

Het eerste deel van het proefschrift bestaat uit 'vingeroefeningen' die illustreren dat informatie over het onbekende alleen van waarde kan zijn wanneer bekend is volgens welk mechanisme deze informatie tot stand is gekomen. In de theoretische kansrekening wordt met informatie omgegaan door ernaar te conditioneren. In de statistische praktijk levert dit problemen doordat er vaak onduidelijkheid heerst over de gezamenlijke verdeling van de toevallige grootheden X en Y die zitten achter de waarnemingen

[2] R.A. Fisher, *Statistical Methods and Scientific Inference*, third edition, Macmillan, New York, 1973, p. 107

x en de onbekende y. Eerst wordt dit uitvoerig uitgewerkt voor een voorbeeld met een worp met een zuivere dobbelsteen. Zo mag uit de informatie 'er is een even aantal ogen gegooid' niet automatisch de conclusie getrokken worden: 'de kans dat er een zes gegooid is, is één derde'. De wijze waarop de informatiebron werkt, moet meegenomen worden in het statistische model. Vervolgens wordt een gelijksoortig vraagstuk, het twee enveloppen probleem, bekeken. Net als in het eerste voorbeeld speelt de moeilijkheid omtrent de numerieke specificatie van conditionele kansen een grote rol.

Het tweede en belangrijkste deel gaat over de situatie waar men een steekproef $x_1, \ldots, x_n$ heeft uit een verdeling met een kansdichtheid f. Men wil deze steekproef gebruiken om een schatting van f te vormen of, wat vrijwel hetzelfde is, om een verdelingsuitspraak ('distributional inference') te doen over $y = (x_{n+1})$. Een nieuwe methode wordt beschreven om de kansdichtheid f te schatten, waarbij 'voorkennis' met betrekking tot f in de beschouwing betrokken wordt. Dit gebeurt door een kansdichtheid ψ als 'initial guess' voor f te specificeren. (Tevens dient de mate van vertrouwen in deze voorkennis te worden vastgelegd.) Op basis van de steekproef x, de beginschatting ψ (en de mate van vertrouwen er in) wordt door middel van een multi-modale aanpak, die het midden zoekt tussen de klassieke en de Bayesiaanse statistiek, een schatting $\hat{f}$ van f gemaakt. Wanneer ψ geen onredelijke beginschatting is, dan presteert de bijbehorende dichtheidsschatter, in het algemeen, beter dan de gebruikelijke kernschatters. Dat is geen wonder want de kernschatter maakt geen gebruik van ψ. Hoe de vergelijking uitpakt als de kernschatter wordt aangepast aan ψ is nog onbekend.

Om de toepasbaarheid van deze methode te beschouwen, is gekeken naar omvangrijke data over de mate van verontreiniging van oppervlaktewater in Nederland. Eerder onderzoek van deze data heeft uitgewezen dat de concentraties van vervuilende stoffen doorgaans redelijk goed te beschrijven zijn met lognormale verdelingen. Een complicatie is dat de concentratie niet gemeten kan worden wanneer deze onder een bepaald waarnemingsniveau ligt. De dichtheidsschattingstheorie uit dit proefschrift, aangepast aan genoemde complicatie, is gebruikt om de 'voorkennis' van bij benadering lognormale verdelingen aan te passen aan de metingen. Dit levert betere dichtheidsschattingen dan wanneer lognormale verdelingen aan de data worden aangepast.

Deze dichtheidsschattingsmethode blijkt ook zeer geschikt te zijn voor de zogenaamde goodness of fit problematiek. Er wordt aan de hand van de data een uitspraak gedaan over de hypothese H_0: $f = \psi$. De resulterende toetsingsmethoden hebben interessante raakvlakken met bestaande goodness of fit toestingsmethoden, zoals de χ^2-toets, Kolmogorovs goodness of fit methode en de 'smooth tests' van Neyman.

Om te beklemtonen dat het doen van verdelingsuitspraken van belang is in de praktijk, is een extra voorbeeld toegevoegd uit het grensgebied tussen de multivariate analyse en de analyse van tijdreeksen.