

Statistical Inference: 600 Exercises, 100 Stories, and Solutions

Nils Lid Hjort and Emil Aas Stoltenberg

Department of Mathematics, University of Oslo

Here: Preface, Chapters 7, 8, 9, 10

– This version last touched by Nils and Emil, 16 August 2026 –

©Nils Lid Hjort and Emil Aas Stoltenberg, 2026

Some technical stuff

ISBN - Numbers numbers

The Kioskvelter Project

This is a draft of our book-to-be and it may not be reproduced
or transmitted, in any form or by any means, without permission.

1234-5678

To my somebody
– N.L.H.

To my somebody
– E.Aa.S.

Preface

There are various books on various shelves concerning the broader themes of *statistical inference*, both generally and with its many specialised themes. These include Bayesian analysis (parametric and nonparametric), testing theory, optimal estimation, confidence distributions, large-sample theory with the necessary probability on board, survival analysis and event history analysis, Markov chains and processes, optimal decision theory, time series, machine learning, processing of big data, image analysis, multivariate analysis, model selection and averaging, goodness of fit methods, stochastic processes, change-point identification, prediction, etc. This book is decidedly different, in several key respects.

600 Exercises

First of all, we change gears from the traditional narrative of ‘theory development first, then a few exercises and illustrations’, to ‘crisp introduction text to a theme, then jump into exercises mode’. So when we come to Bayesian analysis in Ch. 7, for example, readers & doers are pushed rather quickly into working with and understanding from our generous lists of exercises. Thus theory is developed as we go, through problems and their solutions, from simpler statements and illustrations to the needed lemmas and theorems, as opposed to the more traditional ‘theory first’ approach. This is in line with what Aristotle wrote, in *Nicomachean Ethics*; “for the things we have to learn before we can do them, we learn by doing them”; people learn to cook by cooking. Somewhat more pointedly, the psychologist and progressive educator John Dewey held, “learn to know by doing, and to do by knowing”. Our approach is hence correlated with educational principles around Experimental Learning, and also the so-called Flipped Classroom.

The introductory Ch. 1 works through a generous list of parametric models, their construction and properties, from the more famous to various extensions and connections to yet other models. A healthy segment of our readers would know the basics of these models already, but may use the chapter for reference, as do we for the rest of the book. Ch. 2 then builds the apparatus for large-sample approximations, the uses of which truly abound in probability theory and both theoretical and applied statistics, with the Law of Large Numbers, the Central Limit Theorem with extensions, the delta method, moment-generating functions, etc. Chs. 3 and 4 then develop the fundamental concepts and methods for statistical estimation, confidence, and testing, with the required Neyman–Pearson theory and other excursions.

The important Ch. 5, on likelihoods and their ramifications, can of course be worked through and appreciated separately, but is also used in various guises in several later

exercises,
exercises

335 b.C.

in 1914

chapters, specifically for the Bayesian methods of Ch. 7, the survival and event history Ch. 11, for model selection in Ch. 12, and for Markov chain models in Ch. 13. Ch. 6 goes into minimum divergence inference, generalising and in various situations robustifying likelihood inference. In Ch. 8 we build confidence, with the frequentist no-prior methodology for summarising inference in terms of distributions for focus parameters, with consequences also for meta-analyses, ways of combining information across diverse sources. Themes related to loss, risk, optimal estimation, minimaxity, admissibility are worked with in Ch. 9.

Often enough statisticians need for modelling and inference purposes to go beyond the finite-dimensional world of data points and vectors. Ch. 10 develops the appropriate tools and results for classes of stochastic processes, with Donsker theorems, Brownian motion and bridges, along with statistically relevant constructions of running t-tests, monitoring processes for models, log-likelihood maxima bridges, and more. Some of these are also used in the survival chapter Ch. 11.

Model selection themes have been part of statistical practice since the early models were invented, but have become more important and ubiquitous in the modern world of computing, access to an increasing volume of datasets, and available software to run through thousands of model candidates. Ch. 12 develops these themes, with emphasis on the established acronyms AIC, BIC, FIC (the Akaike, the Bayes, the Focused Information Criterion). Markov chains and processes are the themes of Ch. 13, where also the important Markov Chain Monte Carlo methods are tended to; the MCMC algorithms play an important role both in the finite-dimensional Bayes world of Ch. 7 but also in our final Ch. 14, on Bayesian nonparametrics, with infinite-dimensional models.

Notably, each chapter is rounded off with a ‘Notes and Pointers’ section, commenting on how key ideas have evolved historically, and with connections to other chapters and yet further themes.

Notes &
Pointers

100 Statistical Stories

Secondly, we, with the active participation of you, the readers & doers, illustrate and showcase concepts, modelling, theory development via a generous list of Statistical Stories. These range from the perhaps simple illustrations and applications to the rather more challenging terrain, where theory needs further development or new twists, in order to solve real-life real-data problems. Our stories are sorted into the broader categories of

stories, stories

- (i) Demography, Epidemiology, Medicine;
- (ii) Biology, Climate, Ecology;
- (iii) Art, History, Literature, Music;
- (iv) Economics, Political Science, Sociology;
- (v) Sports;
- (vi) Simulated stories;
- (vii) Miscellaneous.

Significantly, also the stories are told in the spirit of exercises, from start observations concerning the context of and the problems at hand, to developing the right modelling and then the tools required for solving these. Students and readers are then meant to get involved in the problem solving process from the start, as opposed to listening to a polished final lecture by someone who with struggle and efforts might finally having solved the problem. Ploughing through the pieces of these stories will also provide good training in the subtle tasks of statistical modelling, checking, and selection.

It is to be noted that most stories engage tools from not merely one chapter at a time, so to speak, but often via several themes from several chapters. So as opposed to illustrating a certain Bayesian method, for example, there might be alternative and equally relevant approaches, depending on the context, drawn from other chapters in our book. Also, for some stories, theory needs further development as we go.

solutions,
solutions

Importantly, we also provide *solutions* to all 600 Exercises and 100 Statistical Stories. There is not space in the physical book to include these, however, so they are placed online in the GitHub site pointed to below. That site also has *all datasets* used through the book, with **R** and **Python** code when required.

Our methodological horizon is wide, so also in that sense our book is on the ambitious side of the spectrum. We have chapters on a fair range of methodological themes, from basic models to estimation and testing, on large-sample theory with central limit theorems and so on, empirical processes, parametric and nonparametric Bayes, confidence distributions, survival analysis, Markov chains and processes, and modern methods for statistical learning. Though we do manage to develop several of these themes rather broadly, to the state of strong applicability also for challenging uses, the book is not intended to be a full-semester full-course book for each of its subthemes. Several types of courses can however easily be built from the book, as we offer advice on below.

Appendixes A, B, C, D, E

At the end of the book are four appendixes with exercises that takes the reader through some of the required mathematical and probability background. Our list of exercises are selective, and we have strived to only include material that is strictly necessary for solving and understanding the exercises in other chapters, and that hopefully also sheds some light on the short & crisp exercises as well as on the statistical stories.

Appendix **A** gives a concise and selective introduction to metric spaces. and includes results about such spaces that are used several places in the book, but in particular in Ch. **10**. Statistics builds on probability, and probability is a branch of measure theory. Appendix **B** gives an introduction to measure theory. This includes exercises on σ -algebras, Carathéodory's extension theorem, Dynkin's lemma, the Lebesgue integral, the convergence theorems, the Radon–Nikodym theorem, and so on. Crucially, the theory developed in this appendix underlies all of the statistical theory in the book, even though for most of the exercises and stories this theory is not actively present, so to speak. The measure theory appendix provides the background for Appendix **C** on probability theory. This appendix gives a measure theoretic introduction to probability theory, and develops such familiar concepts as the expectation, independence, and conditional expectation, to mention some, in a rigorous fashion. Appendix **D** contains exercises on martingale theory in discrete and continuous time. Some martingale theory is present in Ch. **10**, and is used very actively in Ch. **11** on survival analysis. Martingale theory is a subject that deserves book length treatises (of which there are many, see Notes & Pointers), in our martingale appendix, however, we have attempted to keep the theory to the bare minimum required to work with survival analysis from the counting process point of view, without having to take too many mathematical results for granted.

Finally, and crucially, Appendix **E** goes through the datasets used for our Statistical Stories, with references and brief backgrounds and descriptions, along with pointers to where in our book the datasets are used.

This book, its students and readers, and its GitHub site

As made clear above the book is broad-horizoned, and will have different strata of readers. The required start level for this book to be readable and useful is meant to be at end-of-bachelor and above, as in Master's and PhD levels (and above!). Prerequisites, for the core of readers, should be basic mathematics, including calculus, linear algebra with matrices and so on; familiarity and experience with programming in or via suitable statistical or mathematical software systems, such as **R** and **Python**, and specifically with the ability to code up things on one's own, outside the terrain of the application of already fine-tuned algorithms of packages; and having operating knowledge of classic statistical models and their parameters. Our start chapter 1 goes through such models, though, along with combinations, variations, extensions, so it is perfectly possible to start almost from scratch, with a certain uphill parameter. Otherwise the already flagged intention and spirit is 'come along', learning away as one proceeds, chapter by chapter, story by story.

Our intended strata of readers are roughly as follows. These boxes are of course far from accurately defined, and our students are invited to cross over to other categories as they might be inclined to. who are you?

(i) The *Linear Readers*, who will benefit from having the stamina to work through chapter by chapter (ideally also exercise for exercise) and appropriate subsets of our stories. These readers will be at a Master's or PhD level. We even hope and trust that experienced scholars above these levels will find things to like, to be surprised and inspired by; some of the material, also in the collection of Stories, is advanced.

(ii) The *Chapter Samplers*, who need to freshen up or learn more about certain themes, like confidence distributions or survival analysis, without necessarily having the time to or the interest for all the other chapters. Such readers will also be guided to and benefit from relevant subsets of our 100 Statistical Stories.

(iii) Long live our lovely linear readers, but the book is also fine for the *Statistical Stories Readers and Samplers*, for those who already know the basics on statistical models, parameter estimation and testing, some Bayes, etc. As noted, our book has Part One, the theory development in chapters, then Part Two, the stories, but could by a segment of our readers be worked through in the opposite order: read stories, and check what you need to understand and do all details. The [McCullagh \(2022\)](#) book, with aspirations partly parallelling ours, is a case in point.

(iv) Our book is also for the *Specialists Inside Themes*, who wish to learn even more and to see how these subfields are applied in new fashions, possibly in combination with methods and themes from other areas.

The GitHub site **Hjort-Stoltenberg-Book** (xx to be fixed), already pointed t above, is seen as a vital companion to the physical book. It has all datasets used; code in **R** and **Python**, i.e. scripts used to generate our figures, tables, and analyses; and, importantly, solutions to all our 600 Exercises and all our 100 Statistical Stories. The site will be actively maintained, and will with time also include follow-up exercises, snippets of course notes, etc.

the Hjort-
Stoltenberg
book website

We have spent efforts making the book largely self-contained, and desert island readers, with few other books on their shelves, can in principle 'learn nearly everything' from this one (given the prerequisites alluded to in the previous paragraph). Just a few exceptions are as follows. (i) The modern Bayesian statistician needs Markov Chain Monte Carlo, as explained and exemplified in Ch. 7 (and used in various stories), on this

occasion pointing to the later-in-the-book Ch. 13 for basic theory. But lifting that basic theory, from discrete settings to continuous ones, would have taken seven extra pages, so there we instead point to relevant sources. (ii) In Ch. 10 on random processes we scrupulously go through the details of the $C[0, 1]$ space of continuous functions on the unit interval, and also need for several further developments, and applications in certain stories, similar but more complicated mathematical details for the bigger function space $D[0, 1]$. Again, all of this would need eleven more pages, so we point to relevant literature, but do sum up the required details and criteria such that we can proceed to further vital developments and applications, like for Poisson, Gamma, and Beta processes, etc. (iii) [xx is there more? completeness theorem for exponential family. Emil Doob-Meyer. xx]

Declaration

Since the work with this endeavour was finished in 2026, with humankind arguably at one of its cross-roads, a small *declaration of independence* might be uttered: the authors have absolutely not used so-called AI or versions thereof when it comes to the actual writing of our texts, sentences, paragraphs, nor for the many cases, on different scales, where ideas have been hatched and developed, modified, finessed, or abandoned, etc. Occasionally we have been helped by modest-level chat-gpt to find certain datasets and references from older literature, however.

We would similarly hope that our readers and users apply themselves, their cleverness, imagination, enthusiasm, perhaps with flecks of fruitful frustration, when working through exercises and stories – i.e. without being too quick to ask machines to solve problems for them. We hold these truths to be self-evident, that thoughtful thinking (fast and slow) & productive struggle remain the best algorithms for deep learning, genuine creativity, and satisfaction.

Road maps for courses to give based on our book

as many as ten
courses can be
taught from our
book

Several types of courses can be taught from this book. Depending on the course parameters of ambition, time allotted, and background level, lecturers could supplement material with course notes and pointers to specialised literature. This is a list of such, reflecting both the book's general ambitions and our own teaching experiences at the Master and PhD level for statistics and data science at the Department of Mathematics, University of Oslo. Below we provide the skeletons of relevant Road Maps, for the different courses. For each of these, a reasonable going-through of the long list of statistical models of Ch. 1 would be required and useful, without necessarily working through all of them, so below we omit further mention of that chapter. The other chapters point back to that initial cavalcade of models when needed.

- A A basic course in probability models and statistical modelling: more of Ch. 1 than for the other courses, and most of Ch. 3. A suitable selection of these Stories (xx more, edit xx): #1, #2, #17, #24, #27, #44, #61, #71, #76, #74, #75, #77, #79, #85, #86, #87, #90. The lecturer and the students could then for these stories focus on the modelling and interpretation aspects, and perhaps leave the inference aspects for other rainy days.
- B Hard-core classical statistical inference, with parametric models, estimation, testing, risk functions, etc.: half of Ch. 2, then most of Chs. 3, 4, half of 5, 7, 9. A

- suitable selection of these Stories: #1, #2, #6, #17, #22, #26, #27, #28, #31, #39, #41, #44, #47, #53, #62, #76, #77, #78, #83, #88, #99.
- C Modern Bayesian statistics, prior to posterior in classical setups and with MCMC, excursions into nonparametric terrain: the essence of Chs. 3, 5, then good portions of Chs. 7, 14; this could also be coupled with addition emphasis on estimation and decision theory, risk functions, with Ch. 9. A suitable selection of these Stories: #1, #4, #18, #19, #23, #31, #41, #52, #56, #60, #81, #86. #90, #101.
- D Confidence distributions, perhaps alongside classical Bayes, with meta-analysis and combination across information sources: the essence of Ch.s. 3, 5, then healthy portions of 7, 8, 14. A suitable selection of these Stories: #1, #3, #4, #6, #8, #12, #13, #30, #33, #45, #46, #52, #59, #60, #77, #90.
- E Large-sample theory, the careful probability theory leading to CLT and more, ideally coupled with parts of empirical processes, convergence, approximations, applications in statistics: most of Chs. 2, 5, 6, 10, perhaps even with parts of Chs. 11, 13. A suitable selection of these Stories (xx more, edit xx): #15, #21, #22, #24, #27, #34, #44, #46, #50, #51, #60, #77, #83, #85, #87, #93, #96, #100.
- F Survival and event history analysis: models and methods for survival data, counting processes, some martingales, Kaplan–Meier and Nelson–Aalen, Cox and Aalen regression: the essence of Chs. 2, 3, 4, 5, then the full 11. A suitable selection of these Stories: #4, #5, #9, #10, #15, #19, #21, #48, #51.
- G Model selection and model averaging, AIC, BIC, FIC, tolerance radii, cross-validation, applications: the essence of Chs. 2, 3, 4, 5, then the full Ch. 12. A suitable selection of these Stories: #5, #8, #16, #17, #22, #24, #26, #27, #28, #37, #38, #45, #51, #50, #53, #55, #58, #61, #76, #87, #98.
- H Markov chains and processes, recurrence times, stationary distributions, MCMC, with applications: half of Ch. 5, then Chs. 10, 13. A suitable selection of these Stories: #3, #14, #15, #25, #46, #61, #86.
- I Finally we point to the possibility of building a different type of course, on statistical modelling and applications, with emphasis on the stories, and less concern for all the technical details. This is also a way to learn about and appreciate classical and modern statistics, concepts, modelling, computation, scientific context. The selection of stories, in volume and in detail, would of course depend on the time given and the interests of the lecturer and the audience.

Other books

There is a veritable ecosystem of books in good libraries about statistical inference, from the basics to more advanced texts and to specialist treatments. The recurrent core themes remain model building with the required portions of probability theory, parametric estimation, confidence intervals, testing of hypotheses, approximate normality. There are also visible Poisson rates for publishing in new categories, in the directions of machine learning and AI, along with still new takes on classic methodology, related to their active uses with emerging types of new data, and the general ‘combination of many types of accessible information’. With the various parameters one may attempt to measure, related to ambition level, attention to details, the scale from theory to applications (or is it the other way around), horizon width, and the pedagogical lines taken, we trust the Mahalanobis distance from our book to most others is significant and positive.

Classic texts on the basis of the theory include [Lehmann \(1959, 1983\)](#), [Casella and](#)

Berger (2013), Bickel and Doksum (2015), Davison (2003), (xx point to 1-2 more, Brazzale et al. (2007), xx).

Books in this landscape differ when it comes to emphasis on applications, where there naturally is a range from ‘simple illustration to check how things work’ to diving into the scientific context and associated key questions. Since it takes ample time to develop theory, concepts, and methodology, many very good texts are perhaps modest when it comes to application details. Some are more ambitious in that regard, however, giving readers Statistical Stories akin to ours, as e.g. Gelman et al. (2020) with its well-told regression stories and Brazzale et al. (2007) where statistical fine-tuning methods are delved into via a list of such stories. In this regard, our own book is admittedly not ten times more ambitious than McCullagh (2022)’s excellent book, even though we have a hundred stories and that book ten; he goes more deeply into several of these. (xx more from the theory-with-applications books. xx)

After having developed the solid groundwork of statistical inference, in chapters 1–5, our book dares to go decently far also into more specialised topics, expanding the horizon of modern statistical modelling and inference, in chapters 6–14. For these themes there are naturally other more specialised texts; here we may point to Brown (1986), McCullagh and Nelder (1989), and Efron (2023) for exponential families and generalised linear models; Basu et al. (2011) for robustness and minimum divergence; Robert (2007) xx and Gelman xx for Bayes; Schweder and Hjort (2016) for confidence distributions and meta-analysis; Billingsley (1968) and Pollard (1984) for processes and convergence; Andersen et al. (1993); Aalen et al. (2008) with emphasis on survival and event history analysis; Claeskens and Hjort (2008b) for model selection; Ghosh and Ramamoorthi (2003); Hjort et al. (2010); Ghosal and van der Vaart (2017) for Bayesian nonparametrics (xx also mueller xx). We place further comments in the Notes and Pointers sections at the end of our chapters.

(xx we point to *some* other books in similar terrain, though they are quite some distance away from our book. xx) Gelman et al. (2020) (yes, story telling), Ferguson (1967, 1996), Li et al. (2027), Lehmann (1959, 1983, 1999); Lehmann and Casella (1998), Keener (2011), ...

Go forth!

поехали!, as
Gagarin said

After having worked your way through significant chunks of this book, both chapters and well-sampled stories, you’ll be fit to do wonders and be sought after on the job market, the segments associated with the words statistics, data science and analysis, modelling, machine learning, prediction (with uncertainty), and even AI.

Thanks

The authors owe special thanks to Céline Cunen, Gudmund Hermansen, Tore Schweder, for having contributed significantly to several of our Statistical Stories, and also for always pleasant and inspiring long-term collaborations. Deep thanks are also due to a long list of colleagues and friends, who have taken part in discussions and rounds of clarification of relevance to various exercises and stories in our book: Marthe Aastveit, Patrick Ball, Bear Braumoeller, Gerda Claeskens, Aaron Clauset, Dennis Christensen, Riccardo de Bin, Ingrid Dæhlen, Arnaldo Frigessi, Abhik Ghosh, Ingrid Glad, Håvard Hegre, Aliaksandr Hubin, Ingrid Hobæk Haff, Kristoffer Hellton, Anders Dahl Hjort,

Martin Jullum, Vinnie Ko, George Lowther, Ian McKeague, Per Mykland, Per August Moen, Jonas Moss, Håvard Møkleiv Nygård, Lars Olsen, Sonia Petrone, Oskar Høgberg Simensen, Catharina Stoltenberg, Gunnar Taraldsen, Ingrid Van Keilegom, Lars Walløe, Jonathan Williams, Lan Zhang, Anti C. Pated.

anticipated

We have also benefitted, directly and indirectly, through the collective efforts of several grander wide-horizoned funded projects: the *FocuStat: Focus Driven Statistical Inference with Complex Data* 2014-2019 project (led by Hjort) at the Department of Mathematics, University of Oslo, funded by the Norwegian Research Council; the *Stability and Change* 2022-2023 project (led by Hjort and Hegre), funded by and hosted at the Centre for Advanced Study (CAS), Academy of Science and Letters, Oslo; and *Integreat: The Norwegian Centre for Knowledge-Driven Machine Learning* 2023-2033 Centre of Excellence (led by Frigessi and Glad), Oslo, funded by the Norwegian Research Council. We finally acknowledge with gratitude a partial support stipend from the Norwegian Non-Fiction Writers and Translators Association (Norsk faglitterær forfatter- og oversetterforening).

Nils Lid Hjort and Emil Aas Stoltenberg
Blindern, some day in 2026

Contents

Preface	iii
Contents	xi
I 600 Exercises	1
1 Statistical models	3
2 Large-sample theory	43
3 Parameters, estimators, precision, confidence	83
4 Testing, sufficiency, power	113
5 Maximum likelihood	145
6 Minimum divergence	195
7 Bayesian inference and computation	213
8 CDs, confidence curves, combining information	245
9 Loss, risk, performance, optimality	281
10 Converging processes, Brownian motion, and bridges	313
11 Survival and event history analysis	355
12 Model selection and model averaging	385
13 Markov chains and Markov processes	413
14 Bayesian nonparametrics	433
II 100 Statistical Stories	459
15 Demography, Epidemiology, Medicine	461
16 Biology, Climate, Ecology	509

17 Art, History, Literature, Music	543
18 Economics, Political Science, Sociology	589
19 Sports	625
20 Simulated stories	659
21 Miscellaneous stories	683
III Appendix	711
A Sets, functions, and metric spaces	713
B Measure and integration	717
C Probability	733
D Martingales	755
E Overview of stories and data	767
References	787
Name index	801
Subject index	803

Part I

600 Exercises

Bayesian inference and computation

In frequentist parametric inference, there is a fixed underlying true parameter value, say θ_0 , and methods aim at estimating this value, perhaps along with confidence regions or testing. Bayesian inference is radically different, conceptually and operationally. It starts with a prior distribution for the model parameter θ , and proceeds via Bayes theorems to produce the posterior distribution, of the full θ or of relevant focus parameters. Thus ‘not knowing θ well’ is expressed in terms of probability distributions. This chapter goes through these concepts and operations, including also computational schemes to simulate outcomes from the posterior distributions. One prominent class of such schemes amounts to setting up a Markov Chain Monte Carlo algorithm for the given problem, where the stationary distribution for the chain is precisely the required posterior. This makes Bayesian inference possible in a host of complicated setups, without needing to rely on mathematically feasible formulae.

Key words: Bayes solutions, Bernshtein–von Mises theorems, conjugate priors, Jeffreys and vip-Jeffreys prior, loss functions, MCMC, prior to posterior distributions

For any two events A and B with positive probabilities, the $\Pr(A \cap B)$ can be expressed in the two ways $\Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B)$. The Bayes formula in its simplest form is an easy consequence of this, telling us how to modify an initial assessment of A when we’re told that B has taken place:

$$\Pr(A|B) = \Pr(A)\Pr(B|A)/\Pr(B).$$

To better appreciate some of its implications, assume $A_1, \dots, A_k$ are a disjoint union of events, which when combined is the full probability space, so with $p_j = \Pr(A_j)$ we have $\sum_{j=1}^k p_j = 1$. When informed that B has happened, these before-news probabilities are then modified to after-news ones,

$$p_j^* = \Pr(A_j|B) = p_j \Pr(B|A_j)/c \quad \text{for } j = 1, \dots, k, \quad (7.1)$$

with c the normalising constant, i.e. the sum $\sum_{j=1}^k p_j \Pr(B|A_j)$. With more information, say learning that events $B_1, B_2, \dots$ are observed, the A_j probabilities are updated further; we’re learning more and more about the p_j , and with good observations they will become more and more precise, even if the initial p_j were not correctly set up.

The continuous version of the above relates to the two ways of expressing the joint density of (θ, y) , say, with $\pi(\theta)f(y|\theta) = \pi(\theta|y)\bar{f}(y)$. In the Bayesian context, the θ

this is the
Bayes theorem

is the parameter vector, y the data with its likelihood function $L(\theta) = f(y|\theta)$, and $\bar{f}(y) = \int \pi(\theta)L(\theta) d\theta$ the normalisation constant in

$$\pi(\theta|y) = \pi(\theta)L(\theta)/\bar{f}(y). \quad (7.2)$$

The $\bar{f}(y)$ is also the marginal distribution of the data, i.e. the likelihood averaged over the prior. In this representation, both θ and y may be complicated and perhaps high-dimensional. The Bayesian paradigm is to formulate uncertainty about model parameters in this fashion, through probability distributions. The $\pi(\theta)$ is the prior, what one knows before data, and $\pi(\theta|y)$ the posterior, the modified distribution for the unknowns after having seen the data.

Consider for illustration and clarification the classical coin flipping experiment, with θ the probability of ‘head up’. With n independent flips we have $Y \sim \text{binom}(n, \theta)$. The frequentist postulates that there is an underlying true θ_0 , uses perhaps the estimator $\hat{\theta} = Y/n$, reaches the 95 percent interval $I_n = \hat{\theta} \pm 1.96 \{\hat{\theta}(1 - \hat{\theta})\}^{1/2}/\sqrt{n}$, etc. The key property here is $\Pr_{\theta_0}(\theta_0 \in I_n) \doteq 0.95$; so I_n is a random interval, covering the true θ_0 in 95 percent of actual cases. The Bayesian viewpoint is strikingly different, starting with a prior density $\pi(\theta)$ to reflect what might be considered understanding of θ before the first flip. Post flipping, the Bayesian has reached $\pi(\theta|y) \propto \pi(\theta)f(y|\theta)$, with the binomial likelihood. This may e.g. be used to construct a 95 percent posterior interval J_n for θ , with $\Pr(\theta \in J_n | y) = 0.95$. The Bayesian is then not interested in ‘independent repeated experiments’, but just in the data at hand. She is also allowed the statistical luxury of putting prior knowledge into the analysis; if it can be considered known that θ must be close to 0.50, with values outside $[0.40, 0.60]$ less likely than 2 percent, that can effectively be utilised in Bayesian analysis, but not so easily in frequentist analysis.

In this chapter we go through the basics of such constructions and methods, conceptually and operationally. We also uncover conditions under which the frequentist and Bayesian might actually (approximately) agree, in their final inference statements. The two 95 percent intervals I_n and J_n in the previous paragraph will e.g. tend to be very similar, at least with increasing n .

An attractive feature of Bayesian analysis is that the answer, to a sufficiently well-posed inference question, is crystal clear, without having to study competing methods, carrying out performance and comparison analyses, etc. Essentially, if you give a Bayesian (i) data, (ii) a model, (iii) a list of possible actions, (iv) a prior for the unknown parameters, and (v) a loss function, there is a Master Recipe for the very best action.

Modern Bayesian statistics has flourished since around 1980, partly through computer power and algorithms, making calculations possible that would have been too hard for previous generations. The operational goal is often to be able to generate samples from the posterior distribution, and we give methods for accomplishing this, including Markov Chain Monte Carlo (MCMC).

The Bayesian Master Recipe, examples, applications

Ex. 7.1 *Learning in one or two steps.* Let the setup be as in (7.1).

- (a) Prove the formula. Then write out the consequence for say $p_j^{**} = \Pr(A_j | B_1, B_2)$, where B_1 and B_2 are events with positive probabilities.
- (b) Above we have reached the two-pieces-of-news updated probabilities, from initial p_j to p_j^{**} , in one go. Now work through the details to establish that the very same p_j^{**} are

obtained if we view them as coming from two steps of learning; first from p_j to p_j^* , after seeing B_1 , and then from p_j^* to $(p_j^*)^*$, after being told that also B_2 has occurred.

(c) Generalise: you start out Monday morning with $p_1, \dots, p_k$ for the A_j , and update these steadily, day by day, resulting in $\tilde{p}_1, \dots, \tilde{p}_k$ on Sunday evening. Show that these are the same as those obtained by using the (7.1) formula for the full week's data $B_1, \dots, B_7$.

Ex. 7.2 *Poisson data with gamma priors.* This exercise illustrates the basic prior to posterior updating mechanism in a simple Poisson setting. Suppose $Y_1, Y_2, \dots$ are i.i.d. Poisson with unknown mean θ .

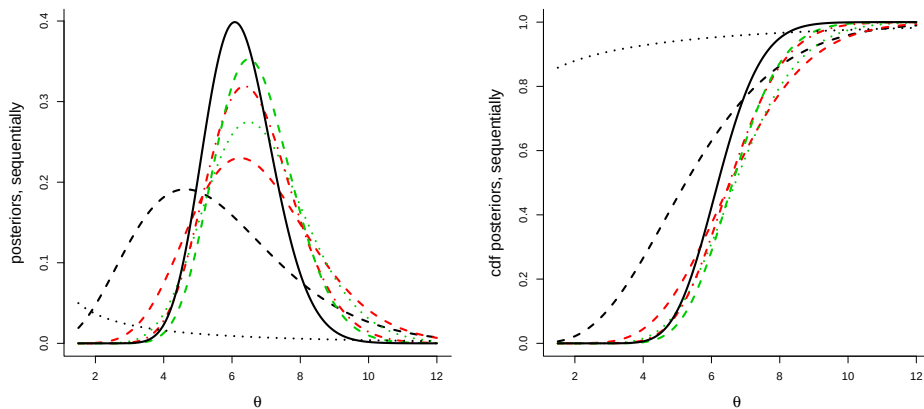


Figure 7.1: Seven curves are displayed, corresponding to the $\text{Gam}(0.1, 0.1)$ initial prior for the Poisson parameter θ , along with the six first updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. The distributions become tighter as more data come in. The smooth black curve is after six data points, the most tight distribution so far. Left panel: the densities; right panel: the c.d.f.s.

(a) Recall definition and properties of the Gamma distribution from Ex. 1.9. In the present Bayesian context, let $\theta \sim \text{Gam}(a, b)$. The prior mean and variance are $a/b = \theta_0$ and $a/b^2 = \theta_0/b$. In particular, low and high values of b signify high and low variability, respectively. Explain how (a, b) may be set from values of prior mean and prior variance. To exemplify, if these are (5.5, 7.7), find (a, b) .

(b) With a single observation Y , which is $\text{Pois}(\theta)$ given θ , show that $\theta | y \sim \text{Gam}(a + y, b + 1)$.

(c) Then suppose there are repeated observations $y_1, \dots, y_n$, being i.i.d. $\sim \text{Pois}(\theta)$ for given θ . Use the above result repeatedly, e.g. interpreting $p(\theta | y_1)$ as the new prior before observing y_2 , etc., to show that $\theta | y_1, \dots, y_n \sim \text{Gam}(a + y_1 + \dots + y_n, b + n)$. Also derive this result directly, i.e. without necessarily thinking about the data having emerged sequentially.

(d) Suppose the prior used is a rather flat $\text{Gam}(0.1, 0.1)$ and that the Poisson data are 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. Reconstruct a version of Figure 7.1 in your computer, plotting the six first posterior densities $p(\theta | \text{data}_j)$ (left panel), where data_j is $y_1, \dots, y_j$, along with the prior density; in the right panel we have the corresponding posterior cumulatives. Complement with another figure also including updated densities 7, 8, 9, 10, for the four last observations, and comment. Also compute the ten Bayes estimates $\hat{\theta}_j = E(\theta | \text{data}_j)$ and the posterior standard deviations, for $j = 1, \dots, 10$.

(e) The mathematics turned out to be rather uncomplicated in this situation, since the Gamma continuous density matches the Poisson discrete density so nicely. Suppose instead that the initial prior for θ is a uniform over $[5.0, 11.0]$. Try to compute posterior distributions, Bayes estimates and posterior standard deviations also in this case, and compare with what you found above. Note also that we in this exercise cared about and managed to reach answers for the posterior distributions $\theta | y_1, \dots, y_n$, without needing to deal with the also implied marginal distribution for $(y_1, \dots, y_m)$. These are not Poisson any longer, with the variability in θ taken into account; see Ex. 7.36.

Ex. 7.3 *The Bayesian Master Recipe.* The general setup is as follows. We have data y_{obs} , seen as the outcome of a random Y over the sample space $\mathcal{Y}$, generated from a model $f_\theta(y) = f(y, \theta)$, with θ having a prior distribution $\pi(\theta) d\theta$ over its parameter space Θ . There is a decision a to be reached, with a belonging to an appropriate action space $\mathcal{A}$, along with a loss function $L(\theta, a)$, a measure of the consequences of decision a if the truth is θ . Whereas the frequentist can attempt different methods for deciding $\hat{a} = \hat{a}(Y)$, then compare risk functions, etc., there is a unique optimal strategy for the Bayesian.

(a) Work through the details of (7.2), including those related to the *marginal distribution* $\bar{f}(y)$. One typically writes that equation as $\pi(\theta | y) \propto \pi(\theta) f_\theta(y)$, which reads ‘posterior is proportional to prior times likelihood’.

(b) A decision $a = \hat{a}(y)$ needs to be reached, as a function of the data. The *Bayes risk* for such a decision is the associated expected loss, $\text{BR}(a, \pi) = E L(\theta, \hat{a}(Y))$, involving randomness on two levels; θ has a prior, and $\hat{a}(Y) | \theta$ is random. Show that it may be expressed in two informative ways:

$$\text{BR}(a, \pi) = \begin{cases} E \{ E L(\theta, \hat{a}(Y)) | \theta \} = \int_{\Theta} R(\theta, \hat{a}) \pi(\theta) d\theta, \\ E \{ E L(\theta, \hat{a}(Y)) | Y \} = \int_{\mathcal{Y}} \{ \int_{\Theta} L(\hat{a}(y), \theta) \pi(\theta | y) d\theta \} \bar{f}(y) dy. \end{cases} \quad (7.3)$$

The first expression involves the frequentist risk function $R(\theta, \hat{a}) = E_{\theta} L(\theta, \hat{a}(Y))$, then averaged with respect to the prior. The inner expectation of the second expression is $E_{\pi} \{ L(\theta, \hat{a}(y)) | Y = y \}$, that is, the expected loss given data.

(c) Show then that the optimal Bayes strategy, the one minimising the Bayes risk, sometimes called the Bayes solution, is achieved by using

$$\hat{a}_B = \operatorname{argmin}_{a \in \mathcal{A}} E_{\theta} \{ L(\theta, a) | y_{\text{obs}} \} = \text{the value of } a \text{ minimising } E_{\theta} \{ L(\theta, a) | y_{\text{obs}} \}.$$

the Bayesian
Master Recipe

Here $E_{\theta} \{ L(\theta, a) | y_{\text{obs}} \}$ is the expected posterior loss. The function g is evaluated and minimised over all a , for the given data $y = y_{\text{obs}}$. This is the Bayes recipe. Note that using the recipe in practice only concerns the observed data y_{obs} , and that one does not need to evaluate its risk function. The *minimum Bayes risk* $\text{MBR}(\pi) = \text{BR}(\hat{a}_B, \pi)$ is the resulting Bayes risk when using this optimal strategy.

minimum Bayes
risk

(d) Above results have been presented and reached in terms of prior and posterior densities, $\pi(\theta)$ and $\pi(\theta | y)$, partly for notational convenience. Show that the arguments go through also for more general priors; these may in particular be mixtures over continuous and discrete measures.

Ex. 7.4 *Some loss functions and their associated Bayes rules.* The Master Recipe of Ex. 7.3 is completely general, and can be applied in new and complicated situations, as long as we have data, a model, a prior for the unknowns, and a loss function. In the Bayesian setup finding or evaluation of the posterior distribution of the parameters is

always important, carrying separate weight, but if clear decisions are needed one needs also the loss function, say $L(\theta, a)$. Here we go through a short list of commonly used loss functions.

(a) For estimating a one-dimensional θ , with squared error loss $L(\theta, a) = (a - \theta)^2$, show that the Bayes estimator is $\hat{\theta}_B = E(\theta | y)$, the posterior mean. Show also that the minimum Bayes risk, with this loss function, can be expressed as $MBR(\pi) = E \text{Var}(\theta | Y)$.

(b) If the loss function is $L(\theta, a) = w(\theta)(a - \theta)^2$, show that the Bayes estimator is

$$\hat{\theta}_B = E\{w(\theta)\theta | \text{data}\} / E\{w(\theta) | \text{data}\}.$$

In particular, when estimating a positive θ using loss $(a - \theta)^2/\theta$, show that the Bayes estimator is $1/E\{(1/\theta) | \text{data}\}$, the inverse of the inverse mean.

(c) For the Poisson-gamma setup of Ex. 7.2, we found $\theta | y \sim \text{Gam}(a + n\bar{y}, b + n)$, with $\bar{y}$ the average of data $y_1, \dots, y_n$. Show that the Bayes estimator is $\hat{\theta}_A = (a + n\bar{y})/(b + n)$ for squared error loss, but $\hat{\theta}_B = (a + n\bar{y} + 1)/(b + n)$ for loss function $\theta(a - \theta)^2$.

(d) Consider the natural absolute loss function, $L(\theta, a) = |a - \theta|$. Show that the Bayes solution becomes the posterior median, i.e. $\hat{\theta}_B = G^{-1}(\frac{1}{2} | \text{data})$, where $G(\theta | \text{data})$ is the posterior c.d.f.

(e) A differently spirited loss function is $L_\varepsilon(\theta, a) = I(|a - \theta| > \varepsilon)$, for some small ε ; zero loss if you're very close, but then a fixed positive loss if you err with more than ε . Show that the Bayes solution is the a which makes the probability $\Pr(\theta \in [a - \varepsilon, a + \varepsilon] | \text{data})$ as big as possible. Explain that this for small ε means finding the mode of the posterior density. For the Poisson setup again, show that with $\theta \sim \text{Gam}(a_0, b_0)$, the Bayes estimator becomes $\hat{\theta}_C = (a_0 + n\bar{y} - 1)/(b_0 + n)$.

(f) Suppose one needs the joint estimation of several parameters, say all of $\theta = (\theta_1, \dots, \theta_p)$, via the loss function $L(\theta, a) = (a - \theta)^t M (a - \theta)$, for an appropriate full-rank symmetric matrix M . Show that the Bayes solution again is the posterior mean, but now for the full vector, i.e. $E(\theta | \text{data})$. In particular, the Bayes solution does not depend on the M matrix, though the actual posterior expected loss, and the Bayes risk, do.

(g) In the previous subquestions the framework has been that of estimating a one-dimensional θ . Check that you understand how these results and insights, for the Bayes solutions, change when the situation is changed to that of estimating a *focus parameter*, say $\mu = g(\theta_1, \dots, \theta_p)$, a function of the full model parameter.

Ex. 7.5 *How many streetcars in San Francisco?* The streetcars in this city are numbered $1, \dots, N$. You observe $Y = 203$ and wonder what N is.

(a) Supposing Y has the uniform distribution on $1, \dots, N$, set up the likelihood, and identify the ML estimate. With the prior $\pi(N)$ proportional to $1/N$, say for $N = 1, 2, \dots, N_{\max}$ for a suitably high $N_{\max}$, find the posterior distribution, along with its mean and median.

(b) Suppose you after having seen no. 203 also see nos. 157, 222. Update your posterior, and again find the mean and median, qua updated estimates.

(c) The Bayesian setup allows any choice for the prior, so think for a minute and construct your own $\pi_{\text{you}}(N)$. Do the updating, for the data 203, 157, 222, and compare with the posterior found above. What is your best estimate of N , if your loss function is of 0-1 type, with $L(N, \hat{N})$ being 0 if $\hat{N} = N$ and 1 if $\hat{N} \neq N$?

all or nothing

for a non-Bayes
treatment, see
Ex. 8.9

Ex. 7.6 *A Bayesian take on hypothesis testing.* Assume the model parameter θ is either in Θ_0 , which we may call the null hypothesis, or not, i.e. in its complement Θ_0^c . Suppose also that the statistician needs to make a decision, either to reject the null, or to accept it. This is the basic framework of hypothesis testing, see Ch. 4, but we now consider the problem from a Bayesian viewpoint.

(a) The decision space is {accept, reject}. For the loss function, take $L(\theta, \text{accept})$ equal to 0 or L_0 , if θ is inside or outside Θ_0 , and $L(\theta, \text{reject})$ equal to 0 or L_1 , if θ is outside or inside Θ_0 . Show that

$$\mathbb{E}\{L(\theta, \text{accept}) \mid \text{data}\} = L_0 p(\text{data}), \quad \mathbb{E}\{L(\theta, \text{reject}) \mid \text{data}\} = L_1 \{1 - p(\text{data})\},$$

where $p(\text{data}) = \Pr(\theta \in \Theta_0^c \mid \text{data})$, the probability that the null is wrong, as measured by the Bayesian posterior distribution.

(b) Deduce that one should reject the null when the probability $p(\text{data})$ for its falseness is sufficiently overwhelming, namely when $p(\text{data}) \geq L_1/(L_0 + L_1)$. – If this threshold is 0.95, for example, show that this corresponds to $L_1/L_0 = 19$. Briefly discuss ways of assigning losses L_0 and L_1 .

(c) Suppose now there is a ‘doubt option’, refraining from deciding on the either-or dichotomy. The setting is as above, with losses L_0 and L_1 , but the action list is {accept, reject, doubt}, with a fixed loss L_d for deciding to be in doubt. Explain that the expected losses given the data are $L_0 p(\text{data})$, $L_1 \{1 - p(\text{data})\}$, L_d , and spell out what the Bayes solution is.

Ex. 7.7 *Which subset does the model parameter belong to?* Consider a setup with data from a model with model parameter θ inside its region Θ . Suppose you need to take one of five different possible decisions, $D_1, \dots, D_5$, and that these are related to where the underlying parameter θ is positioned; if $\theta \in \Theta_j$ the best decision would be D_j , for $j = 1, \dots, 5$. Here the Θ_j are disjoint and their union is the full parameter region.

(a) Suppose the loss function $L(\theta, D_j)$ is 0 if $\theta \in \Theta_j$ and 100 if $\theta \notin \Theta_j$. Show that $\mathbb{E}\{L(\theta, D_j) \mid \text{data}\} = 100\{1 - p_j(\text{data})\}$, where $p_j(\text{data}) = \Pr(\theta \in \Theta_j \mid \text{data})$. Hence show that the optimal Bayes strategy is to take the decision associated with the highest posterior probability $p_j(\text{data})$.

(b) Assume there in addition is a ‘doubt option’, associated with a doubt cost $L_d = 10$; this could e.g. mean planning for getting further data. With decision space $\{D_1, \dots, D_5, \text{doubt}\}$, what is now the Bayesian strategy?

(c) Generalise the previous setup, and results, to the case where the costs associated with reaching the wrong decision are not equally balanced, say $L(\theta, D_j) = c_{i,j}$, if $\theta \in \Theta_i$, for $i = 1, \dots, 5$, with $c_{i,i} = 0$ but the other $c_{i,j}$ positive.

Binomial and beta, multinomial and Dirichlet

Ex. 7.8 *The binomial-beta setup.* Let Y given p be a binomial (n, p) , and for p take a Beta(a, b) prior, see Ex. 1.22. There we worked with the marginal distribution of Y , and looked at certain properties, but here our aims are Bayesian.

(a) Show that $p \mid y \sim \text{Beta}(a+y, b+n-y)$. This is the main and always crucial updating step, getting from the prior to the posterior. In the present case the step is an easy one, since there is only one unknown parameter, and since the product of the prior and the likelihood takes an easy form. This is an instance of *conjugate priors*, where (i) the prior

prior updated
to posterior

conjugate
priors, as with
the Poisson-
gamma of
Ex. 7.2

is inside a nice class; (ii) the likelihood is nice enough to make also the posterior belong to the same class. Give a description of the posterior also for the not quite so standard case where the prior for p is uniform on $[0.30, 0.70]$.

(b) Going back to the $\text{Beta}(a, b)$ prior again, show that the Bayes estimator, under squared error loss, is $\hat{p}_B = (1 - w_n)p_0 + w_n y/n$, where $p_0 = a/(a + b)$ is the prior mean and $w_n = n/(a + b + n)$. For the case of a uniform prior, show that this leads to $(y + 1)/(n + 2)$. Compute the risk functions $r(p) = E_p(\hat{p} - p)^2$, for the classic frequentist Y/n and for the this $(Y + 1)/(n + 2)$, and find the interval where the latter is better than the former.

(c) It is helpful to parametrise (a, b) as $(cp_0, c(1 - p_0))$, with prior mean p_0 and prior variance $p_0(1 - p_0)/(c + 1)$. Explain that for the Bayes estimator and the posterior variance we then have

$$\hat{p}_B = E(p|y) = \frac{cp_0 + y}{c + n}, \quad \text{Var}(\theta|y) = \frac{\hat{p}_B(1 - \hat{p}_B)}{n + c + 1}.$$

Use results from Ex. 1.22 to show that with its marginal distribution, $\text{Var } Y = np_0(1 - p_0)(c + n)/(c + 1)$.

(d) Then use $\text{Var } p = E \text{Var}(p|Y) + \text{Var } E(p|Y)$ to find that with the $\text{Beta}(cp_0, c(1 - p_0))$ prior, the minimum Bayes risk is $\text{MBR} = p_0(1 - p_0)c/\{(c + 1)(c + n)\}$. Show that the maximum of the minimum Bayes risk happens for the prior $\text{Beta}(\frac{1}{2}\sqrt{n}, \frac{1}{2}\sqrt{n})$, with value $\frac{1}{4}/(\sqrt{n} + 1)^2$, and consequent Bayes estimator

$$\hat{p}_B = \frac{\frac{1}{2}\sqrt{n} + Y}{\sqrt{n} + n} = \frac{1}{\sqrt{n} + 1} \frac{1}{2} + \frac{\sqrt{n}}{\sqrt{n} + 1} \frac{Y}{n}. \quad (7.4)$$

This is the *least favourable prior*; explain that if Nature chooses p from this distribution, then this is the worst of all Beta scenarios for the statistician. The $\hat{p}_B$ estimator has a corresponding *minimax* property, no competing estimator has a smaller maximum risk; see material on this in Ch. 9.

(e) It is useful to see the minimum Bayes risk in two ways, via (7.3), i.e. also via the risk function of the Bayes solution. Show that this risk function becomes

$$r_B(p) = E_p(\hat{p}_B - p)^2 = \frac{1}{(c + n)^2} \{np(1 - p) + c^2(p - p_0)^2\}.$$

Then take the mean of this with respect to the prior to find the MBR formula above again. For the least favourable prior, show that the risk function is constant, equal to $\frac{1}{4}/(\sqrt{n} + 1)^2$.

(f) If Y is from the binomial (n, p_{true}) model, show that Y/n and the Bayes estimator $\hat{p}_B$ are large-sample equivalent, with $\sqrt{n}(Y/n - \hat{p}_B) \rightarrow_{\text{pr}} 0$, as long as $c = a + b$ stays fixed or bounded. Deduce that they have the same limit distribution. This does not hold for the least favourable prior, however, as its parameters $(\frac{1}{2}\sqrt{n}, \frac{1}{2}\sqrt{n})$ then move with sample size; find the limit distribution of $\sqrt{n}(\hat{p}_B - p_{\text{true}})$ for this case.

Ex. 7.9 Binomial-beta with weighted loss function. One's best decisions depend on how the consequences are to be evaluated. Consider as in Ex. 7.8 the $Y \sim \text{binom}(n, p)$ situation, but now with the weighted loss function $L(p, \hat{p}) = (\hat{p} - p)^2/\{p(1 - p)\}$.

(a) Discuss briefly what such a loss function entails, compared to the traditional squared error loss, regarding situations where the p probabilities are close to the boundaries. For the ML estimator Y/n , show that the risk function is the constant $1/n$.

(b) With p having any prior, show that the Bayes estimate is

$$\hat{p}_B = E\left(\frac{1}{1-p} \mid y\right) / E\left(\frac{1}{p(1-p)} \mid y\right).$$

If p has a $\text{Beta}(a, b)$ distribution, show that the formula for $E\{1/(1-p)\}/E\{1/p(1-p)\}$ becomes $(a-1)/(a+b-2)$, provided $a > 1$ and $a+b > 2$. With the uniform prior, show that this again leads to $\hat{p}_B = y/n$, first for cases $y = 1, \dots, n-1$, then also for cases $y = 0$ and $y = n$, where some extra care is required for the two boundary cases.

Ex. 7.10 *Bayes estimators are biased.* Except for a few particular cases, perhaps involving special loss functions, the Bayes estimator is not unbiased; it stems from a prior, which is making its inherently biased presence felt. To be specific, suppose a one-dimensional θ is to be estimated with quadratic loss $(a - \theta)^2$, based on observations Y , and that one starts with a proper prior for the θ . We assume, for the arguments to go through smoothly, that θ has a finite prior variance.

(a) We have seen in Ex. 7.4 that the Bayes estimator is $\hat{\theta}_B(y) = E(\theta \mid y)$. Suppose it is unbiased, $E_\theta \hat{\theta}_B(Y) = \theta$ for all θ . Go through the details of the following arguments, where the expectation operator refers to the joint distribution of (θ, Y) :

$$E(\hat{\theta}_B - \theta)^2 = E\theta^2 - E\theta\hat{\theta}_B - E\hat{\theta}_B\theta + E\hat{\theta}_B^2 = E\theta^2 - E\theta^2 - E\hat{\theta}_B^2 + E\hat{\theta}_B^2 = 0.$$

Conclude that unbiasedness can only take place in the trivial case where $\hat{\theta}_B = \theta$ with probability 1, which would mean θ is known.

(b) There are various situations where the posterior mean estimator is close to being unbiased, but by the argument above there will be some remaining bias. In the binomial setup of Ex. 7.8, the Bayes estimator is $(c\theta_0 + Y)/(c + n)$, which becomes the familiar unbiased Y/n for $c \rightarrow 0$; this limiting case, corresponding to a prior proportional to $\theta^{-1}(1-\theta)^{-1}$, is not a proper prior (finite integral), however. We learn that posterior mean estimators computed under certain improper priors (infinite integral) might be unbiased. Note that the arguments used in (a) involve double expectation rules, which presuppose a proper prior.

(c) Explain why deriving the Bayes estimator Y/n in Ex. 7.9 does not contradict what is found above.

Ex. 7.11 *The multinomial-Dirichlet setup.* Here we extend the setup and result of binomial-Beta, to the case of three or more categories. We start with $Y = (Y_1, \dots, Y_k)$ which for given $p = (p_1, \dots, p_k)$ is a multinomial $(n, p_1, \dots, p_k)$. For p we take the $\text{Dir}(a_1, \dots, a_k)$ prior. For details regarding the multinomial and the Dirichlet, see Ex. 1.5 and 1.20.

(a) Show the important and useful result that $(p_1, \dots, p_k) \mid (y_1, \dots, y_k) \sim \text{Dir}(a_1 + y_1, \dots, a_k + y_k)$.

(b) Show that the Bayes estimator under squared error loss becomes

$$\hat{p}_{i,B} = E(p_i \mid \text{data}) = \frac{a_i + y_i}{a + n} = (1 - w_n)p_{0,i} + w_n(y_i/n)$$

convex
combination of
prior mean and
sample mean

for $i = 1, \dots, k$, with prior means $p_{0,i} = a_i/a$, with $a = a_1 + \dots + a_k$, and weight $w_n = n/(a + n)$. Find also the posterior variance and posterior correlation between p_i and p_j .

(c) Find the Bayes estimators, under quadratic loss, for the geometric mean parameters $(p_1 p_2)^{1/2}$ and $(p_1 \dots p_k)^{1/k}$.

Ex. 7.12 *Eventually you will learn the truth.* Set up a suitable simulation experiment in your computer, with true probabilities $p_{\text{true}} = (p_{\text{true},1}, \dots, p_{\text{true},k})$ for boxes $1, \dots, k$, along with prior guesses $p_0 = (p_{0,1}, \dots, p_{0,k})$. You may e.g. take $k = 20$ and something like $p_{\text{true},j} = \sin^2(j)/d_k$, or some other interesting shape, and perhaps very different from your prior guesses.

(a) For the Dirichlet prior with parameters $(cp_{0,1}, \dots, cp_{0,k})$, generate one multinomial observation $(y_1, \dots, y_k)$ at the time (i.e. only one of these is a 1, the other zeros), do the updating, and plot the curves p_{true} and $\hat{p}_B$ on your screen, sequentially, with $\hat{p}_B$ the best guess for p_{true} given what has been observed up to that step.

(b) Run such a code with different versions of prior, prior strength c , etc., to see how much data it takes in order for the $\hat{p}_B$ to eventually come close to p_{true} . Illustrate in such a fashion that even if you are strongly wrong in the beginning, the data will reveal the truth with enough data.

Ex. 7.13 *Gott würfelt nicht.* For the multinomial-Dirichlet setup of Ex. 7.11, we reached the posterior characterisation $p | \text{data} \sim \text{Dir}(a_1 + y_1, \dots, a_k + y_k)$. The importance of this lies in the easy usefulness of simulations, where the posterior distribution of any functions of $(p_1, \dots, p_k)$ may be read off.

there are also R
packages with
direct
rdirichlet

(a) Explain how you may simulate e.g. 10^5 vectors $p = (p_1, \dots, p_k)$ from the posterior distribution, using the characterisation from Ex. 1.20. Concretely, show that one may use $p_1 = G_1/G, \dots, p_k = G_k/G$, with independent $G_1 \sim \text{Gam}(a_1 + y_1), \dots, G_k \sim \text{Gam}(a_k + y_k)$, and sum $G = G_1 + \dots + G_k$.

(b) Suppose you throw a certain and perhaps not entirely standard die 50 times and have counts (2, 7, 4, 12, 14, 11) of outcomes 1, 2, 3, 4, 5, 6. Use either of the priors (i) ‘flat’, $\text{Dir}(1, 1, 1, 1, 1, 1)$; (ii) ‘symmetric and more confident’, $\text{Dir}(4, 4, 4, 4, 4, 4)$; (iii) ‘unwilling to guess’, $\text{Dir}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, for the probabilities $(p_1, \dots, p_6)$, to assess the posterior distribution of each of the following quantities:

Un coup de dés
jamais n’abolira
le hasard

$$\alpha = \frac{p_6}{p_1}, \quad \beta = (1/6) \sum_{j=1}^6 (p_j - 1/6)^2, \quad \gamma = (1/6) \sum_{j=1}^6 |p_j - 1/6|, \quad \delta = \frac{(p_4 p_5 p_6)^{1/3}}{(p_1 p_2 p_3)^{1/3}}.$$

For each of $\alpha, \beta, \gamma, \delta$, and for each of the priors, give the 0.05, 0.50, 0.95 quantile points, from 10^5 simulations from the posterior distributions. This leads to *credibility intervals*, different in spirit and interpretation from the frequentist *confidence intervals*, as per (3.3). Carry out the necessary large-sample approximation, with the delta method for the ML estimator $\hat{\delta} = (\hat{p}_4 \hat{p}_5 \hat{p}_6)^{1/3} / (\hat{p}_1 \hat{p}_2 \hat{p}_3)^{1/3}$, to find both

credibility and
confidence

$$\Pr_{\text{true}}(\delta_{\text{true}} \in \hat{\delta} \pm 1.645 \text{sd}/\sqrt{n}) \quad \text{and} \quad \Pr_B(\delta \in [\text{low}, \text{up}] | \text{data}) = 0.90,$$

with the latter to be evaluated for each of the three priors, and for emphasis writing ‘true’ for probabilities under a fixed but unknown parameter vector and ‘B’ for Bayes. Explain these differences in perspectives. Note the relative ease with which credibility intervals may be read off from posterior simulations, without needing separate mathematical efforts like the delta method. You should also plot the posterior densities, for each of the four quantities noting the extent to which the prior influences the results.

(c) For the case of $\alpha = p_6/p_1$, exact numerical simulation is possible, without simulation. Do this, and compare with the answers reached via simulation.

(d) The above priors are slightly artificial in this context, since they do not allow the explicit possibility that the die in question is plain boring utterly simply a correct one,

i.e. that $p = p_0 = (1/6, \dots, 1/6)$. The priors used hence do not give us the possibility to admit that perhaps $\rho = 1, \alpha = 0, \beta = 0, \gamma = 1$, after all. This motivates using a mixture prior which allows a positive chance for $p = p_0$. Redo therefore the Bayesian analysis above, with the same $(2, 7, 4, 12, 14, 11)$ data, for the prior $\frac{1}{2} \delta(p_0) + \frac{1}{2} \text{Dir}(1, 1, 1, 1, 1, 1)$. Here $\delta(p_0)$ is the ‘degenerate prior’ that puts unit point mass at position p_0 . Compute in particular the posterior probability that $p = p_0$, and display the posterior distributions of $\rho, \alpha, \beta, \gamma$.

Ex. 7.14 *Beta-binomial and Dirichlet-multinomial: normal approximations.* (xx do this. seeing the natural similarities. xx) Consider again the Beta-binomial setup of Ex. 7.8, where $Y | \theta \sim \text{binom}(n, p)$ with a Beta(a, b) prior for p . We shall look into certain mirror properties, for the frequentist and the Bayesian inference methods for p . Below let $p^* = Y/n$ be the ordinary estimator for p .

precursor to the
Bernshtein–von
Mises theorems
of Ex. 7.39

(a) As a separate lemma, used below, show that if $X_c \sim \text{Beta}(cp_0, c(1 - p_0))$, then $c^{1/2}(X_c - p_0) \rightarrow_d N(0, p_0(1 - p_0))$. You may use the Scheffé lemma, see Ex. 2.6, or utilise the representation $X_c = U_c/(U_c + V_c)$, where $U_c \sim \text{Gam}(cp_0, 1)$ and $V_c \sim \text{Gam}(c(1 - p_0), 1)$ are independent, along with limiting normality of Gamma variables; confirm that this representation follows from details of Ex. 1.20.

(b) When there is a true p_{true} at work for the data Y , show from Ch. 2, or perhaps via likelihood results of Ch. 5, that $U_n = \sqrt{n}(p^* - p_{\text{true}}) \rightarrow_d U \sim N(0, p_{\text{true}}(1 - p_{\text{true}}))$. Then work with the normalised posterior distribution for $U_{n,B} = \sqrt{n}(p - \hat{p}_B)$, and show that it has the very same normal limit distribution.

(c) Show that $\hat{p}_B$ and p^* are close enough to deduce also that $\sqrt{n}(p - p^*) | \text{data} \rightarrow_d U$. This is the Bayesian mirror result to the frequentist one. Explain that the frequentist confidence intervals and Bayesian credibility intervals, here of the type

$$p^* \pm 1.645 \{p^*(1 - p^*)\} / \sqrt{n}, [\text{Be}^{-1}(0.05, a + y, b + n - y), \text{Be}^{-1}(0.95, a + y, b + n - y)],$$

become first-order equivalent, for growing n .

(d) The close-to-normal lemma for the Dirichlet, generalising the above for the Beta, is as follows. If $(X_{c,1}, \dots, X_{c,k}) \sim \text{Dir}(cp_{0,1}, \dots, cp_{0,k})$, then the vector $c^{1/2}(X_{c,1} - p_{0,1}, \dots, X_{c,k} - p_{0,k})$ tends to a zero-mean multinormal $(U_1, \dots, U_k)$ with covariances $\text{cov}(U_j, U_\ell) = p_{0,j}(\delta_{j,\ell} - p_{0,\ell})$. Note that this matches the multinomial covariance structure; see Ex. 2.51. To prove this you may use the gamma ratio representation $X_{c,1} = V_{c,1}/V_c, \dots, X_{c,k} = V_{c,k}/V_c$, in terms of independent $V_{c,j} \sim \text{Gam}(cp_{0,j}, 1)$ with sum V_c , and then apply the delta method. Then use this for posterior distribution in Dirichlet-multinomial set up of Ex. 7.11.

Ex. 7.15 *Minimax estimation of multinomial parameters.* We shall consider estimation in the general multinomial model, but for concreteness start with the trinomial case. Consider therefore (X, Y, Z) trinomial (n, p, q, r) , with $p + q + r = 1$. We work with the loss function

$$L((p, q, r), (a, b, c)) = (a - p)^2 + (b - q)^2 + (c - r)^2$$

for estimating (p, q, r) jointly.

(a) Show that using a Dirichlet (cp_0, cq_0, cr_0) for (p, q, r) leads to Bayes estimators $\hat{p} = (cp_0 + x)/(c + n)$, $\hat{q} = (cq_0 + y)/(c + n)$, $\hat{r} = (cr_0 + z)/(c + n)$, and that the associated risk function becomes

$$\frac{1}{(c + n)^2} \{np(1 - p) + c^2(p - p_0)^2 + nq(1 - q) + c^2(q - q_0)^2 + nr(1 - r) + c^2(r - r_0)^2\}.$$

(b) Show that the minimum Bayes risk, for such a prior, becomes

$$\text{MBR} = \frac{c}{(c+1)(c+n)} \{p_0(1-p_0) + q_0(1-q_0) + r_0(1-r_0)\}.$$

Show further that this is maximised over c , for $c = \sqrt{n}$, and then over all prior means for $(p_0, q_0, r_0) = (1/3, 1/3, 1/3)$. Verify that the maximised minimum Bayes risk becomes $(2/3)/(\sqrt{n}+1)^2$. This invites using the $\text{Dir}(\sqrt{n}(1/3, 1/3, 1/3))$ prior, with ensuing estimators $(\hat{p}, \hat{q}, \hat{r})$ equal to

$$\frac{1/3}{\sqrt{n}+1} + \frac{\sqrt{n}}{\sqrt{n}+1} \frac{x}{n}, \quad \frac{1/3}{\sqrt{n}+1} + \frac{\sqrt{n}}{\sqrt{n}+1} \frac{y}{n}, \quad \frac{1/3}{\sqrt{n}+1} + \frac{\sqrt{n}}{\sqrt{n}+1} \frac{z}{n}.$$

(c) Show that this estimation scheme has constant risk function, equal to $(2/3)/(\sqrt{n}+1)^2$, i.e. identical to the maximised minimum Bayes risk. As for the binomial parallel (7.4), this $(\hat{p}, \hat{q}, \hat{r})$ method has a nice minimax property, see Ex. 9.13.

(d) Generalise the above, from the trinomial to the general multinomial setting.

Sampling from the posterior via Markov Chain Monte Carlo

Ex. 7.16 MCMC, I: simulating from a given distribution. Suppose we need to simulate realisations from a given and possibly complicated density $f(x)$ on some domain. The Metropolis algorithm works as follows. The task is to form a long chain $x_1, x_2, \dots$ in your computer, where its limiting distribution is precisely the given f . After having generated x^{old} , decide on x^{new} by (i) deciding on a proposal, x^{prop} , drawn from a suitable distribution symmetric in $(x^{\text{old}}, x^{\text{prop}})$, and then (ii) accepting this proposal with probability $p_{\text{accept}} = \min(1, f(x^{\text{prop}})/f(x^{\text{old}}))$. It is to be noted that any factors of proportionality, as with $f(x) = cf_0(x)$ with a perhaps complicated c , cancel out and disappear. In algorithmic terms,

$$x^{\text{new}} = (1 - \text{ok}) x^{\text{old}} + \text{ok} x^{\text{prop}}, \quad \text{where } \text{ok} = I(\text{accept}) = I(u \leq p_{\text{accept}}),$$

writing u for a simulated uniform. Markov chain theory of Ch. 13, see Ex. 13.19 and 13.21, secures that this scheme works, in the sense that the chain converges in distribution to that of π . Here, handling distributions with densities, explain that we have both $\Pr(x_n \in A) \rightarrow \int_A f(x) dx$ for all sets A , and $\bar{g}_n = n^{-1} \sum_{i=1}^n g(x_i) \rightarrow_{\text{pr}} g_0 = \mathbb{E}_f g(X) = \int g f dx$ for each bounded g .

(a) Set up such a scheme to generate outcomes from the standard normal (ignoring the existence of simpler direct algorithms). Start at any x_0 , then draw proposals $x^{\text{prop}} \sim \text{unif}[x^{\text{old}} - a, x^{\text{old}} + a]$, with acceptance probabilities set up via the general scheme above. Run the chain for a suitably long time and check with a fine histogram that the distribution matches the normal. Keep track of the acceptance probabilities and the overall acceptance rate. The theory works and says that the chain will converge to the standard normal, for any positive fine-tuning parameter a ; explain however in which ways too small or too large values of a will be ineffective.

(b) Consider a somewhat harder challenge, simulating realisations from the density f of $0.05 \text{N}(-2, 1) + 0.90 \text{N}(0, 1) + 0.05 \text{N}(2, 1)$ via MCMC. Set up a chain that converges to this f in distribution; check that the output produces a fine histogram matching this f .

(c) The algorithm works also in higher dimension. Set up a chain that produces outcomes (x, y, z) from the model on the unit cube with density $f(x, y, z) = 1 + \theta(x - \frac{1}{2})(y - \frac{1}{2})(z - \frac{1}{2})$,

“Equation of
State
Calculation by
Fast Computing
Machines”,
1953

MCMC

called
ergodicity

with $\theta = 3.45$. From the output, read off the correlations, and the probability that $Z \geq (XY)^{1/2}$.

(d) With such MCMC output one typically reads off key quantities, like means, quantiles, probabilities for certain sets, etc. The crucial point is that sample averages converge, even though the chains being sampled do not produce i.i.d. output. Check with Ex. 3.41 to explain that with such a chain x_j , with $x_j \rightarrow_d x$ in distribution, we have $m^{-1} \sum_{j=1}^m h(x_j)$ converging to $h_0 = \int h(x)f(x) dx$, for each function h , though perhaps with higher sampling variability than if the x_j had been pure i.i.d. Explain also that a histogram, say based on 10^4 values of x_j , is a bona fide estimator for the underlying density $f(x)$. These matters of autocorrelated chain output do not concern us much as long as computers give us long chains.

Ex. 7.17 MCMC, II: simulating from the posterior. The Metropolis algorithm of Ex. 7.16 is generic, and finds extensive use in Bayesian statistics, where the object is to sample from the posterior distribution, given the data at hand.

(a) With the posterior density written as $\pi(\theta | \text{data}) = c \pi(\theta)L(\theta)$, in terms of prior times likelihood and a proportionality constant c , show that the Metropolis algorithm can be described as follows. Form a long Markov chain $\theta_1, \theta_2, \dots$ in your computer, starting anywhere in the parameter region. After having generated θ^{old} , (i) get hold of a proposal θ^{prop} , from a proposal distribution symmetric in $(\theta^{\text{old}}, \theta^{\text{prop}})$, and then (ii) accept it with probability $p_{\text{accept}} = \min(1, r)$, with ratio

$$r = \frac{\pi(\theta^{\text{prop}} | \text{data})}{\pi(\theta^{\text{old}} | \text{data})} = \frac{\pi(\theta^{\text{prop}})L(\theta^{\text{prop}})}{\pi(\theta^{\text{old}})L(\theta^{\text{old}})} = \frac{\pi(\theta^{\text{prop}})}{\pi(\theta^{\text{old}})} \exp\{\ell(\theta^{\text{prop}}) - \ell(\theta^{\text{old}})\}.$$

Note that the proportionality constant c has disappeared, so for both the prior and the likelihood we merely need to give these up to a proportionality factor.

(b) In the Poisson-gamma setup of Ex. 7.2 we could find the posterior distribution directly. Suppose however that the prior is outside the gammas, say uniform on $[4.0, 8.0]$. Set up an MCMC to sample from the posterior, given data 6, 8, 7, 6, 7, 4, 11. Compute also the mean, median, standard deviation.

(c) As an illustration, generate first $y_1, \dots, y_n$ from the uniform distribution on the unit interval. For these data, consider the log-linear density model $f(y, \theta_1, \theta_2) = \exp\{\theta_1(y - \frac{1}{2}) + \theta_2(y - \frac{1}{2})^2 - k(\theta_1, \theta_2)\}$, briefly studied in Ex. 1.59, and use independent zero-mean priors for θ_1, θ_2 , with standard deviations 1.0, 0.5. Set up an MCMC for the posterior of (θ_1, θ_2) . Plot say 25 densities drawn from the posterior model in this way.

Bayesian analysis for normal models

Ex. 7.18 The Woodbury matrix identity and some special cases. It will be useful for some of the linear algebra involved in linear normal models to have a few matrix identities on board first. The *Woodbury identity* says that

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1},$$

with A and C invertible matrices of size $n \times n$ and $p \times p$, along with U and V of sizes $n \times p$ and $p \times n$. The point for us is to have them and use them, for likelihoods and for Bayesian work; this is not a linear algebra book. after all

(a) Prove the Woodbury, starting by multiplying $A + UCV$ with the right hand side. For the determinant, the point partly being to come down from a high n to a lower p ,

show that

$$|A + UCV| = |C^{-1} + VA^{-1}U| |C| |A| = |A| |C| |I_p + CVA^{-1}U|.$$

(b) For u and v vectors of size n , show that

$$(A + uv^t)^{-1} = A^{-1} - A^{-1}uv^t A^{-1} / (1 + u^t A^{-1}v),$$

$$|A + uv^t| = |A| (1 + u^t A^{-1}v).$$

Ex. 7.19 *The normal prior and posterior with normal data.* Here we go through the basic steps and results for situations with normal data and normal priors for unknown mean parameters. More elaborate constructions and technical issues are needed when there in addition are unknown parameters in the variance and covariance structure, to be pursued in Ex. 7.24, 7.37.

(a) Assume iqs in a city are known to be distributed as $N(\xi_0, \sigma_0^2)$, with $(\xi_0, \sigma_0) = (100, 15)$. Suppose now that you're visiting a special school, with iqs thought to have a $N(\xi, \frac{1}{2}\sigma_0^2)$ distribution, with as yet an unknown ξ . What is your view of this distribution, when the first pupil you test has an iq of 130?

(b) There are things to think through and to learn from, by working through this very simple setup first. (i) For a single observation Y assume it comes from the $N(\xi, \sigma^2)$, and take σ as known; (ii) for the unknown mean ξ assume it comes from the prior $N(\xi_0, \tau_0^2)$, with specified prior parameters ξ_0, τ_0 . Show that this leads to a binormal joint distribution for parameter and observation,

$$\begin{pmatrix} \xi \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \xi_0 \\ \xi_0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma^2 \end{pmatrix}\right).$$

(c) Use general conditioning results from Ex. 1.44 to infer that

$$\xi | y \sim N(\xi_0 + w(y - \xi_0), w\sigma^2), \quad \text{with } w = \tau_0^2 / (\tau_0^2 + \sigma^2).$$

So w and $1 - w$ are the weights given to the data-based estimate y and the prior guess ξ_0 , respectively. Also, w is the reduction factor with which the variance of the prior-free estimator Y , from σ^2 to $w\sigma^2$. Verify also that *the precision*, the inverse variance, goes from $1/\tau_0^2$ to $1/\tau_0^2 + 1/\sigma^2$ with this extra datapoint.

precision =
inverse variance

(d) An easy but important extension of the above is to the case of a full sample $Y_1, \dots, Y_n$ from the $N(\xi, \sigma^2)$ distribution, independent given the ξ , again with σ taken known and the normal prior $N(\xi_0, \tau_0^2)$ for the unknown mean. Show that $\bar{Y} = (1/n) \sum_{i=1}^n Y_i$ is sufficient (xx xref here xx), and that

$$\begin{pmatrix} \xi \\ \bar{Y} \end{pmatrix} \sim N_2\left(\begin{pmatrix} \xi_0 \\ \xi_0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma^2/n \end{pmatrix}\right).$$

Show from this that

$$\xi | \text{data} \sim N(\xi_0 + w_n(\bar{y} - \xi_0), w_n\sigma^2/n), \quad \text{with } w_n = \frac{\tau_0^2}{\tau_0^2 + \sigma^2/n} = \frac{n\tau_0^2}{n\tau_0^2 + \sigma^2}.$$

Again, w_n is both the weight given to the data-based estimate and the factor with which the neutral estimator's variance is reduced, from σ^2/n to $w_n\sigma^2/n$. Note that $w_n \rightarrow 1$; discuss how this may be seen as 'the data wash out the prior'. In generalisation of the above, convince yourself that precision is additive: $(w_n\sigma^2/n)^{-1} = 1/\tau_0^2 + 1/(\sigma^2/n)$.

(e) Discuss the case of a ‘flat prior’, where τ is taken large, for $\xi \sim N(\xi_0, \tau^2)$. Show indeed that this leads to the intuitively plausible $\xi | \text{data} \sim N(\bar{y}, \sigma^2/n)$.

(f) In addition to having a coherent updating machine, changing the prior to the posterior, for each new data point, the Bayesian structure implies positive dependence among the observations. From $E(Y_i | \xi) = \xi$, $\text{Var}(Y_i | \xi) = \sigma^2$, $E(Y_i Y_j | \xi) = \xi^2$, show that

$$E Y_i = \xi_0, \quad \text{Var } Y_i = \sigma^2 + \tau_0^2, \quad \text{cov}(Y_i, Y_j) = \tau_0^2, \quad \text{corr}(Y_i, Y_j) = \frac{\tau_0^2}{\sigma^2 + \tau_0^2},$$

Show also that $\text{Var } \bar{Y}_n = \sigma^2/n + \tau_0^2$, and discuss what this means for large n .

(g) Prove first the convenient formula

$$v(\xi - \xi_0)^2 + n(\xi - \bar{y})^2 = (v + n)(\xi - \xi^*)^2 + d_n(\bar{y} - \xi_0)^2,$$

$$\text{where } \xi^* = \frac{v\xi_0 + n\bar{y}}{v + n} \text{ and } d_n = \frac{vn}{v + n} = (v^{-1} + n^{-1})^{-1},$$

which also may be written and interpreted via $1/d_n = 1/v + 1/n$.

(h) Show that with any prior $p(\xi)$ for ξ , the marginal density of $(Y_1, \dots, Y_n)$ can be written

$$\bar{f}(y_1, \dots, y_n) = \int (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{1}{2} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \xi)^2\right\} p(\xi) d\xi.$$

For the case of $N(\xi_0, \tau_0^2)$ worked with above, let first $Q_0 = \sum_{i=1}^n (y_i - \bar{y})^2$, and verify that $\sum_{i=1}^n (y_i - \xi)^2 = Q_0 + n(\xi - \bar{y})^2$. Writing for mathematical convenience $1/\tau^2 = v/\sigma^2$, and $1/d_n = 1/n + 1/v$, show that

$$\bar{f}(y_1, \dots, y_n) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp(-\frac{1}{2} Q_0/\sigma^2) \left(\frac{v}{v + n}\right)^{1/2} \exp\{-\frac{1}{2} d_n(\bar{y} - \xi_0)^2/\sigma^2\}.$$

(i) With the marginal density seen as a function of the two fine-tuning parameters (ξ_0, τ_0^2) , find the maximum marginal likelihood estimators $\hat{\xi}_0$ and $\hat{\sigma}$.

(j) Argue directly that $Y = (Y_1, \dots, Y_n)^t$ must be multinormal, with $Y \sim N_n(\xi_0 \mathbf{1}, \sigma^2 I_n + \tau_0^2 \mathbf{1}\mathbf{1}^t)$, with $\mathbf{1}$ the vector $(1, \dots, 1)^t$; the variance matrix has $\sigma^2 + \tau_0^2$ on the diagonal and τ_0^2 outside. Show that this agrees with the marginal density formula reached above, perhaps using Woodbury identities, see Ex. 7.18.

Ex. 7.20 *Bayes for linear regression.* Consider the linear regression setup of Ex. 3.33, with $Y \sim N_n(X\beta, \sigma^2 I_n)$, in terms of a $n \times p$ covariate matrix X , with classical frequentist theory worked out there and in Ex. 3.34. We recall that the least squares (and ML) estimator is $\hat{\beta} = (X^t X)^{-1} X^t Y = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i y_i$, with variance matrix $(\sigma^2/n) \Sigma_n^{-1}$. Here we work through Bayesian parallels, so far with σ taken known or well estimated, and with more to come in Ex. 7.24, 7.25.

(a) Let $\beta \sim N_p(\beta_0, \tau^2 K)$, i.e. with uncertainty level $\tau^2 K$ around prior guess β_0 . Show that

$$\begin{pmatrix} \beta \\ Y \end{pmatrix} \sim N_{p+n} \left(\begin{pmatrix} \beta_0 \\ X\beta_0 \end{pmatrix}, \begin{pmatrix} \tau^2 K & \tau^2 K X^t \\ \tau^2 X K & \sigma^2 I_n + \tau^2 X K X^t \end{pmatrix} \right).$$

(b) Again using general multinormal conditioning results from Ex. 1.44, show that the Bayes estimator, along with posterior variance, is

$$\begin{aligned}\hat{\beta}_B &= \beta_0 + \tau^2 K X^t (\sigma^2 I_n + \tau^2 X K X^t)^{-1} (y - X \beta_0), \\ \text{Var}(\beta | \text{data}) &= \tau^2 K - \tau^4 K X^t (\sigma^2 I_n + \tau^2 X K X^t)^{-1} X K.\end{aligned}$$

(c) These answers are satisfactory and computable in any given setup, giving Bayes estimates and credibility intervals for the β_j and for linear combinations. Matters become clearer, however, in terms of *precision matrices*, the inverse covariance matrices. Show via Woodbury identities of Ex. 7.18 that the posterior precision is $\Lambda_{\text{post}} = (1/\sigma^2) X X^t + (1/\tau^2) K^{-1}$, with posterior covariance matrix

$$\begin{aligned}V_{\text{post}} &= \Lambda_{\text{post}}^{-1} = \{(1/\sigma^2) X^t X + (1/\tau^2) K^{-1}\}^{-1}, \\ \hat{\beta}_B &= V_{\text{post}} \{(1/\sigma^2) X^t y + (1/\tau^2) K^{-1} \beta_0\}.\end{aligned}$$

(d) Verify that for τ becoming big, i.e. the prior is flattened, we have a Bayesian mirror result to the classical one: $\beta | \text{data} \sim N_p(\hat{\beta}, (\sigma^2/n) \Sigma_n^{-1})$. Check also how matters simply in the mathematically fortuitous case that the prior variance matrix K is a constant k times Σ_n^{-1} .

Ex. 7.21 *The linex loss function, normal case.* (xx nils: edit & clean, and calibrate with what is in Ch9. xx) When estimating a one-dimensional θ with a $\tilde{\theta}$, the most traditional loss function is that of squared error, $(\tilde{\theta} - \theta)^2$, which in particular is symmetric, treating over- and underestimation as equally important. A more flexible loss function is the so-called linex loss, with

$$L_c(\theta, \tilde{\theta}) = \exp\{c(\tilde{\theta} - \theta)\} - 1 - c(\tilde{\theta} - \theta).$$

The c is a fine-tuning loss parameter, for the statistician to set, balancing over- against underestimation. Note that both positive and negative values of c are allowed here. There is more on the linex loss in Ex. 7.22, 9.23, 9.24, but here we deal with the basics for a normal setup.

(a) Convince yourself that the L_c is always nonnegative. Show that $c > 0$ means penalising overestimation more than underestimation, and vice versa for $c < 0$. For small $|c|$, show that $L_c(\theta, \tilde{\theta}) \doteq \frac{1}{2} c^2 (\tilde{\theta} - \theta)^2$, getting back to squared error loss. The constant in front is immaterial for evaluating and comparing loss and risk, and one may use $L_c^*(\theta, \tilde{\theta}) = L_c(\theta, \tilde{\theta}) / (\frac{1}{2} c^2)$ to have a smoother transition to the $c = 0$ case of squared error loss.

(b) We start with the simple stylised setting where $Y \sim N(\theta, 1)$. Show that the risk function for $\hat{\theta} = Y + d$ is

$$R(\theta, Y + d) = E_\theta [\exp\{c(Y + d - \theta)\} - 1 - c(Y + d - \theta)] = \exp(cd + \frac{1}{2}c^2) - 1 - cd.$$

Minimise over d to find that $\hat{\theta} = Y - \frac{1}{2}c$ is the best estimator of this type, with constant risk $\frac{1}{2}c^2$.

(c) Then push Bayes to the table, with prior $\theta \sim N(0, \tau^2)$. From Ex. 7.19, verify that $\theta | y \sim N(\rho y, \rho)$, with $\rho = \tau^2 / (\tau^2 + 1)$. Writing $\theta = \rho y + \rho^{1/2} N$, for a standard normal N show that the expected loss given data can be expressed as

$$E \{L_c(\theta, t) | \text{data}\} = \exp(c(t - \rho y)) \exp(\frac{1}{2}c^2 \rho^2) - 1 - c(t - \rho y).$$

Minimise over t to exhibit the Bayes estimator $\hat{\theta}_B = \rho Y - \frac{1}{2}c\rho$.

it is also
minimax, see
Ex. 9.23

(d) Generalise the above to the case of $Y \sim N(\theta, \sigma^2)$, for some σ . Show that the best estimator of the type $Y + d$ is $\hat{\theta} = Y - \frac{1}{2}c\sigma^2$. Work with the prior $N(0, \tau^2)$ for θ and show that $\theta | y \sim N(\rho y, \rho\sigma^2)$, with $\rho = \tau^2/(\tau^2 + \sigma^2)$. Derive next the Bayes estimator $\hat{\theta}_B = y - \frac{1}{2}c\sigma^2\rho^2$, and verify that the minimum Bayes risk becomes $\frac{1}{2}c^2\rho^2\sigma^2$. If $\bar{Y}$ is the sample mean of a dataset of size n , with $N(\theta, \sigma^2)$ observations, explain (i) that the best frequentist estimator becomes $\bar{Y} - \frac{1}{2}c\sigma^2/n$, with constant risk $\frac{1}{2}c^2\sigma^2/n$; and (ii) that the Bayes estimator with the $N(0, \tau^2)$ prior becomes $\hat{\theta}_B = \rho_n \bar{Y} - \frac{1}{2}c^2\rho_n^2\sigma^2/n$, now with $\rho_n = \tau^2/(\tau^2 + \sigma^2/n)$.

Ex. 7.22 *The linex loss function, general case.* Consider again the linex loss function $L_c(\theta, \tilde{\theta})$ worked with in Ex. 7.21. Here we work with the general case, i.e. outside normal data with normal priors.

(a) Show that the expected loss given data can be expressed as

$$\begin{aligned} E\{L_c(\theta, t) | \text{data}\} &= E[\exp\{c(t - \theta)\} - 1 - c(t - \theta) | \text{data}] \\ &= \exp(ct)M(-c) - 1 - c(t - \hat{\xi}), \end{aligned}$$

where $\hat{\xi}$ is the posterior mean and $M(-c) = E\{\exp(-c\theta) | \text{data}\}$, the m.g.f. of θ given data, computed at $-c$.

(b) Show that this is minimised for the t_0 where $\exp(ct_0)M(-c) = 1$, or $ct_0 + \log M(-c) = 0$, so that the Bayes estimator becomes $\hat{\theta}_B = -(1/c) \log M(-c)$. This may be computed numerically, perhaps by simulation, in cases where no clear formula exists for $M(-c)$. Show also that the expected posterior loss, using the Bayes solution, is

$$\min E\{L_c(\theta, t) | \text{data}\} = -c(t_0 - \hat{\xi}) = \log M(-c) + c\hat{\xi}.$$

(c) Using approximations for m.g.f.s close to zero, show that $M(-c) \doteq 1 - c\hat{\xi} + \frac{1}{2}c^2(\hat{\xi}^2 + \hat{\sigma}^2)$, with $\hat{\xi}$ and $\hat{\sigma}^2$ the posterior mean and variance. Deduce that $\hat{\theta}_B \doteq \hat{\xi} - \frac{1}{2}c\hat{\sigma}^2$.

(d) In situations where the posterior is based on a sample of size n , the posterior mean $\hat{\xi}_n$ stays stable whereas the posterior variance $\hat{\sigma}^2$ goes down with speed $1/n$, i.e. as $\hat{\sigma}_0^2/n$, for the relevant $\hat{\sigma}_0^2$. In such cases, $\hat{\xi}_n - \frac{1}{2}c\hat{\sigma}_0^2/n$ becomes the approximation to the Bayes linex estimator $\hat{\theta}_B$.

Ex. 7.23 *When underestimation doesn't matter so much.* Suppose θ is to be estimated, where underestimation is less important than overestimation. One version of this is to use the loss function $L(\theta, a) = L_0(a - \theta)$, with $L_0(u) = u^2$ for $u \geq -c$, but then $L_0(u) = c^2$, for all $u \leq -c$, for a relevant negative threshold $-c$. Thus a big error to the left doesn't matter more than a moderate error to the left.

(a) For any random θ , with density f , c.d.f. F , $E_1(x) = \int_{-\infty}^x \theta f(\theta) d\theta$, $E_2(x) = \int_{-\infty}^x \theta^2 f(\theta) d\theta$, show that

$$\begin{aligned} R(a) = E L_0(a - \theta) &= \int_{-\infty}^{a+c} (a - \theta)^2 f(\theta) d\theta + c^2 \int_{a+c}^{\infty} f(\theta) d\theta \\ &= a^2 F(a+c) - 2a E_1(a+c) + E_2(a+c) + c^2(1 - F(a+c)), \end{aligned}$$

and that its derivative simplifies to

$$R'(a) = 2aF(a+c) - 2E_1(a+c).$$

Check that the solution a_0 to setting this to zero is equal to the mean of θ in the case of $-c$ being far to the left.

(b) Explain that the Bayes estimator, for a given dataset, is the solution $\hat{a}$ to the equation $aF(a + c | \text{data}) = E_1(a + c | \text{data})$, featuring the posterior c.d.f. $\Pr(\theta \leq a + c | \text{data})$ and $E_1(a + c | \text{data}) = \int_{-\infty}^{a+c} v\pi(v | \text{data}) dv$.

(c) For the case of $y_1, \dots, y_n$ i.i.d. from the $N(\xi, 1)$, with a flat prior for ξ , show first that $\xi | \text{data} \sim N(\bar{y}, 1/n)$, which may be represented as $\xi = \bar{y} + N/\sqrt{n}$, with a standard normal N . Show then that the equation for finding the Bayes estimator is

$$(\hat{a} - \bar{y})\Phi(\sqrt{n}(\hat{a} - \bar{y} + c)) = -\phi(\sqrt{n}(\hat{a} - \bar{y} + c))/\sqrt{n}.$$

Explain that this means that the Bayes estimator pushes the ML estimator $\bar{y}$ to the left, with an amount diminishing with sample size n .

(d) Invent an illustration, ideally with real data.

Ex. 7.24 *The gamma-normal prior and posterior.* Let data $y_1, \dots, y_n$ for given parameters ξ and σ be i.i.d. $N(\xi, \sigma^2)$. We have seen in Ex. 7.19 that when σ may be taken as a known quantity, the canonical class of priors for ξ is the normal one. When both parameters are unknown, however, as in most practical encounters, a more elaborate analysis is called for.

(a) Show that the likelihood function may be written as being proportional to

$$L_n(\xi, \sigma) = \exp\left[-n \log \sigma - \frac{1}{2} \frac{1}{\sigma^2} \{Q_0 + n(\xi - \bar{y})^2\}\right],$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $Q_0 = \sum_{i=1}^n (y_i - \bar{y})^2$.

(b) With *any* given prior $p(\xi, \sigma)$, explain how you may set up a Metropolis type MCMC to draw samples from the posterior distribution. Try this out in practice, using the prior that takes ξ and $\log \sigma$ independent and uniform on say $[-5, 5]$ and $[-10, 10]$, with data that you simulate for the occasion from a $N(2.345, 1.234^2)$, with $n = 25$. Note that this approach does not need more mathematical algebra as such, apart from the likelihood function above (along with basic programming skills).

(c) There is however a popular and convenient conjugate class of priors for which posterior distributions become particularly clear, with the appropriate algebraic efforts. These in particular involve placing a Gamma prior on the inverse variance $\lambda = 1/\sigma^2$. Say that (λ, ξ) has the gamma-normal distribution with parameters (a, b, ξ_0, v) , and write this as $(\lambda, \xi) \sim \text{GN}(a, b, \xi_0, v)$, provided $\lambda = 1/\sigma^2 \sim \text{Gam}(a, b)$ and $\xi | \sigma \sim N(\xi_0, \sigma^2/v)$. Show that the prior can be expressed as

$$p(\lambda, \xi) \propto \lambda^{a-1} \lambda^{1/2} \exp\left[-\lambda\{b + \frac{1}{2}v(\xi - \xi_0)^2\}\right].$$

What is the unconditional prior variance of ξ ?

(d) Using the identity from Ex. 7.19(g), show that if the prior is $(\lambda, \xi) \sim \text{GN}(a, b, \xi_0, v)$, then

$$(\lambda, \xi) | \text{data} \sim \text{GN}(a + \frac{1}{2}n, b + \frac{1}{2}Q_0 + \frac{1}{2}d_n(\bar{y} - \xi_0)^2, \xi^*, v + n).$$

(e) The special case of a ‘flat prior’ for ξ , corresponding to letting $v \rightarrow 0$ above, is particularly easy to deal with. Show that then

$$(\lambda, \xi) | \text{data} \sim \text{GN}(a + \frac{1}{2}n, b + \frac{1}{2}Q_0, \bar{y}, n).$$

Find the posterior mean of σ^2 under this prior.

(f) For an illustration, consider the cigarette consumption data $x_1, \dots, x_n$, see Story #17, for $n = 44$ US states (actually, 43 states plus the District of Columbia), from the early 1960ies, taken here to form a sample from $N(\xi, \sigma^2)$. The x_i is the consumption in state i , in hundreds per year, ranging then from 14.0 to 42.4 (which translates to from 3.84 per day to 11.61 per day, in the adult population). Fix the prior that takes (i) $\lambda = 1/\sigma^2 \sim \text{Gam}(a, b)$, with (a, b) taken to correspond to 0.10 and 0.90 prior quantiles for σ being 1.0 and 10.0; (ii) given σ , ξ is a $N(20.0, \sigma^2/v)$ with $v = 0.50$. Find the precise posterior distribution for (λ, ξ) . Give posterior means and 90 percent credibility intervals for ξ , σ , and for the probability $p = \Pr(X \geq 40)$.

(g) For the same data and setup as in the previous point, carry out a second Bayesian analysis, with the simpler prior that corresponds to $v \rightarrow 0$. Comment briefly on the differences in results.

Ex. 7.25 *The gamma-multinormal prior for linear regression models.* The aim of the present exercise is to generalise the gamma-normal conjugate prior class above to the linear-normal regression model. The model is the very classical one of Ex. 3.33, where

$$y_i = x_{i,1}\beta_1 + \dots + x_{i,k}\beta_k + \varepsilon_i = x_i^t \beta + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

with the ε_i taken i.i.d. $N(0, \sigma^2)$. Write X for the $n \times k$ matrix of covariates (explanatory variables), with $x_i = (x_{i,1}, \dots, x_{i,k})$ as its i th row, and use y and ε to indicate the vectors of y_i and ε_i . Then $y = X\beta + \varepsilon \sim N_n(X\beta, \sigma^2 I_n)$ is a concise way to write the full model.

(a) Show that the likelihood function may be written as being proportional to

$$L_n(\beta, \sigma) = \sigma^{-n} \exp \left[-\frac{1}{2} \frac{1}{\sigma^2} \{Q_0 + n(\beta - \hat{\beta})^t M_n(\beta - \hat{\beta})\} \right],$$

in which

$$\Sigma_n = (1/n)X^t X = n^{-1} \sum_{i=1}^n x_i x_i^t \quad \text{and} \quad \hat{\beta} = (X^t X)^{-1} X^t y = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i y_i.$$

Also,

$$Q(\beta) = \|y - X\beta\|^2 = Q_0 + n(\beta - \hat{\beta})^t \Sigma_n(\beta - \hat{\beta}),$$

with $Q_0 = \sum_{i=1}^n (y_i - x_i^t \hat{\beta})^2$ the minimum value of Q over all β . Note that $\hat{\beta}$ is the classical least squares estimator (and the ML estimator), which in the frequentist framework is unbiased with variance matrix equal to $\sigma^2(X^t X)^{-1} = (\sigma^2/n)\Sigma_n^{-1}$. This is the basis of all classical methods related to the widely popular linear regression model.

(b) Let $p(\beta, \sigma)$ be *any* prior for the $(k+1)$ -dimensional parameter of the model. Set up formulae for a Metropolis type MCMC algorithm for drawing samples from the posterior distribution of (β, σ) .

(c) In spite of the possibility of solving problems via MCMC (or perhaps acceptance-rejection sampling), as with the previous exercise it is very much worthwhile setting up explicit formulae for the case of a certain canonical prior class. Write $(\lambda, \beta) \sim \text{GN}_k(a, b, \beta_0, \Sigma_0)$ to indicate the gamma-normal prior where

$$\lambda = 1/\sigma^2 \sim \text{Gam}(a, b) \quad \text{and} \quad \beta | \sigma \sim N_k(\beta_0, \sigma^2 \Sigma_0^{-1}).$$

Show that this prior may be expressed as

$$p(\lambda, \beta) \propto \lambda^{a-1} \lambda^{k/2} \exp \left[-\lambda \left\{ b + \frac{1}{2} (\beta - \beta_0)^t \Sigma_0 (\beta - \beta_0) \right\} \right].$$

(d) When multiplying the prior with the likelihood it is convenient to use the following linear algebra identity about quadratic forms, which you should prove first. For symmetric and invertible matrices A and B , and for any vectors a, b, x of the appropriate dimension,

$$(x - a)^t A (x - a) + (x - b)^t B (x - b) = (x - \xi)^t (A + B) (x - \xi) + (b - a)^t D (b - a),$$

where $\xi = (A + B)^{-1}(Aa + Bb)$ (a weighted average of a and b) and D is a matrix for which several equivalent formulae may be used:

$$\begin{aligned} D &= A(A + B)^{-1}B = B(A + B)^{-1}A \\ &= A - A(A + B)^{-1}A = B - B(A + B)^{-1}B = (A^{-1} + B^{-1})^{-1}. \end{aligned}$$

(e) Prove that if (λ, β) has the $\text{GN}_k(a, b, \beta_0, \Sigma_0)$ prior, then

$$(\lambda, \beta) \mid \text{data} \sim \text{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0 + \tfrac{1}{2}(\hat{\beta} - \beta_0)^t D_n (\hat{\beta} - \beta_0), \beta^*, \Sigma_0 + n\Sigma_n),$$

where

$$\beta^* = (\Sigma_0 + n\Sigma_n)^{-1}(\Sigma_0\beta_0 + n\Sigma_n\hat{\beta}) \quad \text{and} \quad D_n = \Sigma_0(\Sigma_0 + n\Sigma_n)^{-1}n\Sigma_n.$$

This characterisation makes it easy to simulate a large number of (β, σ) from the posterior distribution and hence to carry out Bayesian inference for any parameter of quantity of interest.

(f) Note the algebraic simplifications that result when the Σ_0 in the prior is chosen as being proportional to the covariate sample variance matrix, i.e. $\Sigma_0 = c_0\Sigma_n$. Show that then

$$\beta^* = (c_0\beta_0 + n\hat{\beta})/(c_0 + n) \quad \text{and} \quad D_n = c_0n/(c_0 + n).$$

In this connection c_0 has a natural interpretation as ‘prior sample size’.

(g) A special case of the above, leading to simpler results, is that where β has a flat, non-informative prior, corresponding to very large prior variances, i.e. to $\Sigma_0 \rightarrow 0$. Show that with such a prior,

$$(\lambda, \beta) \mid \text{data} \sim \text{GN}_k(a + \tfrac{1}{2}n, b + \tfrac{1}{2}Q_0, \hat{\beta}, n\Sigma_n).$$

The prior is improper (infinite integral), but the posterior is proper as long as $\hat{\beta}$ exists, which requires $X^t X$ to have full rank, which again means at least k linearly independent covariate vectors, and, in particular, $n \geq k$.

(h) Go again to the dataset worked with in Story #17, for illustration and for flexing your operational muscles. For y use the bladder cancer column of deaths per 100,000 inhabitants and for x use the number of cigarettes sold per capita. Your task is to carry out Bayesian analysis within the linear regression model $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ for $i = 1, \dots, 44$, with ε_i taken i.i.d. $N(0, \sigma^2)$. Specifically, we wish point estimates along with 95 percent credibility intervals for (i) each of the three parameters β_0, β_1, σ ; (ii) the probability that $y \geq 25.0$, for a state with cigarette consumption $x = 35.0$; (iii) the bladder cancer death rates y_{45} and y_{46} , per 100,000 inhabitants, for states with cigarette consumption rates $x_{45} = 10.0$ (low) and $x_{46} = 50.0$ (high). You are to carry out such inference with two priors (xx nils checks details xx): (1) First, the informative one which

takes $1/\sigma^2$ a gamma with 0.10 and 0.90 quantiles for σ equal to 1.0 and 5.0, and β_0 and β_1 as independent normals $(5.0, 0.5\sigma^2)$ and $(0.0, (2.0\sigma)^2)$, given σ . (2) Then, the simpler and partly non-informative one that takes a flat prior for (β_0, β_1) and the less informative one for σ that uses 0.10 and 0.90 prior quantiles 0.5 and 10.0. Finally, compare your results from those arrived at using classical frequentist methods.

Ex. 7.26 Mixture priors. Suppose data Y come from a model $f(y, \theta)$, where different priors $\pi_1(\theta), \dots, \pi_k(\theta)$ can be used, each leading to posterior distributions $\pi_1(\theta|y), \dots, \pi_k(\theta|y)$.

(a) For each of these possible priors (and hence possible posteriors), show that there is a representation $f_j(y, \theta) = \pi_j(\theta)f(y, \theta) = \pi_j(\theta|y)\bar{f}_j(y)$, where $\bar{f}_j(y) = \int f(y, \theta)\pi_j(\theta) d\theta$ is the marginal density of Y , associated with the $\pi_j(\theta)$ prior.

(b) Suppose now that a full mixture prior is assigned to θ , of the type $\pi(\theta) = p_1\pi_1(\theta) + \dots + p_k\pi_k(\theta)$, with probabilities $p_1, \dots, p_k$ summing to 1. Explain that this can be interpreted as θ being drawn from prior j with probability p_j . Show also that the marginal distribution of Y can be expressed as $\bar{f}(y) = \sum_{j=1}^k p_j\bar{f}_j(y)$.

(c) Then show that the posterior distribution for θ becomes

$$\pi(\theta|y) = p_1^*\pi_1(\theta|y) + \dots + p_k^*\pi_k(\theta|y),$$

with revised prior probabilities $p_j^* = p_j\bar{f}_j(y)/\sum_{j'=1}^k p_{j'}\bar{f}_{j'}(y)$ for the different types of priors.

(d) Suppose $Y \sim \text{binom}(n, \theta)$, and that the prior used for θ is $0.15\text{Beta}(2, 10) + 0.70\text{Beta}(15, 15) + 0.15\text{Beta}(10, 2)$. Draw this prior in a plot. With $n = 100$, compute the posterior probabilities p_1^*, p_2^*, p_3^* , and draw the posterior distribution for θ , along with the prior, for each of the cases $y = 12$, $y = 48$, $y = 91$.

(e) Revisit the Ex. 7.13 setup, with throws of a die. Now do a mixture prior, of the type $0.50\text{Dir}(1, 1, 1, 1, 1, 1) + 0.50\text{Dir}(s, s, s, s, s, s)$, where s is quite big, reflecting the possibility that the die is perfectly fair with probabilities equal to or very close to $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$.

(f) Generalise the above to the situation where $\pi(\theta) = \int \pi_\alpha(\theta) dG(\alpha)$ is a mixture of $\pi_\alpha(\theta)$ priors, with $dG(\alpha)$ a fully general probability measure over the space of hyperparameter α . The α could be a parameter belonging to a finite set, matching the setup above, or a full continuous mixture. Show that the posterior can be represented as $\pi(\theta|y) = \int \pi_\alpha(\theta|y) dG(\alpha|\text{data})$, where $dG(\alpha|\text{data})$ is the posterior for the hyperparameter, and $\pi_\alpha(\theta|y)$ is the posterior for θ in the setup where α is fixed and known.

Ex. 7.27 de Finetti (and probability probably doesn't exist). Why believe in i.i.d. setups, and why be Bayesian? There are mathematical arguments from a Start Assumption that somehow lead to both these ingredients. Say that a random sequence $(X_1, X_2, \dots)$ is *exchangeable* if, for any finite subset of length k , the distribution of $(X_{i_1}, \dots, X_{i_k})$ stays the same (i.e. regardless of *which* indices are at work, as long as there are k of them). So (X_2, X_7, X_{13}) has the same distribution as (X_{14}, X_8, X_5) , etc.

(a) Consider any density model $f(x, \theta)$ for outcomes x , with θ in a parameter space Θ . Suppose θ is drawn from some π , and that the X_i are i.i.d. $f(x, \theta)$ given that parameter value. Show that, marginally, the density is $f(x_1, \dots, x_n) = \int \{\prod_{i=1}^n f(x_i, \theta)\} d\pi(\theta)$, and that the X_i sequence is exchangeable. Show also, incidentally, that if $T = (T_1, \dots, T_p)$ is sufficient for θ , the the marginal density of $X_1, \dots, X_n$ depends only on T .

(b) Suppose the X_i are i.i.d. Bernoulli variables with probability θ , but that θ has been picked from a background $\text{Beta}(a, b)$. The sequence is hence exchangeable; identify also the distribution of $(X_1, \dots, X_n)$, and show that it only depends on $S_n = \sum_{i=1}^n X_i$. Also, show that

$$\Pr(S_n = s) = \binom{n}{s} \frac{\Gamma(a+b) \Gamma(a+s) \Gamma(b+n-s)}{\Gamma(a) \Gamma(b) \Gamma(a+b+n)} \quad \text{for } s = 0, 1, \dots, n,$$

and that the predictive depends only on the count S_n , with $\Pr(X_{n+1} = 1 | x_1, \dots, x_n) = (a + S_n)/(a + b + n)$.

(c) Then for a partial converse: if the predictive has the above $(a + S_n)/(a + b + n)$ linear form, there *must* be Beta-mixing, i.e. a prior chose the θ and the flips must be i.i.d. given that θ . Show indeed, by induction, that

$$\Pr(X_1 = x_1, \dots, X_n = x_n) = \int_0^1 \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \text{be}(\theta, a, b) d\theta.$$

So if you accept the linear predictive for coin tossing, you *must* accept (i) they are i.i.d. coin tosses given some θ , and (ii) there is a Bayesian prior for that θ . See also Notes & Pointers for more on *de Finetti theorems* generalising this; exchangeability implies this structure.

(d) Let $X_1, X_2, \dots$ be an exchangeable sequence of random variables, and consider the one-step ahead predictive distribution $\alpha_n(A) = \Pr(X_{n+1} \in A | X_1, \dots, X_n)$, where $A \in \mathcal{B}(\mathbb{R})$. Show that for each A , the predictives $\alpha_n(A)$ are such that

$$\mathbb{E}\{\alpha_{n+1}(A) | X_1, \dots, X_n\} = \alpha_n(A), \quad \text{a.s.}, \quad (7.5)$$

that is, they form a martingale with respect to filtration generated by $X_1, X_2, \dots$

(e) Suppose that the predictives $(\alpha_n)_{n \geq 1}$ forms a martingale, i.e., they satisfy (7.5). Show that it then must be the case that for each event A , and any $k \geq 1$

$$\Pr(X_{n+k} \in A | X_1, \dots, X_n) = \Pr(X_{n+1} \in A | X_1, \dots, X_n),$$

almost surely. A sequence of random variables with this property is said to be *conditionally identically distributed*. Show that the converse is also true, namely that if $X_1, X_2, \dots$ are conditionally identically distributed, then their one-step ahead predictives $\Pr(X_{n+1} \in A | X_1, \dots, X_n)$ form a martingale, in the sense of (7.5).

The Jeffreys and the vip-Jeffreys priors

Ex. 7.28 *The Jeffreys prior: the basics.* Consider a smooth parametric model $f(y, \theta)$, with score function $u(y, \theta) = \partial \log f(y, \theta) / \partial \theta$ and finite Fisher information matrix $J(\theta) = \text{Var}_\theta u(Y, \theta)$; see Ex. 5.6. The *Jeffreys prior* is defined as $\pi_{\text{jeff}}(\theta) \propto |J(\theta)|^{1/2}$. If this has a finite integral, over the parameter domain, the proportionality constant is fixed to make the integral 1, but we explicitly allow also improper priors here, i.e. positive functions with infinite integral. It will be seen that in most cases, the passing from $\pi_{\text{jeff}}(\theta)$ to $\pi_{\text{jeff}}(\theta | \text{data})$ via the Bayes theorem (7.2) gives a proper posterior.

(a) Consider a smooth 1-1 reparametrisation, from $\theta = (\theta_1, \dots, \theta_p)$ to $\phi = (\phi_1, \dots, \phi_p)$, with $\theta = h(\phi)$ and $\phi = h^{-1}(\theta)$. Use Ex. 1.12 to show that if θ has the Jeffreys prior, then ϕ has density $\pi^*(\phi) = \pi_{\text{jeff}}(h(\phi)) |G(\phi)|$ with the gradient matrix $G(\phi) = \partial h(\phi) / \partial \phi$.

the sun has
risen n days in
a row; it will
rise tomorrow
too, with
probability
 $(n+1)/(n+2)$

(b) We have seen in Ex. 5.9 that the Fisher information matrix on the ϕ scale is $J^*(\phi) = G(\phi)J(h(\phi))G(\phi)^t$. Deduce that the Jeffreys prior constructed on the ϕ scale becomes $\pi_{\text{jeff}}^*(\phi) = \pi_{\text{jeff}}(h(\phi))|G(\phi)|$. But this is the same as $\pi^*(\phi)$ above. Explain that this demonstrates that the Jeffreys prior is invariant under smooth parameter transformations. Three Bayesian statisticians, with the same data and the same model, will reach the same conclusions, even if they use three different parametrisations, as long as they use the Jeffreys prior construction in their respective model formulations.

invariance
w.r.t. parametrisation

Ex. 7.29 *The Jeffreys prior for normal models.* We go through the Fisher information and Jeffreys calculus for normal models.

(a) For the normal $N(\xi, \sigma^2)$ case, with σ known, show that the Jeffreys prior for ξ is flat on the full line. With data $y_1, \dots, y_n$, show that the Bayes rule gives the nice and interpretable posterior $\xi | \text{data} \sim N(\bar{y}, \sigma^2/n)$; in particular, even for $n = 1$ the posterior is proper though the prior isn't. Generalise this to the linear regression model, with $y_i \sim N(x_i^t \beta, \sigma^2)$, with x_i of dimension p , and again with σ known. Show that Jeffreys says that the β prior is flat on the full $\mathbb{R}^p$. Find the posterior for β , and explain that this is proper as long as there are at least p linearly independent x_i in the data; in particular, one needs $n \geq p$.

(b) Matters are more delicate for the normal models when σ is unknown. For the single $N(\xi, \sigma^2)$ model, show that $\pi_{\text{jeff}}(\xi, \sigma) \propto 1/\sigma^2$, i.e. flat for ξ but $1/\sigma^2$ for σ . Explain that the posterior needs at least two datapoints in order to become proper. For the linear regression setup, show that $\pi_{\text{jeff}}(\beta, \sigma) \propto 1/\sigma^{p+1}$. Connect this to the gamma-normal setup of Ex. 7.25, and identify the proper posterior $\pi(\beta, \sigma | \text{data})$.

Ex. 7.30 *The Jeffreys prior in certain models.* Also in various other models we may derive clear formula for the Fisher information matrix, and hence also for the Jeffreys prior $|J(\theta)|^{1/2}$.

(a) For the binomial (n, p) model, show that the Jeffreys prior is the $\text{Beta}(\frac{1}{2}, \frac{1}{2})$. Compare this with the geometric, with $f(y, p) = (1-p)^y p$, for which you should find $\pi_{\text{jeff}}(p) \propto p^{-1}(1-p)^{-1/2}$, which can be thought of as a $\text{Beta}(\varepsilon, \frac{1}{2})$ with a very small ε . Show that this prior is improper, but that the posterior becomes proper as long as there are $n \geq 1$ observations.

(b) Then consider the trinomial (X, Y, Z) , with probabilities proportional to $p^x q^y (1-p-q)^z$, with $x+y+z = n$. Find the 2×2 Fisher information matrix and its determinant, and show that the Jeffreys prior is proportional to $p^{-1/2} q^{-1/2} (1-p-q)^{-1/2}$, which is the Dirichlet $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ for (p, q, r) . Generalise to the multinomial $(n, p_1, \dots, p_k)$ case, with $p_1 + \dots + p_k = 1$.

(c) For the Poisson with parameter θ , show that the Jeffreys prior is $1/\theta^{1/2}$.

(d) For the Gamma (a, b) , show that the Jeffreys prior takes the form $\pi(a, b) = \pi_1(a)\pi_2(b)$, with $\pi_1(a) \propto \{a\psi'(a) - 1\}^{1/2}$ and $\pi_2(b) \propto 1/b$.

(e) Then work through the details for the Weibull model, with c.d.f. $1 - \exp\{-(t/a)^b\}$, see Ex. 5.39. Via the Fisher matrix found there, show that $\pi_{\text{jeff}}(a, b) \propto 1/a$, i.e. flat in b .

Ex. 7.31 *A simple model for deviations from uniformity.* This exercise illustrates how we can carry out Bayesian analysis for almost any given one-parameter model, via simple numerical techniques; bigger models need bigger tools, e.g. the MCMC ones of Ex. 7.17, if needed coupled with numerical evaluation of the Fisher matrix. Consider then the model $f(y, \theta) = 1 + \theta(y - \frac{1}{2})$ for $y \in [0, 1]$.

(a) Show that this indeed defines a bona fide model, for $\theta \in [-2, 2]$, and with c.d.f. $F(y, \theta) = y + \frac{1}{2}\theta(y^2 - y)$. Show that the Fisher information is

$$J(\theta) = \int_{-1/2}^{1/2} \frac{x^2}{1 + \theta x} dx = \left(\log \frac{1 + \frac{1}{2}\theta}{1 - \frac{1}{2}\theta} - \theta \right) / \theta^3.$$

Compute and display the Jeffreys prior.

(b) Take $\theta_{\text{true}} = 0.333$, simulate say $n = 100$ points from the model, and give a graph for the log-likelihood function. Compute the ML and an approximate 90 percent interval for θ via the methods of Ch. 5. Then, with a uniform prior on $[-2, 2]$, compute and display the posterior distribution for θ .

(c) Using a fine grid, e.g. with grid length 0.0001, sample say 10^5 points from the posterior distribution. From these, provide 0.05, 0.50, 0.95 quantiles. (xx just a bit more; round off; point away. xx)

Ex. 7.32 *The vip-Jeffreys combo prior.* Bayesian analysis for a model requires full specification of a prior for its full parameter vector θ , of dimension say p . In many situations there is prior information available regarding a sub-parameter, with a separately elicited prior $\pi_{\text{vip}}(\mu)$ for that perhaps very important parameter, but otherwise not for other aspects of the full model parameter. Parametrising θ into (μ, γ) , with ensuing Fisher information matrix, we propose using

$$\pi_{\text{combo}}(\mu, \gamma) = \pi_{\text{vip}}(\mu)\pi_{\text{jeff}}(\gamma | \mu) = \pi_{\text{vip}}(\mu)|J(\gamma | \mu)|^{1/2}. \quad (7.6)$$

With the full $p \times p$ Fisher information matrix $J(\mu, \gamma)$, the $J(\gamma | \mu)$ is the $(p-1) \times (p-1)$ lower right corner matrix, which is also the Fisher matrix for the model $f(y, \mu, \gamma)$ when μ is held fixed.

(a) Suppose X and Y are independent $N(\xi_1, \sigma^2)$ and $N(\xi_2, \sigma^2)$, with known σ . If there is a vip-prior $\pi(\delta)$ for $\delta = \xi_2 - \xi_1$, what is the vip-combo prior? Generalise to the setup with $X_i \sim N(\xi_1, \sigma^2)$ and $Y_j \sim N(\xi_2, \sigma^2)$, for $i = 1, \dots, n$ and $j = 1, \dots, m$. Again, with known σ and a vip-prior for δ , find the vip-combo. Then with unknown σ , find the vip-combo priors for the cases of start prior for (i) δ and (ii) δ/σ .

(b) Suppose X_i and Y_j are independent exponentials with parameters θ_1 and θ_2 , for $i = 1, \dots, n$ and $j = 1, \dots, m$. Find first the Jeffreys prior, and then the vip-combo prior when there is a start prior for $\rho = \theta_2/\theta_1$. Work out expressions for $\pi(\rho | \text{data})$.

(c) Consider the Gamma (a, b) model, with vip prior $\pi_{\text{vip}}(\xi)$ for the mean $\xi = a/b$. Parametrise the model as $(c\xi, c)$, and deduce first that the Jeffreys prior for c , for given ξ , is

$$\text{jeff}(c | \xi) \propto \{\xi^2 \psi'(c\xi) - \xi/c\}^{1/2}.$$

Then exhibit the vip-combo prior for (a, b) .

(d) For the $N(\xi, \sigma^2)$ model, find the combo priors $\pi_{\text{vip}}(\xi)\pi_{\text{jeff}}(\sigma | \xi)$ and $\pi_{\text{vip}}(\sigma)\pi_{\text{jeff}}(\xi | \sigma)$, for the cases of having start priors on ξ and σ , respectively.

Ex. 7.33 *More vip-combo.* Here we pursue the vip-combo constructions in a few more situations.

(a) Consider independent Poisson pairs (X_i, Y_i) for $i = 1, \dots, m$, with parameters $(\theta_i, \theta_i\gamma)$. Suppose there is a $\pi_{\text{vip}}(\gamma)$. Show that the combo prior then takes the form

$$\pi_{\text{vip}}(\gamma)(1 + \gamma)^{m/2} \prod_{i=1}^m 1/\theta_i^{1/2}.$$

the vip-Jeffreys
combo

Find an expression for $\pi_{\text{combo}}(\gamma | \text{data})$.

(b) You stare at your die for a minute, before you toss it 25 times. You think you know something about p_6 , the probability of having ‘6’, formalised as a clear $\pi_{\text{vip}}(p_6)$. What is then your combo prior for the full $(p_1, \dots, p_6)$? How can you compute the posterior $\pi(p_6 | \text{data})$?

(c) Consider the binormal model, with information on the correlation ρ . Show first that the Jeffreys in the five-parameter model is the square-root of

$$|J(\theta)| = 8/\{(1 - \rho^2)^3 \sigma_1^4 \sigma_2^4\}.$$

Then, examining the four-parameter model for given ρ , do the calculus and algebra to show that $J(\xi_1, \xi_2, \sigma_1, \sigma_2 | \rho)$ matrix has blocks

$$\frac{1}{1 - \rho^2} \begin{pmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1 \sigma_2) \\ -\rho/(\sigma_1 \sigma_2) & 1/\sigma_2^2 \end{pmatrix} \quad \text{and} \quad 2 \begin{pmatrix} 1/\sigma_1^2 & \rho/(\sigma_1 \sigma_2) \\ \rho/(\sigma_1 \sigma_2) & 1/\sigma_2^2 \end{pmatrix},$$

with zero correlations between these. Deduce from this that the vip-Jeffreys combo prior for the five parameters becomes

$$\pi_{\text{combo}} = \pi_{\text{vip}}(\rho) \frac{(1 + \rho^2)^{1/2}}{1 - \rho^2} \frac{1}{\sigma_1^2 \sigma_2^2}.$$

For which $\pi_{\text{vip}}(\rho)$ does this match the non-informative Jeffreys prior? For all other choices, the vip-combo is different, and indeed more informative.

(d) Look at linear regression $N(a+bx, \sigma^2)$, with information on b . Show first that Jeffreys is $1/\sigma^3$, and then that the vip-Jeffreys is $\pi_{\text{vip}}(b)/\sigma^2$. How can you find $\pi(b | \text{data})$?

(e) Let X and Y be independent normals $N(a, 1)$ and $N(b, 1)$. Reaching clear inference for $\gamma = a/b$ is not easy since that parameter is messed up for b close to zero. With the Jeffreys flat prior for (a, b) , show that $a | \text{data} \sim N(x, 1)$ and $b | \text{data} \sim N(y, 1)$. The posterior for γ is awkward, particularly for y small; try some values of (x, y) and simulate to see this. Suppose however that there is a clear prior $\pi_{\text{vip}}(\gamma)$. Using the principle above, study for fixed γ the one-parameter density $g(x, y, b\gamma, b)$, with log-density $-\frac{1}{2}(x - b\gamma)^2 - \frac{1}{2}(y - b^2)$. Show that the combo prior becomes $\pi_{\text{combo}}(\gamma, b) = \pi_{\text{vip}}(\gamma)(1 + \gamma^2)^{1/2}$; in particular, it's flat in b . Then work out the posterior for γ :

$$\begin{aligned} \pi_{\text{combo}}(\gamma | x, y) &\propto \pi_{\text{vip}}(\gamma)(1 + \gamma^2)^{1/2} \int g(x, y, b\gamma, \gamma) db \\ &= \pi_{\text{vip}}(\gamma) \exp\left\{\frac{1}{2} \frac{(\gamma x + y)^2}{1 + \gamma^2}\right\}. \end{aligned}$$

So with prior information on the ratio we have a much sharper Bayesian solution to this so-called Fieller problem; see Ex. 8.30 for the frequentist take.

Ex. 7.34 *Bayesian trouble with the length problem.* Suppose $Y \sim N_p(\theta, I_p)$, i.e. independent components $Y_i \sim N(\theta_i, 1)$, with interest in the length $\rho = \|\theta\| = (\theta_1^2 + \dots + \theta_p^2)^{1/2}$.

check also
Ex. 8.29

(a) Show that the Jeffreys prior for θ is the flat one, and that $\theta | y \sim N_p(y, I_p)$. Deduce that $\rho^2 | y$ has the distribution $\chi_p^2(\|y\|^2)$. Explain that the Bayes estimator for this squared length $\lambda = \rho^2$ parameter is $\hat{\lambda} = p + \|y\|^2$, with expected value $2p + \rho^2$ – so this Bayesian prior leads to potentially drastic overshooting.

(b) Now suppose there is prior information $\pi_{\text{vip}}(\rho)$ for the length. Explain that to work out the combo prior we must examine the $(p-1)$ -parameter model, with log-density

$$\log g_\rho = -\frac{1}{2} \sum_{i=1}^{p-1} (y_i - \theta_i)^2 - \frac{1}{2} (y_p - \theta'_p)^2, \quad \text{with } \theta'_p = (\rho^2 - \theta_1^2 - \dots - \theta_{p-1}^2)^{1/2}.$$

Show that the score function becomes

$$u_j = (y_j - \theta_j) - (\theta_j / \theta'_p)(y_p - \theta_p) \quad \text{for } j = 1, \dots, p-1.$$

Writing for clarity γ for this reduced model parameter $(\theta_1, \dots, \theta_{p-1})$, show that the Fisher information matrix becomes

$$J_\rho(\gamma) = I_{p-1} + (1/(\theta'_p)^2) \gamma \gamma^\top, \quad \text{with } |J_\rho(\gamma)| = 1 + \|\gamma\|^2 / (\rho^2 - \|\gamma\|^2) = \rho^2 / (\rho^2 - \|\gamma\|^2).$$

Conclude that the combo prior becomes

$$\pi_{\text{combo}}(\theta) \propto \pi_{\text{vip}}(\rho) / (1 - \|\gamma\|^2 / \rho^2)^{1/2}.$$

Illustrate this, for $p = 2$, plotting the posterior for ρ under the neutral Jeffreys and the informative vip-combo prior, for $y = (y_1, y_2)$ not far from zero, and with your own choice for $\pi_{\text{vip}}(\rho)$.

Ex. 7.35 *Parameter invariance for the vip-Jeffreys prior.* Within the framework of Ex. 7.44, for the vip-combo prior $\pi_{\text{combo}}(\mu, \gamma)$, consider a smooth parameter transformation of the non-vip part, say $\gamma = h(\phi)$ with $\phi = h^{-1}(\gamma)$.

(a) Starting with the combo prior (7.6), show that the implied density for (μ, ϕ) becomes

$$\pi^*(\mu, \phi) = \pi_{\text{combo}}(\mu, h(\phi)) \left| \frac{\partial(\mu, h(\phi))}{\partial(\mu, \phi)} \right| \propto \pi_{\text{vip}}(\mu) |J(h(\phi))|^{1/2} |G_2(\phi)|,$$

say, with $\partial h(\phi) / \partial \phi$ the $(p-1) \times (p-1)$ Jacobi matrix of the $h(\phi)$ transformation.

(b) For the reparametrised density $f(y, \mu, h(\gamma))$, with μ kept fixed, show that the score function is

$$u^*(y, \mu, \phi) = \partial \log f(y, \mu, h(\phi)) / \partial \phi = G_2(\phi) u(y, \mu, h(\gamma)).$$

From this, show that the Fisher matrix for ϕ becomes

$$J^*(\phi | \mu) = E_\phi u^*(Y, \mu, \phi) u^{*\top}(Y, \mu, \phi) = G_2(\phi) J(h(\phi)) G_2(\phi)^\top.$$

(c) Explain that the vip-combo method, for this reparametrised model, says that we should use

$$\pi^{**}(\mu, \phi) = \pi_{\text{vip}}(\mu) |G_2(\phi) J(h(\phi)) G_2(\phi)^\top|^{1/2}.$$

Convince yourself that this is, luckily, the same as $\pi^*(\mu, \phi)$; that is, the vip-combo method is invariant with respect to parameter transformation of the non-vip part.

Marginal distributions and Bernshtein–von Mises theorems

Ex. 7.36 *The marginal distribution.* Suppose we have data $y_1, \dots, y_n$ from a model $f(y, \theta)$, with a prior $\pi(\theta)$. Most of the time Bayesians care about the posterior distribution, but on occasion, also in connection with bigger setups, one needs the marginal distribution, which is $\bar{f}(y_1, \dots, y_n) = \int L_n(\theta) \pi(\theta) d\theta$, in terms of the likelihood function $L_n(\theta)$. In various setups there will be a clear formula for this $\bar{f}$, as seen below; see also Ex. 7.38 for a very useful approximation method for more complex cases.

(a) Let $y_1, \dots, y_n$ be independent Bernoulli variables with $\Pr(y_i = 1 | \theta) = \theta$, and let $\theta \sim \text{Beta}(a, b)$. Writing $z = \sum_{i=1}^n y_i$ for the number of 1s, show that

$$\bar{f}(y_1, \dots, y_n) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+z)\Gamma(b+n-z)}{\Gamma(a+b+n)}.$$

(b) Let then $y_1, \dots, y_n$ be independent $\text{Pois}(\theta)$, with a $\text{Gam}(a, b)$ prior, as with Ex. 7.2. Show that

$$\bar{f}(y_1, \dots, y_n) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+n\bar{y})}{(b+n)^{a+n\bar{y}}} \frac{1}{y_1! \dots y_n!}.$$

(c) Consider the linear regression setup with $Y \sim N_n(X\beta, \sigma^2 I_n)$, with σ known, and the prior $\beta \sim N_p(\beta_0, \Sigma_0)$. In Ex. 7.20 we found the posterior distribution for β . Now find also that the joint marginal distribution for Y is $N_n(X\beta_0, \sigma^2 I_n + X\Sigma_0 X^t)$. Explain that this can be represented as $Y = X\beta_0 + Xu + \varepsilon$, where $u \sim N_p(0, \Sigma_0)$ is independent of $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

Ex. 7.37 *The gamma-normal induced marginal model.* For the direct Bayesian use one only needs the prior to posterior computation, in this case from the initial $\text{GN}(a, b, \xi_0, v)$ to the updated GN given in Ex. 7.24, and one somehow bypasses the marginal density $\bar{f}(y_1, \dots, y_n)$ of the data, the likelihood with the parameters (ξ, σ) integrated out according to the prior. On occasion this marginal distribution is important, however, and also finds use as a model in its own right, for positively dependent data.

(a) (xx first things with $\xi \sim N(\xi_0, \tau_0^2)$, known σ . two ways of computing and seeing $\bar{f}(y_1, \dots, y_n)$. do marginal moments and correlations. xx)

(b) Then the $\text{GN}(a, b, \xi_0, v)$ gamma-normal prior for (λ, ξ) , as with Ex. 7.24. Show first that the likelihood times the prior, $L_n(\lambda, \xi)p(\lambda, \xi)$, can be expressed as

$$\frac{\lambda^{n/2}}{(2\pi)^{n/2}} \exp[-\frac{1}{2}\lambda\{Q_0 + n(\xi - \bar{y})^2\}] \frac{b^a}{\Gamma(a)} \lambda^{a-1} (v\lambda)^{1/2} \exp[-\lambda\{b + \frac{1}{2}v(\xi - \xi_0)^2\}].$$

Then integrate out the ξ to get

$$\frac{1}{(2\pi)^{n/2}} \frac{b^a}{\Gamma(a)} \lambda^{a+n/2-1} \left(\frac{v}{v+n}\right)^{1/2} \exp\{-\lambda(b + \frac{1}{2}Q_0 + \frac{1}{2}d_n(\bar{y} - \xi_0)^2)\},$$

with $d_n = (1/v + 1/n)^{-1}$ as per Ex. 7.24. Show then that this leads to the marginal density being

$$\bar{f}(y_1, \dots, y_n) = (2\pi)^{-n/2} \left(\frac{v}{v+n}\right)^{1/2} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+n/2)}{\{b + \frac{1}{2}Q_0 + \frac{1}{2}d_n(\bar{y} - \xi_0)^2\}^{a+n/2}}.$$

Ex. 7.38 *Approximating the marginal distribution.* In the setup of Ex. 7.36, we go through a useful type of Laplace approximation for the marginal. Arguments below are also related to the model selection strategy BIC; see Ex. 12.15.

(a) Writing as usual $\ell_n(\theta)$ for the log-likelihood, with maximum value $\ell_{n,\max} = \ell_n(\hat{\theta})$, in terms of the ML estimator, show that

$$\bar{f}(y_1, \dots, y_n) = \exp(\ell_{n,\max}) \int \exp\{\ell_n(\theta) - \ell_n(\hat{\theta})\} \pi(\theta) d\theta.$$

With $J_n = -(1/n)\partial^2 \ell_n(\hat{\theta})/\partial\theta\partial\theta^t$ the normalised observed information matrix, of dimension say $p \times p$, show that the marginal can be approximated with

$$\begin{aligned}\bar{f} &\doteq \exp(\ell_{n,\max}) \int \exp\{-\tfrac{1}{2}n(\theta - \hat{\theta})^t J_n(\theta - \hat{\theta})\} \pi(\theta) d\theta \\ &= \exp(\ell_{n,\max}) \int \exp(-\tfrac{1}{2}s^t J_n s) \pi(\hat{\theta} + s/\sqrt{n}) ds / n^{p/2} \\ &\doteq \exp(\ell_{n,\max}) \pi(\hat{\theta}) (2\pi)^{p/2} |J_n|^{-1/2} / n^{p/2}.\end{aligned}$$

(b) Check out a couple of settings, where you can compute the $\bar{f}$ explicitly, and look into how successful the approximation formula

$$\log \bar{f} \doteq \ell_{n,\max} - \tfrac{1}{2}p \log n + \log \pi(\hat{\theta}) - \tfrac{1}{2} \log |J_n| + \tfrac{1}{2}p \log(2\pi),$$

is, with its two leading terms.

Ex. 7.39 Bernshtein–von Mises approximations. Suppose observations $Y_1, \dots, Y_n$ are i.i.d. from a density $f(y, \theta)$, with $\pi(\theta)$ a prior for the model parameter, of dimension say p . The posterior density can of course be quite complicated, perhaps necessitating numerical efforts, or simulation, for its evaluation. Remarkably, there are generic and simple normal approximations, however. For an informative illustration and precursor for the general results below, check Ex. 7.14.

(a) Show that the posterior density $\pi_n(\theta) = \pi(\theta | \text{data})$ is proportional to $\pi(\theta) \exp\{\ell_n(\theta)\}$, with $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i, \theta)$ the log-likelihood function.

(b) Let $\hat{\theta}$ be the ML estimator, and $J_n = -(1/n)\partial^2 \ell_n(\hat{\theta})/\partial\theta\partial\theta^t$ the normalised observed information, as per likelihood theory of Ch. 5. Show that the density of $Z_n = \sqrt{n}(\theta - \hat{\theta})$ is $g_n(z) = \pi_n(\hat{\theta} + z/\sqrt{n})(1/n^{p/2})$, and that it can be approximated as

$$g_n(z) \propto \pi(\hat{\theta} + z/\sqrt{n}) \exp\{\ell_n(\hat{\theta} + z/\sqrt{n}) - \ell_n(\hat{\theta})\} \doteq \pi(\hat{\theta} + z/\sqrt{n}) \exp(-\tfrac{1}{2}z^t J_n z).$$

(c) Suppose then that the data really are i.i.d. from the model, with an underlying θ_{true} . In particular, then $\hat{\theta} \rightarrow_{\text{pr}} \theta_{\text{true}}$ and $J_n \rightarrow_{\text{pr}} J = J(\theta_{\text{true}})$, by the general likelihood theory of Ch. 5. If $\pi(\theta)$ is continuous in a neighbourhood around θ_{true} , show that $g_n(z)$ tends to the density of a $N_p(0, J^{-1})$, in probability. In concrete terms, show

$$D_n = \int |g_n(z) - \phi_p(z, 0, J^{-1})| dz \rightarrow_{\text{pr}} 0.$$

This is one of several versions of *Bernshtein–von Mises (BvM theorems)*. These are Bayesian mirror versions of the classical maximum likelihood asymptotics results in the frequentist camp:

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_{\text{true}}) &\rightarrow_d N(0, J^{-1}), \\ \sqrt{n}(\theta - \hat{\theta}) | \text{data} &\rightarrow_d N(0, J^{-1}), \text{ in probability.}\end{aligned}$$

(d) From the above, argue that we under mild integrability regularity have (i) $\sqrt{n}(\hat{\theta}_B - \hat{\theta}) \rightarrow_{\text{pr}} 0$, i.e. the Bayesian posterior mean and the ML have identical limit distributions; and (ii) $n \text{Var}(\theta | \text{data}) \rightarrow_{\text{pr}} J^{-1}$. In this connection, check with Ex. 2.5.

(e) Go back to the Beta-binomial and Dirichlet-multinomial setups of Ex. 7.14 and verify that results reached there are special cases of the BvM theorem above. For another clear illustration, consider the case where $Y_1, \dots, Y_n | \theta$ are i.i.d. $\text{Pois}(\theta)$ and $\theta \sim \text{Gam}(a_0, b_0)$. Check what the BvM theorem says in this situation. Choose also n and (a_0, b_0) for some brief numerical investigations.

sometimes
called the
Bayesian
Central Limit
Theorem

(f) Explain and discuss that we are learning here, quite generally, (i) that the effect of the prior is being washed out as the data volume increases, and (ii) that the posterior distributions, even for different start priors, are close to $N_p(\hat{\theta}, \hat{J}_{\text{full}}^{-1})$, with $\hat{J}_{\text{full}} = -\partial^2 \ell_{n, \hat{\theta}} / \partial \theta \partial \theta^t$ the observed Fisher information matrix. Argue that this invites the ‘Lazy Bayesian’ pragmatic view, that one may bypass the prior elicitation and jump directly to the $\theta \mid \text{data} \approx_d N_p(\hat{\theta}, \hat{J}_{\text{full}}^{-1})$, at least in cases with moderate or large datasets compared to the parameter dimension. the Lazy Bayesian

Multiparameter and hierarchical setups

For exercises in this subsection, check also partly related multiparameter exercises in the frequentist and risk function emphasised framework of Ch. 9, as with Ex. 9.25, 9.30.

Ex. 7.40 *Multiparameter inference.* Consider independent $Y_i \sim N(\theta_i, 1)$ for $i = 1, \dots, k$, the task being to estimate the ensemble of means, with loss function $L = \sum_{i=1}^k (a_i - \theta_i)^2$.

(a) The standard estimator is Y_i , simply; show e.g. that this is the ML estimator. Show that the risk function $R(\theta, Y)$ is constant, equal to k . Show also that with this loss function, the Bayes solution, after having started with some prior for $(\theta_1, \dots, \theta_k)$, are the posterior means, $E(\theta_i \mid \text{data})$.

(b) Then let θ_i have the prior $N(0, \tau^2)$, independently of the others, for $i = 1, \dots, k$. Show as with 7.19 that $\theta_i \mid \text{data} \sim N(\rho y_i, \rho^2)$, with $\rho = \tau^2 / (\tau^2 + 1)$. For the Bayes estimator $\hat{\theta}_i(\tau) = \rho Y_i$, show that the risk function is $\sum_{i=1}^k \{\rho^2 + (1 - \rho)^2 \theta_i^2\}$ and that the Bayes risk is $k\rho$.

(c) The ρY_i strategy requires of the Bayesian that τ is specified. When this is a too tall order, a natural method is to use the data to estimate ρ , leading to the *empirical Bayes strategy* $\theta_i^* = \hat{\theta}_i(\hat{\rho}) = \hat{\rho} Y_i$. With the $\theta_i \sim N(0, \tau^2)$ prior, explain that Y_i , marginally, is a $N(0, \tau^2 + 1)$, and that $V = \sum_{i=1}^k Y_i^2 \sim (\tau^2 + 1)\chi_k^2$. Since V/k estimates $\tau^2 + 1$, argue that $\hat{\rho} = \max(1 - k/V, 0)$, and relatives, are sensible estimators of ρ . Technically, this is not a bona fide Bayes method, as $\hat{\rho} Y_i$ is not the posterior mean stemming from a prior on $(\theta_1, \dots, \theta_k)$, but its frequentist performance may be excellent. In particular, versions of this empirical Bayes scheme will succeed in having risk functions below the standard risk level k , in all of the parameter space (provided $k \geq 3$, actually). This is the so-called Stein phenomenon; see connections from the present exercise to Ex. 9.25. empirical Bayes

(d) A full Bayesian solution, without resorting to the empirical Bayes method above, is to give a prior for τ , or equivalently the ρ , along with $\theta_i \mid \tau \sim N(0, \tau^2)$. Show that the joint density of the $2k + 1$ quantities in the game, i.e. $\rho, \theta_1, \dots, \theta_k, y_1, \dots, y_k$, may be written

$$\pi(\rho) \prod_{i=1}^k \left\{ \frac{1}{\tau} \phi\left(\frac{\theta_i}{\tau}\right) \phi(y_i - \theta_i) \right\} = \pi(\rho) \prod_{i=1}^k \left\{ \frac{1}{\rho^{1/2}} \phi\left(\frac{\theta_i - \rho y_i}{\rho^{1/2}}\right) (1 - \rho)^{1/2} \phi((1 - \rho)^{1/2} y_i) \right\}.$$

Integrate out the θ_i to show that $(\rho, y_1, \dots, y_k)$ follow the joint density proportional to $\pi(\rho)(1 - \rho)^{k/2} \exp\{-\frac{1}{2}(1 - \rho)v\}$. Show that the Jeffreys prior is proportional to $1/(1 - \rho)$, with which the posterior for ρ becomes $\pi(\rho \mid \text{data}) \propto (1 - \rho)^{k/2 - 1} \exp\{-\frac{1}{2}(1 - \rho)v\}$ on the unit interval.

(e) Explain that posterior simulation may be carried out via (i) drawing ρ from its posterior, e.g. via one-dimensional MCMC, and then (ii) drawing $\theta_1, \dots, \theta_k$ independently from their $N(\rho y_i, \rho)$. For the loss function L given above the Bayes solutions are the

posterior means; show that these are $\tilde{\theta}_i = \tilde{\rho} y_i$, with $\tilde{\rho} = E(\rho | \text{data})$. This can easily be computed numerically.

(f) As long as one may carry out posterior simulations, one may also find Bayes solutions with respect to more complicated loss functions than the quadratic one. For illustration, suppose the observations are

-1.277 -1.254 -0.791 -0.148 0.178 0.294 0.847 1.229 1.720 1.749

and set up a scheme to simulate say 10^5 realisations of $(\theta_1, \dots, \theta_k)$ given these datapoints, from the prior described above. Then find the Bayes estimators with the loss function $\sum_{i=1}^k |\hat{\theta}_i - \theta_i|$. Check what happens if data point no. 10 is pushed some steps and then farther to the right.

Ex. 7.41 *Multiparameter inference, II.* (xx nils rant. need polish and finish. xx) the natural real thing: independent $Y_i \sim N(\theta_i, \sigma^2)$ for $i = 1, \dots, k$, with unknown means and the same unknown variance. the object is estimating all the means, with loss $L = \sum_{i=1}^k (\hat{\theta}_i - \theta_i)^2 / \sigma^2$. direct estimator is $\bar{Y}_i$, with risk k . with any prior construction, the Bayes solution is $E(\theta_i | \text{data})$. natural prior: $\theta_i \sim N(\theta_0, \tau^2)$. then

$$\theta_i | y_i \sim N((1 - \rho)\theta_0 + \rho y_i, \rho\sigma^2), \quad \text{with } \rho = \tau^2 / (\sigma^2 + \tau^2).$$

here we are at a higher level of complexity compared to Ex. 7.41; there we knew θ_0 and σ , and could concentrate on a prior for ρ . for empirical Bayes, note that marginally, $Y_i \sim N(\theta_0, \sigma^2 + \tau^2)$. the sample mean is $\hat{\theta}_0 = \bar{Y}$, and $V = \sum_{i=1}^k (Y_i - \bar{Y})^2 \sim (\sigma^2 + \tau^2) \chi_{k-1}^2$, leading to an estimator for $\sigma^2 + \tau^2$. the Bayes estimator is

$$\hat{\theta}_i = E\{(1 - \rho)\theta_0 + \rho y_i | \text{data}\}.$$

we need posterior for (θ_0, ρ) , but perhaps need a prior for the full triple (σ, θ_0, τ) . we may write up

$$\begin{aligned} f &= \pi(\sigma, \theta_0, \tau) \prod_{i=1}^k \left\{ \frac{1}{\tau} \phi\left(\frac{\theta_i - \theta_0}{\tau}\right) \frac{1}{\sigma} \phi\left(\frac{y_i - \theta_i}{\sigma}\right) \right\} \\ &= \pi(\sigma, \theta_0, \tau) \prod_{i=1}^k \left\{ \frac{1}{\rho^{1/2} \sigma} \phi\left(\frac{\theta_i - (1 - \rho)\theta_0 - \rho y_i}{\rho^{1/2} \sigma}\right) \frac{1}{(\sigma^2 + \tau^2)^{1/2}} \phi\left(\frac{y_i - \theta_0}{(\sigma^2 + \tau^2)^{1/2}}\right) \right\}. \end{aligned}$$

integrating out the θ_i we have

$$\pi(\sigma, \theta_0, \tau) \prod_{i=1}^k \frac{1}{(\sigma^2 + \tau^2)^{1/2}} \phi\left(\frac{y_i - \theta_0}{(\sigma^2 + \tau^2)^{1/2}}\right).$$

Ex. 7.42 *Multiparameter inference, III.* Consider independent geometric $Y_1, \dots, Y_m$, with distribution $(1 - \theta_i)^{y-1} \theta_i$ for Y_i . The task is the joint estimation of this ensemble of perhaps related probabilities, with these (ordered) observations from $m = 20$ experiments: 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 7, 7, 15, 17, 19, 22, 78.

(a) With $\theta_i \sim \text{Beta}(a, b)$, show that the posterior is $\theta_i | y_i \sim \text{Beta}(a + 1, b + y_i - 1)$, with posterior means $\hat{\theta}_i(a, b) = (a + 1) / (a + b + y_i)$.

(b) To estimate (a, b) from data, show that the marginal distribution of Y_i takes the form

$$\bar{f}(y_i) = \int_0^1 (1 - \theta)^{y-1} \theta \text{be}(\theta, a, b) d\theta = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + 1)\Gamma(b + y - 1)}{\Gamma(a + b + y)} \quad \text{for } y = 1, 2, \dots$$

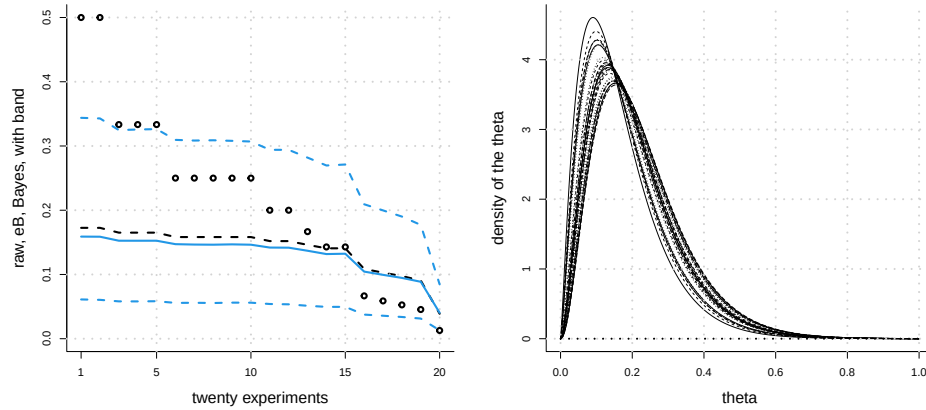


Figure 7.2: Joint estimation of 20 geometric probabilities. Left panel: raw estimates (small circles), empirical Bayes (dashed curve), and full Bayes posterior medians (full curve), with 90 percent band. Right panel: 25 simulated $\text{be}(\theta, a, b)$, with (a, b) from the posterior.

Code up the log-likelihood function to find the ML estimates $(\hat{a}, \hat{b}) = (2.811, 17.274)$. From this compute the empirical Bayes estimates $\hat{\theta}_i(\hat{a}, \hat{b})$, and plot them, along with the raw estimates $1/y_i$.

(c) Then carry out a full Bayesian analysis, using MCMC, with a flat prior for (a, b) in $[0, 100] \times [0, 100]$. This leads to say 10^5 realisations of (a, b) from the posterior distribution. Use these to simulate the same number 10^5 of each θ_i , given data, and read off their (0.05, 0.50, 0.95) quantiles. Construct a version of Figure 7.2, left panel.

(d) Use say 25 posterior simulated (a, b) to visualise Beta densities $\text{be}(\theta, a, b)$, as with Figure 7.2, right panel.

Leftovers and further material

Ex. 7.43 *How many piano tuners in Oslo?* Applications of Bayesian methods, in complicated and perhaps nonstandard cases, particularly in situations with limited data and many model parameters, involve the challenging task of *setting up the priors*. This could often involve translation of different pieces of information into probability distributions, perhaps followed by nontrivial combinations of these. This exercise is meant as a perhaps rough illustration of such themes. – The task is to guess the number N of piano tuners in Oslo, based on perhaps rough approximations and calculations, along with uncertainty assessments. (xx note, ca 2024, from nils piano tuner: $N = 23$ in oslo, 62 in Norway, but there are various half-timers and charlatans. xx)

(a) Such a moderately rough setup takes $N = np_1p_2/k$, with n the population of Oslo; p_1 the ‘piano fraction’ of these, so that np_1 is roughly the number of pianos in Oslo; p_2 the fraction of these again who tune their pianos at least once a year; and k the average number of piano tuning jobs carried about in the course of a year, by a piano tuner in Oslo. – Of course there are other ways to set up such things, with other components, and a clearer chain of arguments could be forced to form more precise definitions and assumptions (one could think in terms of families and their sizes, etc., to get a better grip on p_1 , etc.). Show that with independent priors for n, p_1, p_2, k , then a prior estimate

for N is $E n E p_1 E p_2 E 1/k$.

(b) An attempt is as follows: let

$$\begin{aligned} n &\sim N(\xi_0, \sigma_0^2), \text{ with prior guess } \xi_0; \\ p_1 &\sim \text{Beta}(c_1 p_{1,0}, c_1(1 - p_{1,0})), \text{ with prior guess } p_{1,0}; \\ p_2 &\sim \text{Beta}(c_2 p_{2,0}, c_2(1 - p_{2,0})), \text{ with prior guess } p_{2,0}; \\ k &\sim \text{Pois}(k_0), \text{ with prior guess } k_0. \end{aligned}$$

One then first finds good prior means, for the four components in the setup, and then finetune their precision via σ_0 for n , c_1 for p_1 , and c_2 for p_2 . – Try this setup, with prior guesses $0.60 \cdot 10^6$, 0.07 , 0.40 , $k_0 = 500$, and then variability determined by $\sigma_0 = 0.05 \cdot 10^6$, $c_1 = 15$, $c_2 = 10$. Simulate 10^5 realisations from each of the four distributions, inspect their histograms, and present a final histogram for $N = np_1 p_2 / k$. Also compute the 0.10, 0.50, 0.90 points in this distribution (our simulations for this particular prior lead to 4.1, 26.2, 91.4), and compare in particular the median with the mean.

(c) Play with other prior parameters, for some or all of the four prior components above, using your insights for finetuning of both prior guesses and the prior uncertainties. This should lead you to a perhaps more accurate prior for N than the one given above.

(d) Suppose we use the four prior components as above, but that further information is gathered regarding the two components n and p_2 . For n , the population size of Oslo, along with the perhaps not fully defined suburbs and surroundings (since piano tuners from Oslo might travel to Ski and Ås), assume there is a point estimate $\hat{n} = 0.64 \cdot 10^6$, with standard deviation $0.04 \cdot 10^6$. For p_2 , the fraction of pianos tuned at least once a year, a quick Facebook check with friends and acquaintances (and their friends and acquaintances again) indicates that $y_2 = 55$ of $m_2 = 100$ piano owners tune their instruments at least once a year. Use these extra pieces of information to update the priors of n and p_2 , and then revise the full distribution of $N = np_1 p_2 / k$. Simulate again 10^5 realisations from the N distribution, find out how the 0.10, 0.50, 0.90 quantiles have changed.

(e) (xx rounding this off. mention briefly Fermi. such questions sometimes asked in job interviews, to see how a candidate reasons. also the number of solar systems in a galaxy with intelligent life. illustrate the machinery of changing the final $N = np_1 \cdots p_k$ with more information on a single component. xx)

Ex. 7.44 *Bayes with Jeffreys in more challenging waters.* Carrying out Bayes analysis with the Jeffreys prior becomes harder when there is no simple formula for the latter. Consider a setup with given densities f_0 and f_1 , but unknown mixture proportion p for data coming from $f = (1 - p)f_0 + pf_1$.

(a) Show that the Fisher information becomes

$$J(p) = \int \frac{\{f_1(x) - f_0(x)\}^2}{(1 - p)f_0(x) + pf_1(x)} dx.$$

For a few choices of f_0 and f_1 , compute and plot $\pi_{\text{jeff}}(p)$ numerically. Comment in particular on the case where f_0 and f_1 are normal with the same variance.

(b) Let for concreteness f_0 and f_1 be $N(\xi_0, 1)$ and $N(\xi_1, 1)$, with $(\xi_0, \xi_1) = (0, 1)$, though your code should work for any other pair of densities. Generate $n = 25$ points from the mixture density, using say $p_{\text{true}} = 0.33$. Your task is to simulate realisations from

check also with
Story #89

$\pi(p|\text{data})$. For this use MCMC, see Ex. 7.17, which for the acceptance probability computation requires numerical evaluation of $\pi_{\text{jeff}}(p^{\text{prop}})/\pi_{\text{jeff}}(p^{\text{old}})$. The point is that this is entirely possible, even without having an easier formula for the Fisher information.

(c) For the one-dimensional case, as above, it suffices to being able to evaluate one-dimensional integrals. This is more challenging in the vector case. Explain how you can still get hold of numerical approximations to $\pi_{\text{jeff}}(a, b)$, say, in a two-parameter model, by simulating a number of score vectors at each position (a, b) in the evaluation of the acceptance probability step for the MCMC. Try such a scheme for the case of data sampled from $f = (1 - p)f_0 + pf_1$, this time with $p = 0.33$ known and fixed, but with unknown means for $N(\xi_0, 1)$ and $N(\xi_1, 1)$.

Notes and pointers

point to Ch. 14 for various extensions, quote Lindley 1972, Bayesians had been ‘embarrassingly silent’ about nonparametrics.

for mcmc, point Metropolis 1953, and to Alfred Teller, ‘the father of the hydrogen bomb’.

point to de Finetti, “probabilità non esiste”, and Spiegelhalter 2025.

probability does not exist. also Spiegelhalter Nature 2025. if you accept exchangeability, then you’re a Bayesian, by necessity.

(xx sketching solution, to be pushed to ‘Solutions’. the separate lemma in 7.16. xx) $\sqrt{c}(U_c/c - p_0) \rightarrow_d A$, $\sqrt{c}(V_c/c - (1 - p_0)) \rightarrow_d B$, with variances p_0 and $1 - p_0$; derivatives of $u/(u + v)$ are $(u + v - u)/(u + v)^2$ and $-u/(u + v)^2$; evaluated at $(p_0, 1 - p_0)$ become $1 - p_0$ and $-p_0$; giving by the delta method

$$\sqrt{c}\left(\frac{U_c/c}{U_c/c + V_c/c} - \frac{p_0}{p_0 + 1 - p_0}\right) \rightarrow_d (1 - p_0)A - p_0B,$$

which is a zero-mean normal with variance $(1 - p_0)^2 p_0 + p_0^2 (1 - p_0) = p_0(1 - p_0)$.

(xx mention Varian (1975); Zellner (1986); Claeskens and Hjort (2008a) for the linex loss; but this should perhaps be in Ch. 9. mention Stigler (1983), who uses Bayes’s Theorem to help him spot a candidate for an earlier discoverer of Bayes’ Theorem. mention MCMC revolution since the 1990ies. point to Story #42. xx)

(xx on the Jeffreys prior. the argument used is slightly shaky since the transformation formula is used for possibly improper densities. point to Bernardo-Berger. can push in the nils-emil combo prior; informative on some components, non-informative on the rest. Cencov papers re Jeffreys being only solution to invariance demand. xx)

(xx hierarchical Bayes, with or without empirical Bayes. point to Efron etc. xx)

I.8

Confidence distributions, confidence curves, combining information sources

With μ a focus parameter, a function of the full parameter vector θ , the Bayesian setup gives a posterior distribution. This requires the conceptually and practically difficult task of defining a prior for the full θ , however. Confidence distributions (CDs) are a frequentist parallel, yielding post-data distributions for such focus parameters, without any prior. In this chapter we develop theory for CDs and confidence curves, and also find ways of combining CDs across different information sources. Constructing and computing CDs is not an easy or automatic task, but we develop and illustrate several recipes. For the exponential family class, we derive optimal CDs, with their own clear recipes.

Key words: Beta correction, boundary constraints, combining CDs, confidence conversion, confidence curves, confidence distributions, exponential family, meta-analysis, II-CC-FF, prediction CDs, t-bootstrapping

Confidence distributions and confidence curves are fruitful statistical inference summaries. Suppose in general terms that data y stem from a model $f(y, \theta)$, with model parameter $\theta = (\theta_1, \dots, \theta_p)$, and that $\mu = \mu(\theta_1, \dots, \theta_p)$ is a parameter of particular interest. A *confidence distribution* for μ , a CD, for short, is a function $C(\mu, y)$, such that (i) it is a c.d.f. in μ , for each dataset y , and (ii) the distribution of $U = C(\mu_0, Y)$ is uniform at the true value $\mu_0 = \mu(\theta_0)$. In other words,

$$\Pr_{\theta_0}\{a \leq C(\mu_0, Y) \leq b\} = b - a \quad \text{for each } a, b \in [0, 1].$$

Assuming this random c.d.f. has a unique inverse, then, we have

$$\Pr_{\theta}\{C^{-1}(0.05, Y) \leq \mu \leq C^{-1}(0.95, Y)\} = 0.90, \quad (8.1)$$

and of course similarly for other choices of quantiles. This is by definition making $[C^{-1}(0.05, y_{\text{obs}}), C^{-1}(0.95, y_{\text{obs}})]$ a 90 percent confidence interval for the focus parameter μ . The CD concept is hence related to and an extension of the confidence intervals, see Ch. 4. The *confidence curve* is a related summary graph, most often computed from the CD via $cc(\mu, y) = |1 - 2C(\mu, y)|$. It has the practical property that $\{\mu: cc(\mu, y) \leq 0.90\}$ gives the 90 percent interval directly; similarly, all intervals at any desired confidence level can be read off from the confidence curve. We sometimes write simply $C(\mu)$ and $cc(\mu)$, omitting the data argument, when clear from the context what the data are, in $C(\mu) = C(\mu, \text{data})$ and $cc(\mu) = cc(\mu, \text{data})$.

Constructing a CD is not always an easy or automatic task, but we develop several practical recipes, some of which are based on approximate normality, or on more general methods of likelihood theory. Just as tests have detection power, also CDs have power, and theory is developed below to find optimal CDs in classes of situations. This is partly paralleling the optimal testing methodology of Ch. 4. All in all we develop and illustrate the following recipes: (i) Via the c.d.f. of an estimator; (ii) normal approximation; (iii) based on a pivot; (iv) deviance and Wilks theorems; (v) t-bootstrapping; (vi) the optimal CD via conditional distributions, if inside the exponential family.

The CDs are post-data graphical summaries of the level of uncertainty for any focus parameter, and can be seen as frequentist parallels to the Bayesian posterior distributions; here there is no prior, however. We illustrate this ‘clear data-only based posteriors without priors’ aspect of the CDs through theory and applications (xx perhaps point to a few Stories xx).

Combining different information sources is a broad statistical theme, going back to the first meta-analysis concepts and methods of Karl Pearson just after 1900 (Simpson and Pearson, 1904). The more familiar meta-analysis methods aim at combining independent estimators for the same quantity, or for providing a broader population assessment of similar but not identical parameters. CDs are useful for such endeavours, and we provide methods for combining sources more general than the traditional ones.

Recipes One, Two, Three for constructing CDs

Ex. 8.1 *The probability transform.* Some of the following facts are related to various operations for confidence distributions and confidence curves

(a) Suppose X has a continuous and increasing cumulative distribution function F , i.e. $F(x) = \Pr(X \leq x)$. Show that $U = F(X)$ is uniform on the unit interval. Any continuously distributed random variable can hence be transformed to uniformity, via this *probability transform*.

(b) Show that also $U_2 = 1 - F(X)$ and $U_3 = |1 - 2F(X)|$ have uniform distributions.

(c) Simulate a million copies of $x_i \sim N(0, 1)$, and check the histogram of $\Gamma_1(x_i^2)$, where Γ_ν is the cumulative distribution function of a χ_ν^2 . Comment on what you find.

Ex. 8.2 *Recipe One: via the c.d.f. of an estimator.* Suppose θ is a one-dimensional parameter, for which we need a CD, after having observed data y_{obs} . If there is an estimator $\hat{\theta}$, with a distribution depending only on this θ , there is a clear recipe.

(a) Assume therefore that $\hat{\theta}$ has a continuous distribution function $K_\theta(x) = \Pr_\theta(\hat{\theta} \leq x)$; its distribution is here required to depend only on θ , not on other aspects of the underlying model employed. Consider Recipe One, the construction

Recipe One

$$C(\theta, y_{\text{obs}}) = \Pr_\theta(\hat{\theta} \geq \hat{\theta}_{\text{obs}}) = 1 - K_\theta(\hat{\theta}_{\text{obs}}),$$

a curve that can be computed and plotted post-data, where $\hat{\theta}_{\text{obs}} = \hat{\theta}(y_{\text{obs}})$ is the observed estimate. Show that it has the property that the random $C(\theta, Y)$ is uniformly distributed, for each fixed θ .

(b) To illustrate, go through the details for the case of using $\hat{\theta} = 1/\bar{Y}$, with i.i.d. observations $Y_1, \dots, Y_n$ from the exponential $\theta \exp(-\theta y)$. Show first that $2\theta Y_i \sim \chi_2^2$, and derive $K_\theta(x) = 1 - \Gamma_{2n}(2n\theta/x)$, with Γ_{2n} the c.d.f. of the χ_{2n}^2 . Simulate data and plot the CD $C(\theta, y_{\text{obs}}) = \Gamma_{2n}(2n\theta/\hat{\theta}_{\text{obs}})$. From the CD, find a 95 percent interval for θ .

(c) Assume $X_1, \dots, X_m$ are i.i.d. $\text{Expo}(\theta_1)$ and that $Y_1, \dots, Y_n$ are i.i.d. $\text{Expo}(\theta_2)$. Find the distribution of the estimator $\hat{\rho} = \hat{\theta}_1/\hat{\theta}_2$ for the ratio $\rho = \theta_1/\theta_2$, and derive the associated CD.

(d) Generate $n = 25$ datapoints from the double exponential $f(y, \theta) = \frac{1}{2} \exp(-|y - \theta|)$, using your favourite θ_0 . Compute and display the CD for θ based on the median M_n .

(e) For a simpler and more fundamental illustration, suppose $\hat{\theta} \sim N(\theta, \kappa^2)$, with a known κ . Show that Recipe One gives $C(\theta) = \Phi((\theta - \hat{\theta})/\kappa)$. Verify that the famous 95 percent interval $\hat{\theta} \pm 1.96 \kappa$ agrees with this.

Ex. 8.3 *Confidence distribution and confidence curve for the normal standard deviation.* The confidence distribution C and the confidence curve cc are close cousins, and they do not need to be both displayed for each new statistical application. Here is a simple illustration. You observe the $n = 6$ data points 4.09, 6.37, 6.87, 7.86, 8.28, 13.13 from a normal distribution and wish to assess the underlying spread parameter, the standard deviation σ .

(a) For the empirical variance, use $\hat{\sigma}^2 \sim \sigma^2 \chi_m^2/m$, with $m = n - 1$, to build the CD

$$C(\sigma, y_{\text{obs}}) = \Pr_{\sigma}(\hat{\sigma} \geq \hat{\sigma}_{\text{obs}}) = 1 - \Gamma_m(m\hat{\sigma}_{\text{obs}}^2/\sigma^2).$$

Here y_{obs} represents the observed data, and $\hat{\sigma}_{\text{obs}}$ the observed point estimate. Show that $C(\sigma, Y) \sim \text{unif}$, where Y represents a random data set $Y_1, \dots, Y_n$, from the σ in question. In particular, the distribution of $C(\sigma, Y)$ does not depend on σ . Make a graph, also of the associated confidence curve

$$cc(\sigma, y_{\text{obs}}) = |1 - 2C(\sigma, y_{\text{obs}})| = |1 - 2\Gamma_m(m\hat{\sigma}_{\text{obs}}^2/\sigma^2)|.$$

Compute the *median confidence estimate* $\hat{\sigma}_{0.50} = C^{-1}(0.50, y_{\text{obs}})$ and the natural 90 percent confidence interval $[C^{-1}(0.05, y_{\text{obs}}), C^{-1}(0.95, y_{\text{obs}})]$. Find and display also the *confidence density* $c(\sigma, y_{\text{obs}})$, the derivative of the CD.

(b) Compute also the *confidence density* $c(\sigma, y_{\text{obs}})$ associated with the CD. Compute furthermore its mode, say σ^* , and briefly assess its properties as an estimator of σ .

(c) A Bayesian approach to the same problem, i.e. finding a posterior distribution for σ , is to start with a prior $\pi(\sigma)$ and then compute $\pi(\sigma | y_{\text{obs}}) \propto \pi(\sigma)g(\hat{\sigma}, \sigma)$, where $g(\hat{\sigma}, \sigma)$ is the likelihood, here the density function for $\hat{\sigma}$ as a function of σ . When does such a Bayesian approach agree with the confidence density?

(d) Suppose there are two independent normal samples, with standard deviations σ_1 and σ_2 . Construct a CD for $\rho = \sigma_1/\sigma_2$. Invent a second simple small dataset, to complement the first dataset given above, and then compute and display the confidence curve $cc(\rho, \text{data})$.

Ex. 8.4 *CDs via the c.d.f. of a statistic.* There is a mild but practically important extension of Ex. 8.2, as follows.

(a) In the one-parameter setup of that exercise, suppose T is a statistic, stochastically increasing in θ . Show that $C(\theta) = \Pr_{\theta}(T \geq t_{\text{obs}})$ is a CD. So the T does not need to be an estimator, or a fine-tuned estimator, for θ , as long as bigger value of T indicate bigger values of θ .

(b) When constructing the CD for the normal σ , in Ex. 8.3, convince yourself that with any continuously increasing $T = T(\hat{\sigma})$, the $\Pr_{\sigma}(T \geq t_{\text{obs}})$ recipe simply gives the same answer. In particular, there is no need to fine-tune the numerator in $\sum_{i=1}^n (Y_i - \bar{Y})^2/m_n$.

the confidence
curve

(c) Simulate a little dataset from a normal distribution, then compute and plot two CDs: first the one above, associated with $\hat{\sigma}$, then the other, based on the range $T = \max Y_i - \min Y_i$.

Ex. 8.5 *A CD for mean by standard deviation.* As in Ex. 3.38, consider i.i.d. $Y_1, \dots, Y_n$ from a normal (ξ, σ^2) , with inference needed for $\kappa = \xi/\sigma$. The natural estimator is $\hat{\kappa} = \hat{\xi}/\hat{\sigma}$, with the usual sample mean and sample standard deviation. As we saw in the exercise pointed to, $X = \sqrt{n}\hat{\kappa} \sim t_m(\sqrt{n}\kappa)$, with $m = n - 1$ degrees of freedom for the noncentral t .

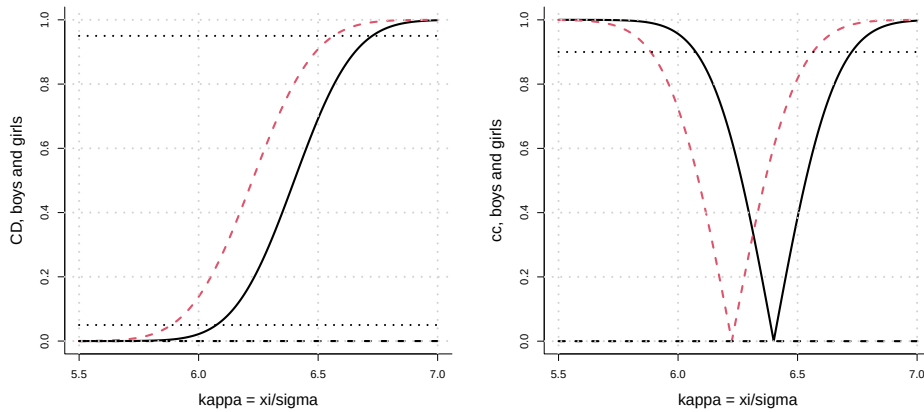


Figure 8.1: For 548 boys and 480 girls born in Oslo, 2001–2008, CDs (left panel) and confidence curves (right panel), for the birthweight ratio $\kappa = \xi/\sigma$ (full curves for boys, slanted curves for girls). One may read off 0.05 and 0.95 confidence points from the CDs, or the 0.90 confidence interval directly from the ccs; 6.075 to 6.726 for boys, 5.886 to 6.564 for girls. Median confidence estimates are 6.400 and 6.225.

(a) We start with a stylised version of the problem, with a single $T \sim t_m(a)$, with known m and unknown a . Show that $C(a, t_{\text{obs}}) = \Pr_a(T \geq t_{\text{obs}}) = 1 - G_m(t_{\text{obs}}, a)$ is a CD for a , in terms of the c.d.f. for the $t_m(a)$ distribution. The associated confidence curve is then $cc(a, t_{\text{obs}}) = |1 - 2G_m(t_{\text{obs}}, a)|$.

(b) Going back to the normal sample, show that the natural CD for $\kappa = \xi/\sigma$ becomes

$$C(\kappa) = \Pr_{\kappa}(\hat{\kappa} \geq \hat{\kappa}_{\text{obs}}) = 1 - G_m(\sqrt{n}\hat{\kappa}_{\text{obs}}, \sqrt{n}\kappa).$$

(c) For an illustration, we use the data from Story #22, with birthweights for 548 boys and 480 girls born in Oslo, 2001–2008. The required sufficient information $(\hat{\xi}, \hat{\sigma})$ is (3.669, 0.573) for the boys and (3.494, 0.561) for the girls. Compute and display the CDs and confidence curves, as with Figure 8.1. Find also 99 percent confidence intervals for the κ parameter in the two groups.

(d) Consider a linear regression setup, with $y_i \sim N(x_i^t \beta, \sigma^2)$ for $i = 1, \dots, n$. For two covariate vectors x_a and x_b , examine the mean difference $E(Y_b | x_b) - E(Y_a | x_a) = (x_b - x_a)^t \beta$, along with $\kappa = (x_b - x_a)^t \beta / \sigma$, a natural quantity informing us about the essential statistical difference between the two. Construct a CD for this parameter. Start out with $Y_b - Y_a \sim N(d^t \beta, 2\sigma^2)$, with $d = x_b - x_a$, so $p = \Pr(Y_b > Y_a) = \Phi(\kappa/\sqrt{2})$, with

$\kappa = d^t \beta / \sigma$ (xx check this xx). Explain that $\hat{\kappa} = d^t \hat{\beta} / \hat{\sigma}$, leading to

$$\sqrt{n} \hat{\kappa} / \tau_n = \frac{\sqrt{n} d^t \hat{\beta} / (\sigma \tau_n)}{\hat{\sigma} / \sigma} = \frac{N(\sqrt{n} \kappa / \tau_n, 1)}{(\chi_m^2 / m)^{1/2}}$$

and to the CD

$$C(\kappa) = \Pr_{\kappa}(\sqrt{n} \hat{\kappa} \geq \sqrt{n} \kappa_{\text{obs}}) = 1 - G_m(\sqrt{n} \hat{\kappa}_{\text{obs}} / \tau_n, \sqrt{n} \kappa / \tau_n).$$

Ex. 8.6 *Computing a CD with simulation and isotonic repair.* Suppose one observes $y_1, \dots, y_n$ from the one-parameter Weibull distribution with c.d.f. $F(y, b) = 1 - \exp(-y^b)$, with sample size $n = 25$, and computes the data mean $\bar{y}_{\text{obs}} = 1.313$.

(a) Though we do not actually need this in the CD computations here, find an estimate of b based on $E Y_i = \Gamma(1 + 1/b)$; see Ex. 1.61. Show that $C(b) = \Pr_b(\bar{Y} \leq \bar{y}_{\text{obs}})$ is a CD for b ; this is related to what we learn in Ex. 8.4.

(b) The practical obstacle here is that $\bar{Y}$ does not have a simple distribution. But we're saved by simulation. Show that the simulation recipe $Y_i^* = V_i^{1/b}$ produces outcomes from the Weibull $F(y, b)$, where the V_i are unit exponential. For a grid of b values, e.g. from 0.20 to 1.20, compute the simulation based $C^*(b)$, the proportion of B cases where the simulated $\bar{Y}^*$ is below $\bar{y}_{\text{obs}}$. Compute also the confidence curve $cc^*(b) = |1 - 2C^*(b)|$. For this simple example it is easy to accomplish this with a high B , say 10^5 , to make $C^*(b)$ and $cc^*(b)$ smooth and very close to the real $C(b)$ and $cc(b)$; for this illustration, however, make the simulation size as relatively small as $B = 100$, and plot the curves, as in Figure 8.2.

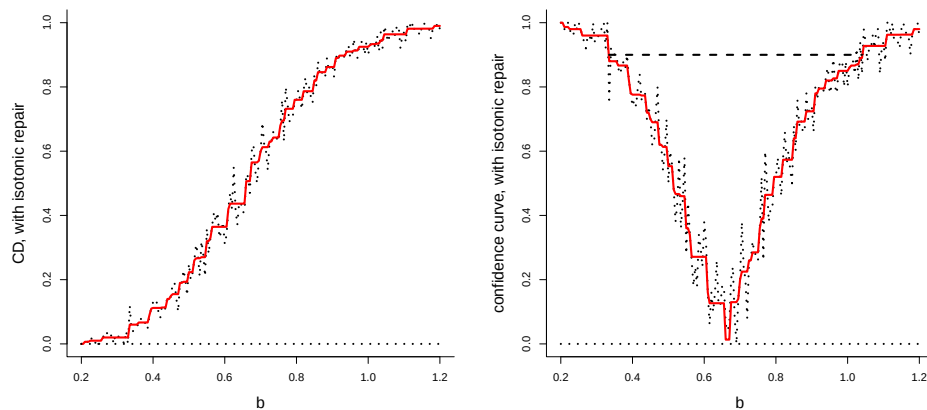


Figure 8.2: Simulation based confidence distribution $C^*(b)$ and confidence curve $cc^*(b)$ for the Weibull parameter b , based on the observed sample mean $\bar{y}_{\text{obs}} = 1.313$ for $n = 25$ data points, along with isotonic repairs. The simulation size here is the low $B = 100$.

(c) We learn that with a low or moderate simulation size B , the $C^*(b)$ and $cc^*(b)$ will be wiggly. We can do better, using the prior knowledge that $C(b)$ is increasing. There are several repair mechanisms, which from the potentially wiggly $C^*(b)$ create a monotonically increasing curve. A simple scheme is so-called *isotonic regression*, the details of which we do need to get into here. Supposing you have first created `bval` and `Cval` in your R session, you may use `Cvaliso=isoreg(bval,Cval)$yf`, which repairs

your $C^*(b)$ and $cc^*(b)$ to ensure monotonicity. Produce versions of Figure 8.2, left and right panels.

Ex. 8.7 *An extension of Recipe One.* In Ex. 8.2 we saw that the simple construction $C(\theta, y) = \Pr_\theta(\hat{\theta} \geq \hat{\theta}_{\text{obs}})$ gives a CD, in the case of one-dimensional setups with a well-defined estimator $\hat{\theta}$.

check Stories
#60 and #89,
where isotonic
repair is used

(a) When working with estimators, finetuning efforts are often exuded to trim away biases, getting the scaling right, etc. In a sense this is not needed here, when constructing the CD. Show that if $\hat{\alpha} = g(\hat{\theta})$, with any smooth increasing g , the recipe $C^*(\theta) = \Pr_\theta(\hat{\alpha} \geq \hat{\alpha}_{\text{obs}})$ gives precisely the same CD as without the g transformation.

(b) So this CD recipe relies merely on having an informative statistic, say Z , with a distribution stochastically increasing in θ ; it does not really have to be an estimator for that parameter. Show that $C(\theta, y) = \Pr_\theta(Z \geq z_{\text{obs}})$ is a bona fide CD.

(c) Show also that the construction works, if there are other parameters at play too, as long as the distribution of the chosen Z only depends on θ . Go through the details for the case of the Y_i being $N(\mu, \sigma^2)$, with $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2$, and also for $Z' = \sum_{i=1}^n |Y_i - M_n|$, where M_n is the empirical median. Compute, display, compare both CDs, based on Z and on Z' , for the simple dataset of Ex. 8.3 (with $n = 6$). For the Z case, there is a formula, but for the Z' case you would need simulation, for a grid of σ values; see Ex. 8.6.

(d) Suppose $X_1, \dots, X_m$ are independent with $X_j \sim N(\xi_j, 1)$, and that the hypothesis H_0 of equal means is to be checked. This is related to assessing the spread parameter λ , with $\lambda^2 = \sum_{j=1}^m (\xi_j - \bar{\xi})^2$, with $\bar{\xi}$ their average. With $W = \sum_{j=1}^m (X_j - \bar{X})^2$, show that $C(\lambda) = 1 - \Gamma_{m-1}(W_{\text{obs}}, \lambda^2)$ is a full CD for λ , with the noncentral chi-squared. Plotting that CD is rather more informative than merely saying yes-or-to a test.

Ex. 8.8 *The endpoint for a uniform.* Consider i.i.d. $Y_1, \dots, Y_n$ from the $\text{unif}[0, \theta]$. How can we make CDs for θ ?

(a) Show that $M_n = \max_{i \leq n} Y_i$ is sufficient, and that $(1+1/n)M_n$ is the natural unbiased estimator. Explain that the CD based on M_n , without needing to finetune constants or the like, is $C(\theta) = \Pr_\theta(M_n \geq m) = 1 - (m/\theta)^n$ for $\theta \geq m = M_{n,\text{obs}}$. Simulate, with small or moderate n , to see how this CD looks like. Show that the median confidence estimator is $2^{-1/n}M_n \doteq (1 - 0.693/n)M_n$.

(b) Show that the confidence density is $c(\theta) = C'(\theta) = nm^n/\theta^{n+1}$ for $\theta \geq m$, and that this matches the Bayesian posterior with the Jeffreys prior $\pi(\theta) \propto 1/\theta$.

(c) One may also build other CDs (though not as sharp as the one above). Using $\bar{Y}$ (which estimates $\frac{1}{2}\theta$), work out the CD $C^*(\theta) = \Pr_\theta(\bar{Y} \geq \bar{y}_{\text{obs}})$, via a normal approximation. Simulate say $n = 10$ datapoints and plot both CDs.

Ex. 8.9 *How many streetcars in San Francisco (again)?* In Ex. 7.5 we were Bayesians in San Francisco, estimating the number N of streetcars there, after having seen numbers 203, 157, 222, and using a prior $\pi(N)$. Here we're frequentists.

(a) With observed streetcars numbered $y_1, \dots, y_m$, taken to be i.i.d. and uniform on $1, \dots, N$, the Bayes posterior is $\pi(N | \text{data}) \propto \pi(N)(1/N)^m$ on $N \geq y_{\text{max}} = \max y_i$. Using the $\pi(N) \propto 1/N$ prior, Construct (again) these probabilities, pointwise and cumulatives, shown as dashed curves in Figure 8.3, for this simple $m = 3$ case.

(b) Argue that the CD solution is $C(N) = \Pr_N(Y_{\text{max}} \geq y_{\text{max,obs}}) = 1 - (y_{\text{max,obs}}/N)^m$ for $N \geq y_{\text{max,obs}}$. Compute this c.d.f., along with the confidence probabilities $c(N) = C(N) - C(N-1)$. Construct the plots of Figure 8.3, and comment on how close these

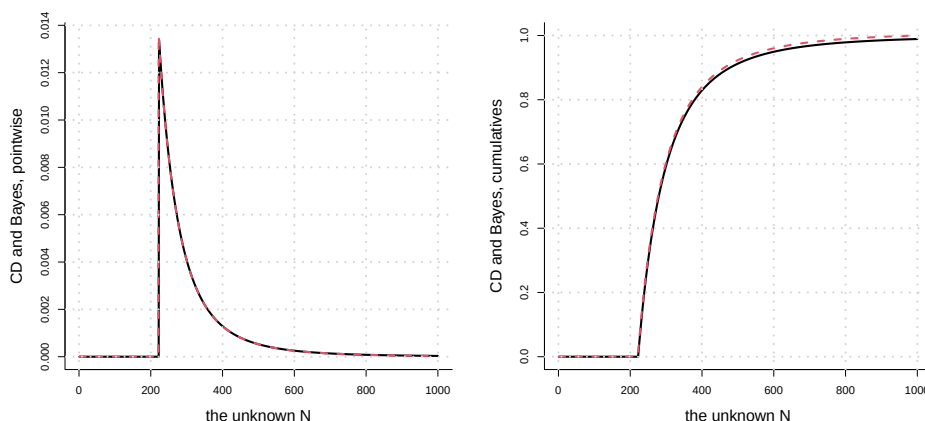


Figure 8.3: *Estimating the number of streetcars in San Francisco, after having seen numbers 203, 157, 222: pointwise probabilities (left panel), cumulative probabilities (right panel), for the frequentist (full line) and the Bayesian (dashed line). They agree very well.*

are to the Bayesian answers, already for $m = 3$. Compute and display also these curves after the first datum 203, and after having observed 203, 157.

Ex. 8.10 *Recipe Two: the normal approximation CD.* Applying Recipe One of Ex. 8.2 to the case of the estimator having a normal distribution leads as we saw there to a clear CD, provided the variance is known. But this is at least approximately so, for large classes of situations, as we've seen in Chs. 2 and 5.

(a) Suppose in general terms that $\hat{\theta}$ estimates θ , and that its distribution is approximately a $N(\theta, \kappa^2)$. Explain that $C(\theta) = \Phi((\theta - \hat{\theta})/\kappa)$ then is an approximate CD for θ . More formally, if $(\hat{\theta} - \theta_0)/\hat{\kappa} \rightarrow_d N(0, 1)$, at the true parameter θ_0 , show that $C(\theta, Y) = \Phi((\theta - \hat{\theta})/\hat{\kappa})$ has the property that it converges in distribution to the uniform, at θ_0 . In typical applications of these arguments, there is a $\sqrt{n}$ scaling in terms of an underlying sample size, with $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \tau^2)$, say, and $\hat{\kappa} = \hat{\tau}/\sqrt{n}$, with $\hat{\tau} \rightarrow_{\text{pr}} \tau$. So this is Recipe Two, the normal approximation CD, most typically of this type $\Phi(\sqrt{n}(\theta - \hat{\theta})/\hat{\tau})$.

Recipe Two

(b) Simulate a moderate or small dataset from a normal distribution. Compute and display two (approximate) CDs for the mean parameter ξ , (i) using the data mean, (ii) using the data median.

(c) We learn in Chs. 5, and in yet other chapters, that approximate normality is highly common, for large classes of estimators, typically along with consistent estimators for the variances. Also, the delta method implies approximate normality of smooth functions of background estimators, see Ex. 2.61, making in its turn approximate normality CDs easily available. For a simple illustration, suppose you throw your nearest die, which has probability p of giving a '6', until you get your first '6'. You carry out this geometric experiment $n = 10$ times, giving you counts $Y_1, \dots, Y_n$ equal to 1, 2, 17, 18, 20, 4, 3, 1, 15, 3. Use the normal approximation for $\bar{Y}$ to give an approximate CD for p . You may also compare this to what one achieves working with the exact distribution of $\bar{Y}$.

(d) Consider logistic regression, where $p(x) = \Pr(Y = 1 | x)$ is modelled as $H(x^t \beta)$, with the logistic transform. Build a CD for $p(x_0)$, the probability in question, for a given

individual with covariates x_0 . You may try this out, for a few mothers and their babies, in Story #6.

Ex. 8.11 *Recipe Three: from a pivot to a CD.* Suppose in general terms that μ is some parameter of interest, in a model for observations Y , and that a function $A = \text{piv}(\mu, Y)$ of the parameter and the data has the property that its distribution does not depend on the model parameters (in particular, therefore, not on μ , which might itself be a function of other model parameters). We call A a pivot, in more pedantic detail a pivot for the parameter μ . a pivot

(a) With $Y_1, \dots, Y_n$ independent from the normal (μ, σ^2) , let $R_n = \sum_{i=1}^n |Y_i - \bar{Y}|$ with the sample mean $\bar{Y}$. Show that $(\bar{Y} - \mu)/R_n$ is a pivot. Invent yet another pivot involving μ , with a different denominator.

(b) With two normal samples, say $X_1, \dots, X_m$ from $N(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_n$ from $N(\mu_2, \sigma_2^2)$, suppose $\rho = \sigma_1/\sigma_2$ is in focus. Show that $(V_1/V_2)/\rho$ is a pivot for ρ , where V_1 and V_2 are the interquartile ranges for the two datasets.

(c) Consider $Y_1, \dots, Y_n$ from the Cauchy model with density $(1/\pi)/\{1 + (y - \theta)^2\}$. Show that $R_n - \theta$ is a pivot, where $R_n = \frac{1}{2}(Q_{n,0.10} + Q_{n,0.90})$ is the average of the 0.10 and 0.90 quantiles.

(d) Back to generalities: consider a pivot $A = \text{piv}(\mu, Y)$ for μ in some model, increasing in μ . Assume the situation is continuous, not discrete, so that the pivot's distribution function K is continuous. Show that $C(\mu, y_{\text{obs}}) = 1 - K(\text{piv}(\mu, y_{\text{obs}}))$ is a proper CD. Recipe Three

(e) In clean cases we may derive the precise distribution for the pivot in question, but the CD recipe given above may be used also in more complicated setups, as long as $A = \text{piv}(\mu, Y)$ may be simulated. Make an illustration of this, with the ratio of standard deviations above. Suppose two normal datasets, both of size $n = 100$, lead to interquartile ranges $V_{1,\text{obs}} = 4.44$ and $V_{2,\text{obs}} = 3.33$. Construct and display $C(\rho)$ and $\text{cc}(\rho)$.

(f) Various constructions, involving large-sample approximations to the normal and to chisquares, lead to *approximate pivots*, and then again to approximate CDs and ccs. In this light, discuss how Recipe One, with $\Phi((\mu - \hat{\mu})/\hat{\kappa})$ can be seen via approximate pivots.

Ex. 8.12 *CDs from the t pivot.* In Ex. 8.2 we saw that the simple construction $C(\theta, y) = \Pr_{\theta}(\hat{\theta} \geq \hat{\theta}_{\text{obs}})$ gives a CD, in the case of one-dimensional setups with a well-defined estimator $\hat{\theta}$.

(a) For a normal sample from $N(\mu, \sigma^2)$, we see that several $\Pr_{\mu, \sigma}(Z \geq z_{\text{obs}})$ schemes work, in that the Z in question has a distribution depending on σ , but not μ . Attempt to work with $C^*(\mu, y) = \Pr_{\mu, \sigma}(\bar{Y} \geq \bar{y}_{\text{obs}})$ – and explain that it will not really work (unless σ is known).

(b) But of course there *are* natural CD constructions for μ here. What is needed is a *pivot*, say $A = \text{piv}(\mu, y)$, a function binding the focus parameter and data together in a way which makes its distribution not depend on the parameters. Study indeed

$$t_n = t_n(\mu, Y) = (\bar{Y} - \mu)/(\hat{\sigma}/\sqrt{n}),$$

with $\hat{\sigma}^2 = (n - 1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ the classical empirical variance. Pretend that you in all your cleverness have not seen this t_n before, and are unaware of its relation to a t distribution – but show that the distribution of t_n , call it K_n , does not depend on (μ, σ) .

(c) Then show that $C(\mu, y_{\text{obs}}) = K_n(t_n(\mu, y_{\text{obs}}))$ is a CD for μ . Even if you do not see the connection to the classic t of Student (1908), see Ex. 1.50, you may still carry

through this, by simulating $B = 10^5$ realisations of t_n , and use

$$C(\mu, y_{\text{obs}}) = K_n^*(t_n(\mu, y_{\text{obs}})) = \frac{1}{B} \sum_{j=1}^B I\{t_{n,j} \leq t_n(\mu, y_{\text{obs}})\}.$$

Show however that by all means K_n is a t_m , with $m = n - 1$, so the canonical CD for μ is and remains $C(\mu, y_{\text{obs}}) = G_m(\sqrt{n}(\mu - \bar{y}_{\text{obs}})/\hat{\sigma}_{\text{obs}})$, with G_m the c.d.f. for the t_m .

Recipes Four, Five for constructing CDs

Ex. 8.13 *Recipe Four: confidence curves via Wilks theorems.* Consider data from a parametric model, leading to the log-likelihood function $\ell_n(\theta)$, and that there is a focus parameter $\mu = g(\theta)$. We have seen likelihood profiling and Wilks theorems in Ch. 5, and know that the deviance $D_n(\mu) = 2\{\ell_{\max} - \ell_{\text{prof}}(\mu)\}$ has the property that $D_n(\mu_0) \rightarrow_d \chi_1^2$ at the true value $\mu_0 = g(\theta_0)$; see Ex. 5.13.

Recipe Four

(a) Recipe Four, utilising the Wilks theorems, is to form $\text{cc}(\mu, y) = \Gamma_1(D_n(\mu))$, with Γ_1 the c.d.f. for the χ_1^2 . Show that $\Pr_{\theta_0}(\text{cc}(\mu_0, Y) \leq \alpha) \rightarrow \alpha$ for each α , and explain that this makes $\text{cc}(\mu, y)$ an approximate confidence curve.

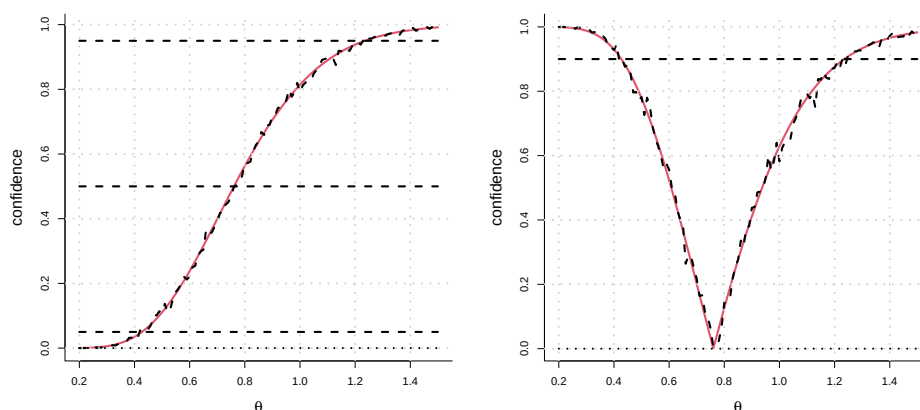


Figure 8.4: For the simple data example of Ex. 8.13: Left panel: confidence distribution $C(\theta)$, via simulations (black and wiggly curve) and via exact calculations (red and smooth curve); right panel: the two versions of the associated confidence curve $\text{cc}(\theta)$. From these we read off the median confidence estimate $\hat{\theta}_{0.50} = 0.76$, and the 90 percent confidence interval $[0.43, 1.24]$.

(b) For an illustration, consider the model $F(y, \theta) = y^\theta$ for observations on $[0, 1]$, where θ is an unknown positive parameter. Write down the log-likelihood function and find a formula for the maximum likelihood (ML) estimator $\hat{\theta}$. Use also theory of Ch. 5 to write down a normal approximation to the distribution of $\hat{\theta}$.

(c) Consider the data set

0.013 0.054 0.234 0.286 0.332 0.507 0.703 0.763 0.772 0.920

Estimate θ and compute the CD $C(\theta) = \Pr_{\theta}(\hat{\theta} \geq \hat{\theta}_{\text{obs}})$, along with the confidence curve $\text{cc}(\theta) = |1 - 2C(\theta)|$, (i) using simulations, (ii) using exact probability calculus. Reproduce a version of Figure 8.4.

(d) Supplement these two curves with approximations based (i) on the normal approximation for $\hat{\theta}$ and (ii) on the chi-squared approximation for the deviance.

(e) The CD and the cc are indeed close cousins, and we have learned in Ex. 8.3 how to get the latter from the former. Show now that we can construct the former from the latter, via $C(\mu) = \frac{1}{2} - \frac{1}{2}cc(\mu)$ for $\mu \leq \hat{\mu}_{0.50}$ and $\frac{1}{2} + \frac{1}{2}cc(\mu)$ for $\mu \geq \hat{\mu}_{0.50}$. This is particularly useful for the deviance based ccs.

Ex. 8.14 *Median age for men and women in Roman Era Egypt.* In Story #50 we work with a rare dataset of lifetimes for 82 men and 59 women in Roman Era Egypt, a century B.C. For the present illustration of constructing confidence curves via log-likelihood profiling, take these lifetimes T_i to have arisen from Weibull distributions, with c.d.f.s $F_m(t) = 1 - \exp\{-(t/a_m)^{b_m}\}$ for men and $F_w(t) = 1 - \exp\{-(t/a_w)^{b_w}\}$ for women.

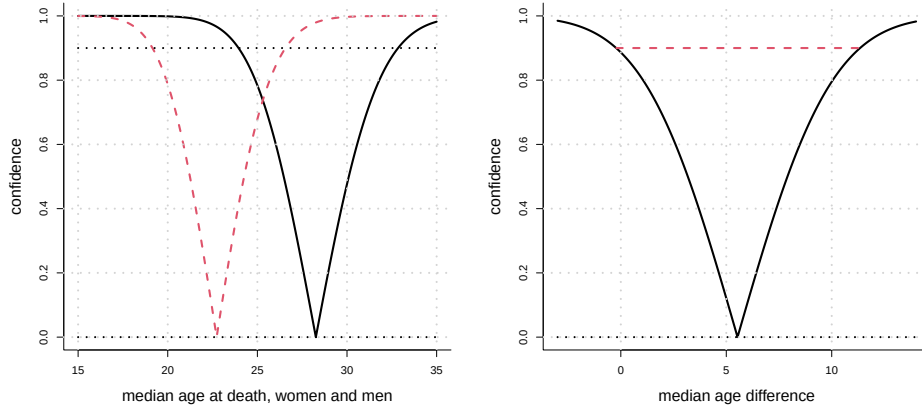


Figure 8.5: *Lifetimes in Roman Era Egypt, a century B.C.:* Left panel: confidence curves for the median age μ_w and μ_m , for women and men, via log-likelihood-profiling from fitting separate Weibull distributions, with 90 percent confidence intervals $[19.95, 26.56]$ and $[23.95, 32.84]$. Right panel: confidence curve for the median age difference $d = \mu_m - \mu_w$, via log-likelihood profiling, and 90 percent confidence interval $[-0.21, 11.34]$.

(a) For a Weibull, with parameters (a, b) , show that the q level quantile becomes $\mu(q) = F^{-1}(q) = ac(q)^{1/b}$, with $c(q) = -\log(1 - q)$. In particular, the median is $a(\log 2)^{1/b}$.

(b) For the men and women separately, write down the log-likelihood functions $\ell_m(a_m, b_m)$ and $\ell_w(a_w, b_w)$, and then carry out profiling to compute the deviance functions, for the medians μ_m and μ_w . Use Recipe Four, from Ex. 8.13, to construct the confidence curves, and make a version of Figure 8.5, left panel; also, read off 90 percent confidence intervals.

(c) It appears indeed that men had longer lives than women, in Egypt 2100 years ago; see Story #50 for more discussion and details. For this exercise, carry out log-likelihood profiling, for the median difference $d = \mu_m - \mu_w$, from the four-parameter model with the two Weibulls. Compute the deviance function and then the confidence curve $cc(d)$, leading to Figure 8.5, right panel. Read off a 90 percent interval for d .

(d) Constructing the confidence curves above involved the somewhat laborious computation of log-likelihood profiles. Compare these curves with the easier ones, computationally speaking, based on normal approximations.

another
illustration:
Story #52(d)

(e) Implementing Recipe Four, in a given setup, necessitates computing the $\ell_{\text{prof}}(\mu)$ for a gridded set of μ . For given μ_{fix} , maximising over all $(\theta_1, \dots, \theta_p)$ such that $\mu(\theta) = \mu_{\text{fix}}$ is helped significantly if a formula can be found, of the type $\theta_1 = h(\mu_{\text{fix}}, \theta_2, \dots, \theta_p)$, in which case numerical optimisation is carried out over all $(\theta_2, \dots, \theta_p)$. In more demanding cases one may turn to the generic trick of maximising $\ell(\theta) - \lambda\{\mu(\theta) - \mu_{\text{fix}}\}^2$, over all $(\theta_1, \dots, \theta_p)$, with a high value for λ . For the Weibull case above, we could use $\mu = a(\log 2)^{1/b}$ for the median to give us the $\ell_{\text{prof}}(\mu)$ relatively easily. Carry out the same task, using Recipe Four to create $\text{cc}(\mu)$ for the men and the women, using the $\text{Gam}(a', b')$ instead. The operative point is that $\mu = G^{-1}(\frac{1}{2}, a', b')$ is not a simple formula, so employ the λ trick.

Ex. 8.15 *The Wilks confidence recipe for the normal model.* Consider the familiar setup with i.i.d. data $Y_1, \dots, Y_n$ from the $N(\xi, \sigma^2)$ model. We have handled the log-likelihood profiles and accompanying Wilks theorems approximations in Ex. 5.22, and can apply results for constructing confidence curves. In the following, let as usual $\hat{\xi}$ and $\hat{\sigma}^2 = Q_0/(n-1)$ be the sample mean and sample variance, with $Q_0 = \sum (Y_i - \bar{Y})^2$. Note that the ML estimator is Q_0/n , a close relative.

(a) Starting with the ξ parameter, explain first that $C(\xi, \text{data}) = G_m(\sqrt{n}(\xi - \bar{Y})/\hat{\sigma})$ is the natural exact CD, with G_{df} the c.d.f. for the t_{df} , with $\text{df} = n-1$ degrees of freedom. Then use the exercise pointed to to show that the Wilks recipe leads to

$$\text{cc}_w(\xi) = \Gamma_1(n \log(1 + (n-1)^{-1} K_n(\xi))),$$

with $K_n(\xi) = n(\xi - \hat{\xi})^2/\hat{\sigma}^2$. Explain also that at the true value of ξ , we have $\text{cc}_w(\xi) = d \log(1 + (n-1)^{-1} t_{\text{df}}^2)$, and argue from this that $\text{cc}_w(\xi) \rightarrow_d \chi_1^2$ for growing n .

(b) Then considering σ , explain that the natural exact CD is $C(\sigma, \text{data}) = 1 - \Gamma_{\text{df}}(\text{df} \hat{\sigma}^2 / \sigma^2)$, with Γ_{df} the c.d.f. for χ_{df}^2 . With $h(x) = x - \log x - 1$, show that the Wilks recipe leads to $\text{cc}_w(\sigma) = \Gamma_1(nh(\hat{\sigma}^2/\sigma^2))$.

(c) We learn that the general Wilks based strategy leads to approximate CDs. Compare these with the exact ones, for ξ and for σ , in the present normal case. One finds that the approximations work well even for small sample sizes.

Ex. 8.16 *Recipe Four: with a Beta correction on top.* As explained in Ex. 8.13, Recipe Four uses $\text{cc}_n(\mu, y) = \Gamma_1(D_n(\mu, y))$, which is guaranteed to work well for moderate to large n in view of the Wilks theorem. Here is a Beta correction method to achieve higher accuracy.

(a) The method is perfect provided $U = \text{cc}_n(\mu_0, Y)$ has the uniform distribution, with Y generated from the model with the true focus parameter being μ_0 . This can be checked via simulation, drawing say 10^5 values U_j of the U . If the situation is not perfect, fit these to the $\text{Beta}(a, b)$ distribution, using ML, with the fitted (a_0, b_0) at least close to $(1, 1)$. Motivate the Beta correction

$$\text{cc}_n^*(\mu) = \text{Be}(\text{cc}(\mu), a_0, b_0).$$

Ideally this Beta modification ought to be carried out for each μ on a grid, but in many cases the (a_0, b_0) are approximately the same for a reasonable interval of values, or even the very same, in pivotal situations when the distribution of U does not depend on parameters.

(b) For an illustration, examine the setup with $Y_1, \dots, Y_n$ from $\text{Expo}(\theta)$. The ML is $\hat{\theta} = 1/\bar{Y}$. With $W_n = \theta/\hat{\theta}$, show $D_n(\theta) \sim 2nh(W_n)$, with $h(w) = w - 1 - \log w$. Note

the BCCC,
Beta corrected
confidence
curve

that this matches and the below answers the corresponding questions for the $cc_w(\sigma)$ part of Ex. 8.15. Work out that the distribution of $U_n = \Gamma_1(D_n(\theta))$ is

$$G_n(u) = \Pr(\Gamma_1(2nh(W_n)) \leq u) = \Pr(h(W_n) \leq \Gamma_1^{-1}(u)/(2n)) = \Pr(c_n \leq \chi_{2n}^2/(2n) \leq d_n),$$

with $c_n < 1 < d_n$ the two solutions to $h(w) = \Gamma_1^{-1}(u)/(2n)$. Numerically compute and display the G_n for U_n , for a few sample sizes. Next simulate 10^6 of U_n to fit these to $\text{Beta}(a_0, b_0)$ via ML; these parameters are fairly close to (1.1). The Beta corrected confidence curve is $cc_n^*(\theta) = \text{Be}(cc_n(\theta), a_0, b_0)$. Explain that the c.d.f. for $U_n^* = cc_n^*(\theta)$ can be coded up and computed:

$$G_n^*(u) = \Pr[\text{Be}(\Gamma_1(cc_n(\theta)), a_0, b_0) \leq u] = \Pr[\Gamma_1(2nh(W_n)) \leq \text{Be}^{-1}(u, a_0, b_0)],$$

i.e. almost as for G_n , but with the Beta repaired $\text{Be}^{-1}(u, a_0, b_0)$ in lieu of u : $G_n^*(u) = G_n(\text{Be}^{-1}(u, a_0, b_0))$. Programme these functions and discover that G_n^* is closer to u than is G_n . You should find that the Beta correction works very well for sample sizes as low as $n = 3$ and almost $n = 2$.

(c) For another concrete example, study the Wilks for ξ in the normal model, as in Ex. 8.15. With the deviance formula found there, show that $U_n = cc(\xi, Y)$ has c.d.f.

$$\begin{aligned} G_n(u) &= \Pr[1 + t_n^2/(n-1) \leq \exp(u/n)] \\ &= \Pr[t_n^2 \leq (n-1)(\exp(u/n) - 1)] = F_{1, \text{df}}((n-1)(\exp(u/n) - 1)). \end{aligned}$$

We fit this distribution to the $\text{Beta}(a_0, b_0)$, with (a_0, b_0) close to (1, 1) for even small n . The Beta repair is $cc_n^*(\xi) = \text{Be}(cc_n(\xi), a_0, b_0)$. Code up these functions and demonstrate that its c.d.f. $G_n^*(u)$ is rather closer to u than is $G_n(u)$, for small n .

Ex. 8.17 Profiling for the correlation coefficient. Consider i.i.d. pairs (X_i, Y_i) from the binormal, with means ξ_1, ξ_2 , variances σ_1^2, σ_2^2 , and with focus on the correlation ρ .

(a) (xx we provide a bit more details and help below, for the algebraic efforts. xx) Write down the full log-likelihood, verifying also that the ML estimators are the usual $\bar{X}, \bar{Y}$, $s_1^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, $s_2^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$, and $\hat{\rho} = R_n = s_{12}/(s_1 s_2)$, with $s_{12} = n^{-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$. Work out that the profiled log-likelihood can be written

$$\ell(\rho) = -\frac{1}{2}n\{2\log(1 - \rho\hat{\rho}) - \log(1 - \rho^2) + 2 + \log(s_1^2) + \log(s_2^2)\},$$

and convince yourself that it is maximised at $\rho = R_n$. Then derive the deviance function

$$D(\rho) = n\{2\log(1 - \rho\hat{\rho}) - \log(1 - \rho^2) - \log(1 - \hat{\rho}^2)\} = n \log \frac{(1 - \rho\hat{\rho})^2}{(1 - \rho^2)(1 - \hat{\rho}^2)}.$$

(b) Now detail the specifics of the Recipe Four based $cc(\rho) = \Gamma_1(D(\rho))$ and its Beta correction $cc^*(\rho) = \text{Be}(cc(\rho), a_0, b_0)$. Try this out for simulated datasets of sizes 5, 10, 15, 20, and for these compute also $q_n = \max |G_n(u) - u|$ and the considerable smaller $q_n^* = \max |G_n^*(u) - u|$, in terms of the c.d.f.s for $U_n = cc(\rho, \text{data})$ and $U_n^* = cc^*(\rho, \text{data})$, with data taken from the binormal with this ρ . Compare also with the confidence curve $cc_{\text{zetaeta}}(\rho)$ constructed using $(n-3)^{1/2}(\hat{\zeta} - \zeta) \approx_d N(0, 1)$, with Fisher's zeta transformation; see Ex. 2.55. Comment: using the basic principle and mechanisms of the log-likelihood profile, we have matched the accurate performance of the zeta transform, without needing to know or care about the limiting distributions and the mathematical efforts associated with the zeta.

Ex. 8.18 *Recipe Five: CDs via approximate pivots and t-bootstrapping.* Consider a parametric model $f(y, \theta)$ for data, with a model parameter of length say p . Suppose there is a focus parameter $\mu = g(\theta)$, estimated as $\hat{\mu} = g(\hat{\theta})$, for which we need a CD.

(a) Suppose first that $\hat{\mu} - \mu$ has some distribution, say G_0 , not depending on θ . Under this simple pivotal assumption for ‘estimator minus estimand’, show that

$$C(\mu) = \Pr_{\theta}(\hat{\mu} \geq \hat{\mu}_{\text{obs}}) = 1 - G_0(\hat{\mu}_{\text{obs}} - \mu).$$

As indicated in e.g. Ex. 8.6, this can be computed even without knowing the form of G_0 , through simulation of many realisations of $\hat{\mu}^* - \hat{\mu}$, where the $\hat{\mu}^*$ is computed from a dataset drawn from the estimated distribution at $\hat{\theta}$. Explain that this recipe works, even if G_0 is nonsymmetric and not centred well at zero. Simulating the distribution of $\hat{\mu} - \mu$ at different points in the θ parameter space may also be helpful for checking the assumption needed for the $C(\mu)$ constructed here to be a clear CD.

(b) For a special and simpler case, if $\hat{\mu} \sim N(\mu, \kappa^2)$, to a good approximation, with known or well estimated κ , show that the general recipe above leads to

$$C(\mu) = \Pr_{\theta}(\mu + \kappa N \geq \hat{\mu}_{\text{obs}}) = \Pr(N \geq (\hat{\mu} - \mu)/\kappa) = \Phi((\mu - \hat{\mu})/\kappa),$$

and argue that this is really a CD. Note that this requires $Z = (\hat{\mu} - \mu)/\kappa$ having distribution equal to or close to a standard normal, regardless of where θ is in its parameter space.

(c) Often the standard deviation in the normal approximation is not that sharply estimated from the available data. Consider a Student type ratio $t = (\hat{\mu} - \mu)/\hat{\kappa}$, with some appropriate scale estimator $\hat{\kappa}$. Assume first that t really is a pivot, i.e. that its distribution G is independent or nearly independent of where θ is in its parameter space. Show that

$$C(\mu) = \Pr_{\theta}((\hat{\mu} - \mu)/\hat{\kappa} \geq (\hat{\mu}_{\text{obs}} - \mu)/\hat{\kappa}_{\text{obs}}) = 1 - G((\hat{\mu}_{\text{obs}} - \mu)/\hat{\kappa}_{\text{obs}})$$

Recipe Five

is a CD. If G is not known, or too difficult to derive, use simulations, of $t^* = (\hat{\mu}^* - \hat{\mu}_{\text{obs}})/\hat{\kappa}^*$, via datasets simulated at position $\hat{\theta}_{\text{obs}}$. We call this a CD computed from t-bootstrapping.

(d) The previous recipe works well if $t = (\hat{\mu} - \mu)/\hat{\kappa}$ is close to pivotal, i.e. its distribution is nearly constant over the parameter region. In other cases we may take the t-bootstrapping argument one step further. Write for emphasis $\theta = (\mu, \gamma)$, perhaps in a reparametrisation, where μ is in focus and γ is of length $p-1$. The t has some distribution, depending on θ , and we write $\Pr_{\theta}(t \leq u) = G(u, \mu, \gamma)$. Show that

$$H(\mu, \gamma) = \Pr_{\mu, \gamma}((\hat{\mu} - \mu)/\hat{\kappa} \geq (\hat{\mu}_{\text{obs}} - \mu)/\hat{\kappa}_{\text{obs}}) = 1 - G((\hat{\mu}_{\text{obs}} - \mu)/\hat{\kappa}_{\text{obs}}, \mu, \gamma).$$

This is not necessarily a CD, in the strict sense, as this probability may depend not only on μ but also on aspects of γ . Often the distribution of t is approximately the same, though, in a neighbourhood around the true value. Argue that this leads to

$$C^*(\mu) = 1 - \hat{G}((\hat{\mu}_{\text{obs}} - \mu)/\hat{\kappa}_{\text{obs}}, \mu) \quad \text{where } \hat{G}(u, \mu) = G(u, \mu, \hat{\gamma}).$$

Such an estimated distribution can be computed via bootstrapping, i.e. simulated datasets at position $\hat{\theta}$ in the parameter space. With B such simulated datasets, leading to simulated values $\hat{\theta}^*, \hat{\mu}^*, \hat{\kappa}^*$, and hence $t^* = (\hat{\mu}^* - \hat{\mu}_{\text{obs}})/\hat{\kappa}^*$.

Recipe Six: CDs in exponential families

Ex. 8.19 *Recipe Six: for exponential family models.* We have worked with the general exponential family in previous chapters, see Ex. 1.57. In particular we learned in Ex. 4.33 that there are uniformly optimal tests, for individual parameters in such models. The same holds in the present framework of CDs. Suppose data stem from a model of the form $f(y, a, b) = \exp\{aU(y) + b^t V(y)\}h(y)$, with U one-dimensional and V of dimension say p . The CD recipe for a is

Recipe Six

$$C^*(a) = \Pr_a\{U(Y) \geq u_{\text{obs}} \mid V(Y) = v_{\text{obs}}\}.$$

This construction is actually broadly optimal, in a power risk function sense, for a broad class of loss functions; see Ex. 8.41. At any rate we can start working with this definition and see how it applies in various situations.

there is a
Neyman–
Pearson
optimality
theory for CDs,
see [Schweder
and Hjort
\(2016, Ch. 5\)](#)

- (a) Verify from arguments in Ex. 4.33 that $C^*(a)$ indeed depends only on a , not on b .
- (b) For an illustration, consider the pair of exponentials of Ex. 4.28. To avoid confusion with the parametrisation, use now $X \sim \text{Expo}(\theta)$, $Y \sim \text{Expo}(\theta + \delta)$, with sum $Z = X + Y$. Show that the joint density is indeed of exponential form, and that the recipe leads to

$$C^*(\delta) = \Pr_\delta(Y \leq y_{\text{obs}} \mid Z = z_{\text{obs}}) = \frac{1 - \exp(-\delta y_{\text{obs}})}{1 - \exp(-\delta z_{\text{obs}})}.$$

Find the positive confidence pointmass at $\delta = 0$.

- (c) Suppose there are m independent pairs of such exponentials, with $X_i \sim \text{Expo}(\theta_i)$, $Y_i \sim \text{Expo}(\theta_i + \delta)$, and sums $Z_i = X_i + Y_i$. We need a CD for the difference parameter δ . Show that the joint density of the $2m$ variables is on the exponential form, and that the resulting CD must be of the form

$$C^*(\delta) = \Pr_\delta(U \leq u_{\text{obs}} \mid Z_1 = z_{1,\text{obs}}, \dots, z_{m,\text{obs}}),$$

with $U = \sum_{i=1}^m Y_i$. There is no clear formula for this conditional distribution, but show that $Y_i \mid z_i$ has density $\delta \exp(-\delta y_i) / \{1 - \exp(-\delta z_i)\}$ for $y_i \in [0, z_i]$. To show how the CD can be computed, via simulations, suppose as in Ex. 4.28 that the data are the three pairs (0.927, 0.819), (1.479, 0.408), (3.780, 1.311). In that exercise we worked with the optimal test for $\delta = 0$ vs. $\delta > 0$, and needed only the null distribution of U given the three sums z_1, z_2, z_3 , i.e. where $\delta = 0$. Now we need to tabulate this conditional distribution also for each $\delta > 0$, however.

Ex. 8.20 *Optimal CD for Gamma shape.* Consider $Y_1, \dots, Y_n$ i.i.d. from $\text{Gam}(a, b)$.

- (a) Give an expression for the log-likelihood function $\ell(a, b)$, and verify that $(U, V) = (\sum_{i=1}^n \log Y_i, \sum_{i=1}^n Y_i)$ is sufficient. Deduce that the optimal CD for the two parameters must be

$$C_1(a) = \Pr_a(U \geq u_{\text{obs}} \mid V = v_{\text{obs}}) \quad \text{and} \quad C_2(b) = \Pr_b(V \geq v_{\text{obs}} \mid U = u_{\text{obs}}).$$

- (b) To find the first of these, show that $Y_i = VD_i$, with $(D_1, \dots, D_n) \sim \text{Dir}(a, \dots, a)$ being independent of V . Argue from this that

$$C_1(a) = \Pr_a(n \log v_{\text{obs}} + \sum_{i=1}^n \log D_i \geq u_{\text{obs}}) = \Pr(W_a \geq u_{\text{obs}} - n \log v_{\text{obs}}),$$

where $W_a = \sum_{i=1}^n \log D_i$. This can then be computed via Dirichlet simulation for each gridded point a .

(c) For an illustration, travel back to Old Egypt, with lifelengths for 82 men and 59 women a century B.C., see Story #50. Modelling these as gamma distributions, compute and display the optimal confidence curves $cc(a_m)$ and $cc(a_w)$, for the two shape parameters. Note that this may be done fully independently of any maximum likelihood approximation theory. Compare nevertheless with the normal approximations, and comment.

Ex. 8.21 *Optimal CDs in the binormal model.* Consider the stylised setup where $(\hat{a}, \hat{b})$ is binormal, with means (a, b) , variances $(1, 1)$, and known correlation ρ . (xx spell out. we do get the natural answer $C(a) = \Pr_a(\hat{a} \geq \hat{a}_{\text{obs}} | \hat{b})$. then multivariate. xx)

(a) Use Recipe Six to find the optimal CD for a .

(b)

(c)

Ex. 8.22 *Bayesian posteriors as approximate CDs.* Consider a setup with data y from a model with parameter $\theta = (\theta_1, \dots, \theta_p)$, and with $\mu = \mu(\theta_1, \dots, \theta_p)$ a focus parameter. A CD for μ has the property $\Pr_\theta\{C^{-1}(0.05, Y) \leq \mu \leq C^{-1}(0.95, Y)\} = 0.90$, etc., as with (8.1), thus delivering confidence intervals with the right coverage. This is also akin to how Bayesian posterior distributions are used. If a Bayesian prior for θ leads to a posterior for θ , and hence for a cumulative $B(\mu | y_{\text{obs}})$, then the Bayesian can read off $[B^{-1}(0.05, y_{\text{obs}}) \leq \mu \leq B^{-1}(0.95, y_{\text{obs}})]$, as with Ex. 7.13. A question of interest and relevance also in Bayesian contexts is whether such intervals make sense also in the frequentist sense.

(a) For a simple start example, consider $Y_1, \dots, Y_n$ which given θ are i.i.d. from the $\text{Pois}(\theta)$, and with a prior $\theta \sim \text{Gam}(a, b)$; see Ex. 7.2. Show that the posterior cumulative for θ becomes $B_n(\theta | \text{data}) = G(\theta, a + n\bar{y}_{\text{obs}}, b + n)$, in terms for the cumulative Gamma, and with $\bar{y}_{\text{obs}}$ the observed data average. Let $\hat{\theta}_B$ and $\hat{\tau}_B$ be the posterior mean and standard deviation. Assume now that data $Y_1, Y_2, \dots$ come from the $\text{Pois}(\theta_0)$, for a certain θ_0 . Show that

$$\begin{aligned} B_n(\theta_0 | Y_1, \dots, Y_n) &= \Pr(\theta \leq \theta_0 | Y_1, \dots, Y_n) = G(\theta_0, a + n\bar{y}, b + n) \\ &\doteq \Phi((\theta_0 - \hat{\theta}_B)/\hat{\tau}_B) \rightarrow_d \Phi(N(0, 1)) \sim \text{unif}, \end{aligned}$$

with probability 1.

(b) For a similar adventure, start with the $\text{Beta}(a, b)$ prior for a binomial probability θ . Show that the posterior cumulative for θ becomes $B_n(\theta | \text{data}) = \text{Be}(\theta, a + y, b + n - y)$, in terms of the Beta cumulative. Assuming that Y_n is really from a binomial with some true θ_0 , show that

$$\begin{aligned} B_n(\theta_0 | Y_n) &= \Pr(\theta \leq \theta_0 | Y_n) = \text{Be}(\theta_0, a + Y_n, b + n - Y_n) \\ &\doteq \Phi((\theta_0 - \hat{\theta}_B)/\hat{\tau}_B) \rightarrow_d \Phi(N(0, 1)) \sim \text{unif}, \end{aligned}$$

with probability 1. Discuss how these findings relate to those of Ex. 7.14.

(c) Consider then the general parametric situation, supposing $Y_1, \dots, Y_n$ being i.i.d. from the density $f(y, \theta)$, with $\mu = \mu(\theta)$ a one-dimensional focus parameter. Let θ_0 and $\mu_0 = \mu(\theta_0)$ denote the true parameters. As with Recipe Two, see Ex. 8.10, there is convergence to the standard normal of the standardised $\sqrt{n}(\hat{\mu}_{\text{ml}} - \mu_0)/\hat{\kappa}$, say, with the ML estimator for μ and $\hat{\kappa}^2$ consistently estimating $\kappa^2 = c^t J^{-1} c$. Now use the Bernshtein–von Mises

a Bernshtein–
von Mises
theorem for
Bayesian
posteriors and
CDs

check with delta
method in Ch. 2

theorem setup of Ex. 7.39 to explain that $\sqrt{n}(\mu - \hat{\mu}_{\text{ml}}) | \text{data}$ tends in distribution to $N(0, \kappa^2)$. The Bayesian posterior has a c.d.f. $B_n(\mu | \text{data})$. Show that at the true value,

$$\begin{aligned} B_n(\mu_0 | \text{data}) &= \Pr(\sqrt{n}(\mu - \hat{\mu}_{\text{ml}})/\hat{\kappa} \leq \sqrt{n}(\mu_0 - \hat{\mu}_{\text{ml}})/\hat{\kappa} | \text{data}) \\ &= \mu(\sqrt{n}(\mu_0 - \hat{\mu}_{\text{ml}})/\hat{\kappa}) + \varepsilon_n, \end{aligned}$$

with $\varepsilon_n \rightarrow_{\text{pr}} 0$. Deduce that $B_n(\mu)$ is a CD in the large-sample sense. In this sense we may think of any sensible Bayesian posterior distribution, in regular parametric models, as Recipe Seven for creating a CD.

Bayes: a
cautious Recipe
Seven

Ex. 8.23 *Confidence distribution for the ratio of explained variation.* Organise the classic linear regression model as $Y_i = \beta_0 + x_i^t \beta + \varepsilon_i$, with the covariates being centred, i.e. $\sum_{i=1}^n x_{i,j} = 0$ for $j = 1, \dots, p$, giving then β_0 a clear interpretation as the overall mean. We also take the ε_i noise terms to be i.i.d. $N(0, \sigma^2)$. Theory was developed in Ex. 4.38 to estimate the fraction $\rho = \beta^t \Sigma_n \beta / (\beta^t \Sigma_n \beta + \sigma^2)$ of the variance explained via the covariates. With $\hat{\mu}_i = \hat{\beta}_0 + x_i^t \hat{\beta}$, we saw there that the statistic

$$R^2 = \text{coefficient of determination} = 1 - \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 / \sum_{i=1}^n (Y_i - \bar{Y})^2$$

can be written as $\hat{\lambda}/(\hat{\lambda} + 1)$, with $\hat{\lambda} = \hat{\beta}^t \Sigma_n \hat{\beta} / \hat{\sigma}^2$, and with $F = n\hat{\lambda}/p \sim F(p, m, n\lambda)$, the non-central F , with $m = n - (p + 1)$ the degrees of freedom for $\hat{\sigma}^2$.

(a) Explain that this leads to a full CD for ρ ,

$$C(\rho) = \Pr_\rho(\hat{\lambda} \geq \hat{\lambda}_{\text{obs}}) = \Pr_\rho(F \geq n\hat{\lambda}_{\text{obs}}/p) = 1 - F(n\hat{\lambda}_{\text{obs}}/p, p, m, n\rho/(1 - \rho)).$$

(b) Explain that the above may be used for a subset A of the covariates at a time, with a modified interpretation for the σ_A^2 part in say $\rho_A = \beta_A^t \Sigma_{A,n} \beta_A / (\beta_A^t \Sigma_{A,n} \beta_A + \sigma_A^2)$. Spell out what this leads to for a single covariate. This is used in Story #6 to assess how much of the birthweight variance is explained by one covariate at a time – the answer is ‘rather little’, for each of ‘smoking’, ‘mother weight’, etc., but then an after all sensible 15 percent when taking all five covariates into account.

Ex. 8.24 *The distance between two normals.* Consider normal densities $N(\xi_0, \sigma^2)$ and $N(\xi_1, \sigma^2)$, for groups 0 and 1, with common σ . A natural measure of their interdistance is $\delta = |\xi_1 - \xi_0|/\sigma$, being invariant with respect to scale changes. For the following, check also with the two-sample t test of Ex. 4.4; the CD for δ gives useful additional information.

the
Mahalanobis
distance (in
dimension 1)

(a) Suppose Y is drawn from one of these two densities, and one classifies this to the distribution closest in mean; that is, one says group 0 if $|Y - \xi_0| < |Y - \xi_1|$, and vice versa. Taking the two groups to be equally likely, show that this classification rule has error rate $\Phi(-\frac{1}{2}\delta)$. You may show this via $\varepsilon = \frac{1}{2}\varepsilon_{0,1} + \frac{1}{2}\varepsilon_{1,0}$, with the conditional error rates given respectively group 0 and group 1.

classification
error rate

(b) Assume in general terms that $\hat{d} \sim N(d, c^2\sigma^2)$, with c known, and that $\delta = |d|/\sigma$ is the parameter of interest. For the case of σ known, using $\hat{\delta}_0 = |\hat{d}|/\sigma$, show that $\hat{\delta}_0^2 \sim c^2\chi_1^2(\delta^2/c^2)$, and that Recipe One then leads to the CD

$$C_0(\delta) = \Pr_\delta(\hat{\delta}_0^2 \geq \hat{\delta}_{0,\text{obs}}^2) = 1 - \Gamma_1(\hat{\delta}_{0,\text{obs}}^2/c^2, \delta^2/c^2),$$

with the c.d.f. for the noncentral chi-squared. For σ unknown, assume there is an estimator with $\hat{\sigma}^2 \sim \sigma^2\chi_{\text{df}}^2/\text{df}$, independent of $\hat{d}$. With $\hat{\delta} = |\hat{d}|/\hat{\sigma}$, derive the CD

$$C(\delta) = \Pr_\delta(\hat{\delta}^2 \geq \hat{\delta}_{\text{obs}}^2) = 1 - F_{1,\text{df}}(\hat{\delta}_{\text{obs}}^2/c^2, \delta^2/c^2),$$

with the c.d.f. for the noncentral $F_{1,\text{df}}$. Discuss the point-mass $C(0)$ at zero and its interpretation.

(c) Suppose we have i.i.d. $X_1, \dots, X_{n_0}$ and $Y_1, \dots, Y_{n_1}$ from groups 0 and 1. For the Mahalanobis distance, use $\hat{\delta} = |\bar{Y} - \bar{X}|/\hat{\sigma}$, with the two-sample based pooled variance $\hat{\sigma}^2 = (Q_0 + Q_1)/(n_0 + n_1 - 2)$, in terms of $Q_0 = \sum_{i=1}^{n_0} (X_i - \bar{X})^2$ and $Q_1 = \sum_{i=1}^{n_1} (Y_i - \bar{Y})^2$. Derive the CD, of the form above, with $c^2 = 1/n_0 + 1/n_1 = N/(n_0 n_1)$, and $N = n_0 + n_1$:

$$C(\delta) = 1 - F_{1,\text{df}}((n_0 n_1 / N) \hat{\delta}^2, (n_0 n_1 / N) \delta^2).$$

Explain how you can find the median confidence estimate for δ .

Ex. 8.25 *The Mahalanobis distance between two multidimensional normals.* For two multinormal populations, say $N_p(\xi_0, \Sigma)$ and $N_p(\xi_1, \Sigma)$, the Mahalanobis distance between them is $\delta = \{(\xi_1 - \xi_0)^t \Sigma^{-1} (\xi_1 - \xi_0)\}^{1/2}$, in generalisation of the one-dimensional distance worked with in Ex. 8.24.

(a) Seeing a y from either f_0 or f_1 , how can guess from which? Work with the natural method which says ‘0’ if $\|y - \xi_0\| < \|y - \xi_1\|$ (and vice versa), with the Σ dictated distance, i.e. with $\|b - a\| = \{(b - a)^t \Sigma^{-1} (b - a)\}^{1/2}$. Via some algebra, demonstrate that choosing ‘0’ or ‘1’ is matter of whether $W = (\xi_1 - \xi_0)^t \Sigma^{-1} (y - \bar{\xi})$ is negative or positive. Further, show that W is normal with variance δ^2 and mean value $-\frac{1}{2}\delta$ or $\frac{1}{2}\delta$, for the two choices. From this, show that the error rate is $\Phi(-\frac{1}{2}\delta)$, independent of dimension.

(b) With samples $X_1, \dots, X_{n_0}$ and $Y_1, \dots, Y_{n_1}$ from these, how can we make inference for this distance? We start studying the case of Σ known or well estimated. Extending arguments from Ex. 8.24, assume $\hat{d} \sim N_p(d, c^2 \Sigma)$, with $\delta^2 = d^t \Sigma^{-1} d$. Show that $\hat{\delta}_0^2 = \hat{d}^t \Sigma^{-1} \hat{d}$ has the distribution $c^2 \chi_p^2(\delta^2/c^2)$, and that this leads to $C_0(\delta) = 1 - \Gamma_p(\hat{\delta}_0^2/c^2, \delta^2/c^2)$. Translate this to the case of two samples, with $c^2 = 1/n_0 + 1/n_1$, and conclude, with $N = n_0 + n_1$, that

$$C_0(\delta) = 1 - \Gamma_p(((n_0 n_1)/N) \hat{\delta}_0^2, ((n_0 n_1)/N) \delta^2).$$

With Σ unknown, use the pooled estimator $\hat{\Sigma} = g_0 \hat{\Sigma}_0 + g_1 \hat{\Sigma}_1$, with $g_0 = (n_0 - 1)/(N - 2)$ and $g_1 = (n_1 - 1)/(N - 2)$. For $\hat{\delta}^2 = (\bar{Y} - \bar{X})^t \hat{\Sigma}^{-1} (\bar{Y} - \bar{X})$ it can then be shown, via multinormal analysis, that there is a clear connection to a noncentral F ,

$$\frac{n_0 n_1}{N} \frac{N - p - 1}{N - 2} \frac{1}{p} \hat{\delta}^2 \sim F_{p, N-p-1}((n_0 n_1 / N) \delta^2).$$

Allowing ourselves to take this for granted, show that this leads to the Mahalanobis CD

$$C(\delta) = 1 - F_{p, N-p-1}\left(\frac{n_0 n_1}{N} \frac{N - p - 1}{N - 2} \frac{1}{p} \hat{\delta}^2, \frac{n_0 n_1}{N} \delta^2\right).$$

Show that for growing n_0, n_1 , this CD becomes close to the simple C_0 above. See Story #53 to see the Mahalanobis distance, analysed in this fashion, between Bach and Reger fugue themes, in six-dimensional space.

CDs for quantiles

Ex. 8.26 *CDs for quantiles.* Let $Y_1, \dots, Y_n$ be independent observations from a smooth density f , with c.d.f. F . How can we construct CDs for its quantiles, the $\mu_q = F^{-1}(q)$? We wish such a CD to be nonparametric, without further assumptions on the f . We go through the main ideas for the case of the median $\mu = F^{-1}(\frac{1}{2})$, before extending methods and results to a general quantile $q \in (0, 1)$.

the
Mahalanobis
distance

see e.g. Mardia
et al. (1979,
Ch. xx)
the
Mahalanobis
CD
see Story #53:
how far is Reger
from Bach

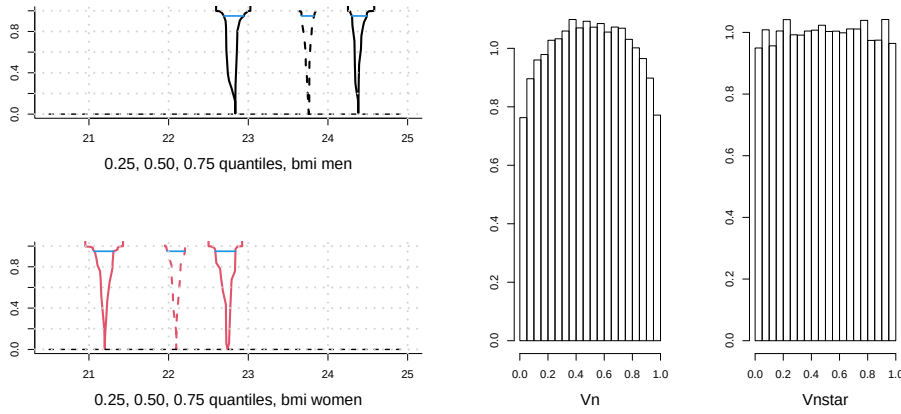


Figure 8.6: Left panel: for Ex. 8.26, confidence curves $cc(\mu_q)$, for the 0.25, 0.50, 0.75 quantiles of the BMI for Olympic speedskaters, men and ladies, computed via the Beta method, with linear interpolation. 95 percent intervals for the three quantiles are indicated via the blue horizontal line. Right panel: for Ex. 8.27, for the case of f being standard normal, $n = 100$, $q = 0.50$: histograms of V_n and V_n^* based on $\text{sim} = 10^5$ simulations. Also in other cases the V_n^* is much closer to uniformity than is V_n .

(a) It is not difficult to construct a first-order correct CD via large-sample results reached in Chapter 3, see in particular Ex. 3.21. With M_n the sample median, we have $\sqrt{n}(M_n - \mu) \rightarrow_d N(0, \frac{1}{4}/f(\mu)^2)$. Show that as long as $\hat{\tau}$ is a consistent estimator for $f(\mu)$, then $\sqrt{n}(M_n - \mu)/(\frac{1}{2}/\hat{\tau}) \rightarrow_d N(0, 1)$, and that this leads to the approximate CD

$$C_n(\mu, y) = \mu(\sqrt{n}(\mu - M_n)/(\frac{1}{2}/\hat{\tau})).$$

approximate
CD for
quantiles

One of several choices is to take $\hat{\tau} = \hat{f}(M_n)$, with $\hat{f}(y) = n^{-1} \sum_{i=1}^n h^{-1} K(h^{-1}(y_i - y))$ a kernel density estimator, with some kernel function K and bandwidth h . The best size for this fine-tuning parameter is of the type $h = c/n^{1/5}$, as seen in Ex. 3.42, and a classic rule of thumb which we typically might resort to here is to take $h = 1.059 \hat{\sigma}/n^{1/2}$, with $\hat{\sigma}$ the empirical standard deviation. Show that the confidence intervals from this CD take the form $M_n \pm z_0(\frac{1}{2}/\hat{\tau})/\sqrt{n}$, with z_0 the relevant normal quantile, like 1.96 for intended 95 percent intervals.

(b) A different idea starts out as follows. For the ordered observations $Y_{(1)} < \dots < Y_{(n)}$, show that

$$\Pr_f(\mu \leq Y_{(i)}) = \Pr(\frac{1}{2} \leq U_{(i)}) = 1 - \text{Be}(\frac{1}{2}, i, n - i + 1) \quad \text{for } i = 1, \dots, n.$$

Here $U_{(i)} = F(Y_{(i)})$; these form an ordered sample from the standard uniform, and we saw in Ex. 3.21 that they have Beta distributions. The $\text{Be}(x, a, b)$ is the c.d.f. of a $\text{Beta}(a, b)$. Define a full CD for μ , say $C_n^*(\mu, \text{data})$, via linear interpolation between the $C_n^*(y_{(i)}, \text{data}) = 1 - \text{Be}(\frac{1}{2}, i, n - i + 1)$ points. This also yields a confidence curve $cc_n^*(\mu, \text{data}) = |1 - 2C_n^*(\mu, \text{data})|$.

(c) Extend the two methods above, constructed there to deal with the median, to a general quantile $\mu_q = F^{-1}(q)$. For the first CD, use $\sqrt{n}(Q_{n,q} - \mu_q) \rightarrow_d N(0, q(1 - q)/f(\mu_q)^2)$, with $Q_{n,q} = F_n^{-1}(q)$ the empirical q quantile, and estimate $\tau_q = f(\mu_q)$ via $\hat{f}(F_n^{-1}(q))$. For the second CD, show first that $\Pr_f(\mu_q \leq Y_{(i)}) = 1 - \text{Be}(q, i, n - i + 1)$,

and use linear interpolation:

$$C_n^*(\mu_q, \text{data}) = \text{interpolation with } 1 - \text{Be}(q, i, n - i + 1) \text{ at } y_{(i)}, \quad (8.2)$$

the Beta
method CD for
quantiles

for $i = 1, \dots, n$. For μ_q inside $(y_{(i)}, y_{(i+1)})$, therefore, the CD value is interpolation between $1 - \text{Be}(q, i, n - i + 1)$ and $1 - \text{Be}(\mu_1, i + 1, n - i)$. We call this *the Beta method CD* for quantiles.

(d) For the dataset `bmi-olympics` with bmi measurements for Olympic speedskaters (1274 men, from 1952 to 2018, 967 ladies, from 1960 to 2018), compute, display, and interpret the confidence curves for the 0.25, 0.50, 0.75 quantiles, using the Beta-with-interpolation method. Reproduce a version of Figure 8.6, left panel. Supplement your curves by adding onto the plots also the direct large-sample approximation CD. You may modify your program to apply it to other datasets, of course, e.g. for comparing the 0.90 quantiles for the birthweights of boys and girls; see Story #22.

Ex. 8.27 *CDs for quantiles: how well do they work?* In Ex. 8.26 we found two non-parametric CD recipes, for any quantile $\mu_q = F^{-1}(q)$. Here we investigate how well they work, in terms of actual coverage probabilities for confidence intervals. For the methods $C_n(\mu_q, \text{data})$ and $C_n^*(\mu_q, \text{data})$, define

$$V_n = C_n(\mu_{q,\text{true}}, Y_1, \dots, Y_n) \quad \text{and} \quad V_n^* = C_n^*(\mu_{q,\text{true}}, Y_1, \dots, Y_n),$$

with $Y_1, \dots, Y_n$ drawn from the density in question. Accurate coverage, at all levels, means that the distribution of these two random CDs, at the true value, should be close to the uniform.

(a) To check the precision of these two CDs, carry out a simple simulation experiment. Take f equal to the standard normal, with $\mu_q = \Phi^{-1}(q)$ to be estimated with uncertainty; use $\hat{\tau} = \hat{f}(Q_{n,q})$ as above, with K the standard normal kernel and bandwidth $h = 1.059 \hat{\sigma}/n^{1/5}$ (which is optimal for the normal case, see Ex. 3.42), using the ordinary standard deviation from the data; and then simulate say $\text{sim} = 10^5$ values of $V_n = C_n(\mu_{q,\text{true}}, Y_1, \dots, Y_n)$ and $V_n^* = C_n^*(\mu_{q,\text{true}}, Y_1, \dots, Y_n)$. Check, perhaps for $n = 50, 100, 500, 1000$, how close the distributions of V_n and V_n^* are to the uniform. Produce a version of Figure 8.6, right panel. – For computing the C_n^* , and hence for executing that part of the simulation experiment, the `approx` algorithm of R is handy, carrying out linear approximation between the two values $1 - \text{Be}(\frac{1}{2}, i, n - i + 1)$ and $1 - \text{Be}(\frac{1}{2}, i + 1, n - i)$, for any μ inside the $[Y_{(i)}, Y_{(i+1)}]$ interval.

(b) Conduct a few similar simulation experiments, to see how close V_n and V_n^* are to uniformity, with different density f , quantile μ_q , sample size n .

(c) To assess ‘closeness to uniformity’ more accurately, use the monitoring processes of Ex. 3.9. For each of your simulation experiments, in addition to displaying histograms of V_n and V_n^* , compute and display the functions $Z_{\text{sim}}(t) = (\text{sim})^{1/2}\{G_{\text{sim}}(t) - t\}$ and $Z_{\text{sim}}^*(t) = (\text{sim})^{1/2}\{G_{\text{sim}}^*(t) - t\}$, where G_{sim} and G_{sim}^* are the empirical distribution functions of the V_n and V_n^* . Compute also $D_{\text{sim}} = \max_t |Z_{\text{sim}}(t)|$ and $D_{\text{sim}}^* = \max_t |Z_{\text{sim}}^*(t)|$. It will transpire (i) that V_n often is not particularly close to uniformity, unless n is rather large; (ii) that V_n^* is often so close to uniformity, even for moderate n and q near 0 or 1, that we cannot see that the distribution is not uniform, even with 10^5 simulated values.

Ex. 8.28 *Large-sample equivalence for two CDs for quantiles.* (xx to come. details for why the two CDs are large-sample equivalent. harder to show clearly that the 2nd is better than the 1st. nils thinks that it is, though, as of 16 August 2026. xx)

CDs in some nonregular setups

Ex. 8.29 *The length problem.* (xx might make a Satellite Collision Story, based on [Cunen et al. \(2020b\)](#). contrasting CD with Bayes. xx) There are several situations, of varying degrees of complexity, where the heart of the matter is, or can be transformed to, the following: with Y having the $N_p(\theta, \Sigma)$ distribution, with unknown mean vector and known or partly known variance matrix, reach inference for the length $\rho = \|\theta\| = (\theta_1^2 + \dots + \theta_p^2)^{1/2}$. See e.g. [Cunen et al. \(2020b\)](#) for an application involving the computation and real-time monitoring of the probability that two satellites will collide.

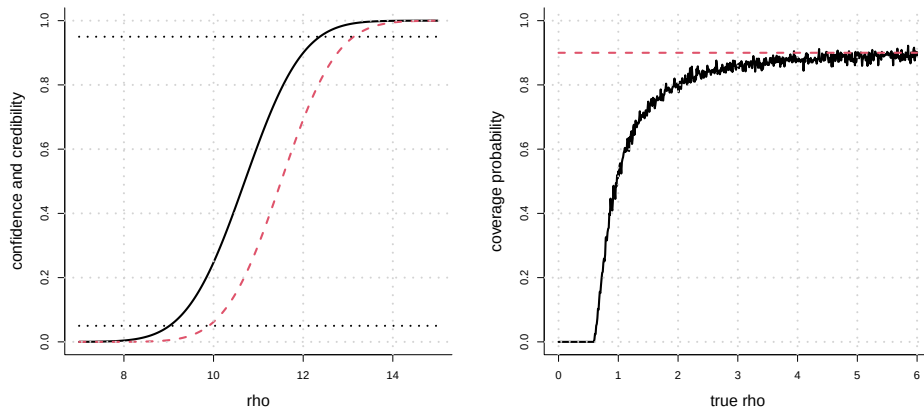


Figure 8.7: Left panel: for dimension $p = 10$ and observed $\hat{\rho} = 11.11$, the correct CD (full curve, with correct 90 percent confidence $[9.02, 12.37]$) and the Bayes cumulative posterior (slanted curve, with 90 percent credibility $[9.89, 13.13]$). Right panel: as a function of the true ρ , for dimension $p = 3$, the curve gives the actual coverage probability of the Bayesian 90 percent credibility interval, based on the posterior stemming from a flat prior for the θ .

- (a) Take first $\Sigma = I_p$, so $y \sim N_p(\theta, I_p)$, which means independent $Y_i \sim N(\theta_i, 1)$ for $i = 1, \dots, p$. Show that the ML estimator of ρ is $\hat{\rho} = \|Y\|$. Show also that $\hat{\rho}^2 \sim \chi_p^2(\rho^2)$, the noncentral chi-squared.
- (b) Deduce that $\hat{\rho}^2$ is overshooting its target ρ^2 , with mean and variance $p + \rho^2$ and $2p + 4\rho^2$. Find also an expression for $E\hat{\rho}$, and show that it overshoots ρ .
- (c) Show that the natural CD becomes $C(\rho, y) = 1 - \Gamma_p(\hat{\rho}^2, \rho^2)$.
- (d) A typical Bayesian analysis would start with a flat prior for $\theta_1, \dots, \theta_p$. Show that $\theta | y \sim N_p(y, I)$, and that this entails $\rho^2 | y \sim \chi_p^2(\hat{\rho}^2)$. For $p = 10$ and $\hat{\rho} = 11.11$, compute and draw both the CD and the Bayesian posterior distribution,

$$C(\rho, y) = 1 - \Gamma_p(\hat{\rho}^2, \rho^2) \quad \text{and} \quad B(\rho, y) = \Gamma_p(\rho^2, \hat{\rho}^2),$$

this CD gives
the correct
coverage

constructing a version of Figure 8.7, left panel. Comment on what you find.

- (e) Simulate to illustrate that the CD by construction works, producing confidence intervals with the correct coverage; $U_C = C(\rho_0, Y) \sim \text{unif}$ when data stem from the model, at position ρ_0 . show however that the Bayesian posterior distribution here risks being very far from producing intervals with the right coverage; $U_B = B(\rho_0, Y)$ is very far from being uniform. Construct a version of Figure 8.7, right panel, for the too low coverage probability of the Bayesian 90 percent credibility interval; the CD based intervals have

exact coverage. (xx link to Bernshtein–von Mises things in Ch. 7; here we’re outside BvM terrain. more on why and how. xx)

(f) More generally, with $Y_1, \dots, Y_n$ being i.i.d. from the $N_p(\theta, \sigma^2 I_p)$, with σ known, show first that $\bar{Y} \sim N_p(\theta, (\sigma^2/n)I_p)$. Then show that $\hat{\rho} = \|\bar{y}\|$ is the ML estimator, with distribution given by $n\hat{\rho}^2/\sigma^2 \sim \chi_p^2(n\rho^2/\sigma^2)$. On the Bayesian side, show that a flat prior for θ leads to $\theta | \text{data} \sim N_p(y, (\sigma^2/n)I_p)$. Show that these statements lead to these generalisations of the above

$$C_n(\rho, y) = 1 - \Gamma_p(n\hat{\rho}^2/\sigma^2, n\rho^2/\sigma^2),$$

$$B_n(\rho | y) = \Gamma_p(n\rho^2/\sigma^2, (n/\sigma^2)\hat{\rho}^2).$$

(g) (xx a bit more, regarding BvM, which holds for fixed p and ρ , with growing n . but misleading picture for finite n . do something to see interplay with n and p . xx) (xx also, briefly, to the case of $Y \sim N_p(\theta, \sigma^2 I_p)$, with σ estimated via an independent $\hat{\sigma}^2 \sim \sigma^2 \chi_m^2/m$. xx)

Ex. 8.30 *The Fieller problem: ratio of normal means.* (xx edit and clean. xx) In various situations one meets problems related to ratios of normal means, so to speak directly or indirectly. These become difficult when the nominator is small, where e.g. delta methods do not work well. Below we construct clean confidence curves for such ratio parameters. (xx point to Stories #35 New Haven and #36 on Oslo climate. xx)

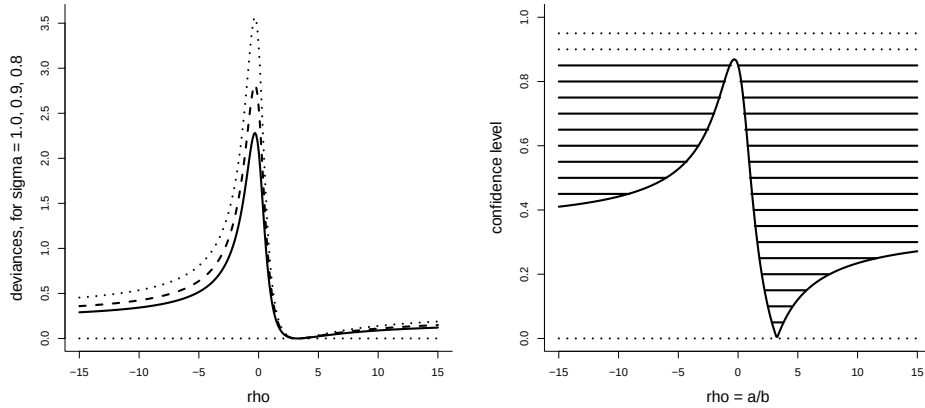


Figure 8.8: Ratio $\rho = a/b$ of normal means, with $(\hat{a}, \hat{b}) = (1.444, 0.444)$ and $\hat{\rho} = 3.252$: Left panel: the deviance function, for σ equal to 1.0, 0.9, 0.8. Right panel: for $\sigma = 1$, the $cc(\rho)$. For confidence levels above 0.87, the interval is the full real line; xxxx.

(a) Consider first the prototype setup for such questions, where $\hat{a} \sim N(a, 1)$ and $\hat{b} \sim N(b, 1)$ are independent, with inference required for $\rho = a/b$. Show first that the log-likelihood is a simple $\ell(a, b) = -\frac{1}{2}Q(a, b)$, with $Q(a, b) = (a - \hat{a})^2 + (b - \hat{b})^2$. Then derive the following formula for the profiled version,

$$\ell_{\text{prof}}(\rho) = -\frac{1}{2} \min\{Q(a, b) : a/b = \rho\} = -\frac{1}{2}(\hat{a} - \hat{b}\rho)^2/(1 + \rho^2),$$

and verify that the deviance function

$$D(\rho) = 2\{\ell_{\text{max}} - \ell_{\text{prof}}(\rho)\} = (\hat{a} - \hat{b}\rho)^2/(1 + \rho^2) \quad \text{is an exact } \chi_1^2.$$

(b) Generalising to the case of $\hat{a} \sim N(a, \sigma^2)$, $\hat{b} \sim N(b, \sigma^2)$, with σ known, explain that

$$cc(\rho) = \Gamma_1\left(\frac{1}{\sigma^2} \frac{(\hat{a} - \hat{b}\rho)^2}{1 + \rho^2}\right)$$

is a clear confidence curve for the $\rho = a/b$ ratio. Construct versions of Figure 8.8, based on having observed $(\hat{a}, \hat{b}) = (1.444, 0.444)$, with $\hat{\rho} = 3.252$. The left panel has deviance functions, for $\sigma = 1.0, 0.9, 0.8$, and the right panel displays the confidence curve $cc(\rho)$, for $\sigma = 1$. Comment on the unusual confidence regions, for different confidence levels α : these are bounded intervals for small α ; right-sided intervals for slightly bigger α ; the union of two infinite intervals for the next levels; and equal to the full real line at level 86.9 and above. All of this is caused by $\hat{b}$ being close to zero. You may play with your code to see how different values of the numerator $\hat{b}$ influence the $cc(\rho)$.

(c) (xx fix this, perhaps with an illustration, to be placed alongside the length problem one. xx) Consider now the classic linear regression setup, where $y_i = a + b(x_i - \bar{x}) + \varepsilon_i$, with the ε_i being i.i.d. $N(0, \sigma^2)$; see Ex. 3.32. Where, or when, will the line reach some threshold level c of interest? Argue that the answer is at $x_0 = \bar{x} + (c - a)/b$, for which the point estimate is $\hat{x}_0 = \bar{x} + (c - \hat{a})/\hat{b}$, in terms of the usual estimators $(\hat{a}, \hat{b})$. With $M_n = \sum_{i=1}^n (x_i - \bar{x})^2$, show that

$$Z_n(x_0) = \frac{\hat{a} + \hat{b}(x_0 - \bar{x}) - c}{\hat{\sigma}(1/n + (x_0 - \bar{x})^2/M_n)^{1/2}} \sim t_{df}, \quad \text{with } df = n - 2.$$

Use this to show that

$$cc(x_0) = F_{1,df}(Z_n(x_0)^2) = F_{1,df}\left(\frac{\{\hat{a} + \hat{b}(x_0 - \bar{x}) - c\}^2}{\hat{\sigma}^2(1/n + (x_0 - \bar{x})^2/M_n)}\right)$$

is a genuine confidence curve for the x_0 parameter. For $|x_0|$ becoming large, show that $cc(x_0) \rightarrow F_{1,df}(M_n \hat{b}^2 / \hat{\sigma}^2)$. Explain that this is $1 - p$, with the p-value for testing $b = 0$ vs. $b \neq 0$. When $t = M_n^{1/2} \hat{b} / \hat{\sigma}$ is not large, therefore, the $cc(x_0)$ is bounded below 1, with the implication that confidence intervals include $\pm\infty$. Discuss this finding.

(d) (xx bioassay. point to application. xx) When is $a + bx \geq c$, in other regression setups? Assume $(\hat{a}, \hat{b})$ is approximately a binormal, with means (a, b) and variance matrix estimated at Σ . The target is again $x_0 = (c - a)/b$. Explain that

$$Z(x_0) = \frac{\hat{a} + \hat{b}x_0 - c}{\{(1, x_0)\Sigma(1, x_0)^t\}^{1/2}} \quad \text{is approximately a standard normal,}$$

leading to the confidence curve $cc(x_0) = \Gamma_1(Z(x_0)^2)$.

Ex. 8.31 *CDs and posterior distributions with boundary constraints.* Here we learn about construction of CDs when there is a boundary condition on the focus parameter. This is sometimes an easy task, involving a natural positive post-data probability on the boundary point. We also compare with Bayesian procedures. Matters may of course be extended and generalised in several directions here, but for simplicity and conciseness we study a very simple prototype situation: y is $N(\theta, 1)$, and $\theta \geq 0$ a priori.

(a) Before we come to the parameter constraint, we deal with the more normal situation where there is no a priori constraint. The classical CD is then $C(\theta, y) = \Phi(\theta - y)$. Show that the Bayesian starting with a flat prior for θ finds the posterior distribution $\theta|y \sim N(y, 1)$, with cumulative $B(\theta|y) = \Phi(\theta - y)$, i.e. identical to the canonical CD. – The point below will partly be that this is *not* the same for the constrained problem.

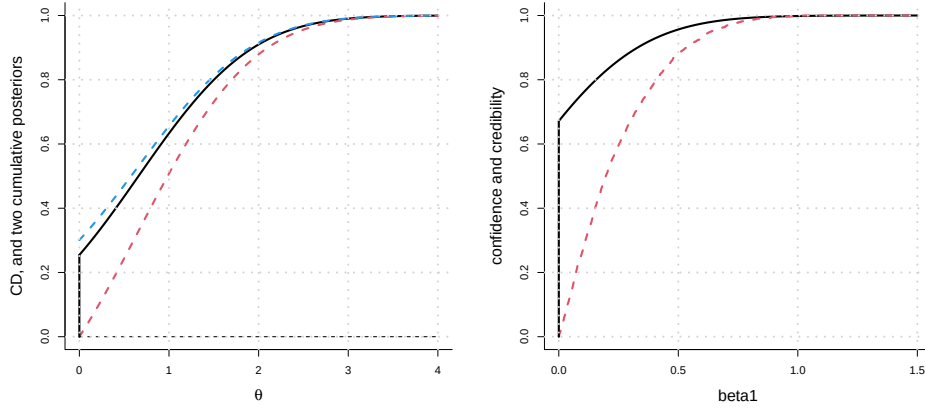


Figure 8.9: Left panel: for Ex. 8.31, with $y_{\text{obs}} = 0.66$ for the $N(\theta, 1)$ model, the black curve is the natural CD, with positive point mass 0.255 at zero. The red and the blue curves are Bayesian posterior distributions, for the flat prior on the halfline, and for the mixture prior with $\frac{1}{2}$ at zero and $\frac{1}{2}$ flat on the halfline, respectively. Right panel: for the linear regression setup of Ex. 8.32, with constraint $\beta_1 \geq 0$, the CD (full line) and the Bayes cumulative posterior (slanted line), with flat priors on $\beta_0, \beta_1, \log \sigma$.

(b) For the remaining points here, assume indeed that $\theta \geq 0$ a priori. Argue that the canonical CD should be $C(\theta, y) = \Phi(\theta - y)$ for $\theta \geq 0$. Its point mass at zero is $\Phi(-y)$. Graph the CD, for the three cases y_{obs} equal to $-0.22, 0.66, 1.99$.

(c) One Bayesian approach in this situation, where $\theta \geq 0$ a priori, is to let θ be flat on $[0, \infty)$. Show that then

$$\theta | y \sim \frac{\phi(\theta - y)}{\int_0^\infty \phi(\theta - y) d\theta} = \frac{\phi(\theta - y)}{\Phi(y)} \quad \text{for } \theta \geq 0,$$

and that the cumulative posterior distribution becomes

$$B(\theta | y) = \frac{\Phi(\theta - y) - \Phi(-y)}{1 - \Phi(-y)} = \frac{\Phi(\theta - y) - \Phi(-y)}{\Phi(y)} \quad \text{for } \theta \geq 0.$$

For the three cases of y_{obs} given above, graph the CD along with the Bayesian $B(\theta | y_{\text{obs}})$, and comment on what you find.

(d) In general terms, for the case of $y | \theta \sim N(\theta, 1)$, let θ have the mixture prior distribution $p_0\pi_0 + p_1\pi_1$, with the sub-priors π_0 and π_1 having their individual posteriors $\pi_0(\theta | y)$ and $\pi_1(\theta | y)$. Show that the posterior has a natural mixture form,

$$\theta | y \sim p_0^*(y)\pi_0(\theta | y) + p_1^*(y)\pi_1(\theta | y),$$

where

$$p_0(y) = \frac{p_0 f_0(y)}{p_0 f_0(y) + p_1 f_1(y)} \quad \text{and} \quad p_1(y) = \frac{p_1 f_1(y)}{p_0 f_0(y) + p_1 f_1(y)},$$

and with $f_0(y) = \int \phi(y - \theta)\pi_0(\theta) d\theta$ and $f_1(y) = \int \phi(y - \theta)\pi_1(\theta) d\theta$ the marginal densities following from the two priors. (This structure generalises to general mixture priors in general models, though that does not concern us just now.)

(e) For the prior $p_0\pi_0 + p_1\pi_1$, with π_0 a unit pointmass at zero and π_1 a flat prior on the halflife, show that $f_0(y) = \phi(y)$ and $f_1(y) = \Phi(y)$. With a 50-50 mixture, show hence that

$$p_0(y) = \frac{\phi(y)}{\phi(y) + \Phi(y)} \quad \text{and} \quad p_1(y) = \frac{\Phi(y)}{\phi(y) + \Phi(y)}.$$

Draw curves of these two posterior probabilities, one for the zero-point and the other for the halflife-based part, as y goes from say -5 to 5 . Show that the posterior cumulative distribution becomes $B^*(\theta|y) = p_0(y) + p_1(y)B(\theta|y)$ for $\theta \geq 0$. In particular, there is a pointmass $p_0(y)$ at zero. Construct a version of Figure 8.9.

(f) Show that there is no choice of (p_0, p_1) which makes the Bayesian cumulative posterior $B^*(\theta|y)$ agree with the CD $C(\theta, y)$. Devise a method for selection (p_0, p_1) such that the distance between $B^*(\theta|y)$ and $C(\theta, y)$ is small, for a relevant range of θ and possible observed y_{obs} .

(g) Generalise the formulae above to the case of $y_1, \dots, y_n$ i.i.d. $N(\theta, \sigma^2)$, with known σ .

Ex. 8.32 *CDs for regression parameters with boundary constraints.* The setup and insights of Ex. 8.31 carry over to other models, when there are constraints on one or more parameters. To illustrate this, consider the ordinary linear regression model with $Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ for $i = 1, \dots, n$, with the ε_i being i.i.d. $N(0, \sigma^2)$. Analysis methods for this model are classic (xx point Ch3 xx), as long as there are no constraints on the parameters. Suppose however that it is a priori known that $\beta_1 \geq 0$.

(a) From $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/M_n)$, with $M_n = \sum_{i=1}^n (x_i - \bar{x})^2$, show that $C_n(\beta_1) = G_{\text{df}}(M_n^{1/2}(\beta_1 - \hat{\beta}_1)/\hat{\sigma})$ is a proper CD, in that $C_n(\beta_1, \text{data}) \sim \text{unif}$, for each $\beta_1 \geq 0$. Here G_{df} is the c.d.f. for the t_{df} , and $\text{df} = n - 2$. The point-mass at zero is $G_{\text{df}}(-M_n^{1/2}\hat{\beta}_1/\hat{\sigma})$. To illustrate this, we have simulated (x_i, y_i) from a setup with normal data and $(\beta_0, \beta_1, \sigma) = (1, 0, 1)$, yielding uncorrelated binormal data, with estimates $(1.042, -0.173, 0.994)$, and $M_n = 6.520$. From this, compute and display this CD, as with Figure 8.9, right panel.

(b) The slanted curve in that figure's right panel is the c.d.f. for a Bayesian's $\beta_1 | \text{data}$ distribution. Specifically, take flat priors on $(\beta_0, \beta_1, \log \sigma)$, actively using the constraint $\beta_1 \geq 0$, which requires MCMC tools from Ch. 7. Complete a construction of the figure in question along these lines. We learn that the Bayes solution is not a good one here, not quite able to discover that β_1 is close to zero (or actually zero, as here). The 90 percent credibility interval is $[0.019, 0.636]$, compared to the correct 90 percent confidence interval $[0, 0.329]$, and with a clear chance of $C(0) = 0.671$ of β_1 being zero. (xx could ask for a better prior, $\frac{1}{2}\text{-}\frac{1}{2}$ for zero or a flat prior, will do better. xx)

(c) The same phenomenon surfaces in more general regression setups. In multiple linear regression $y_i = x_i^t \beta + \varepsilon_i$, with a constraint of the type $\beta_j \geq 0$, for a particular j , give the CD for that parameter. (xx round off. point to Story #60. connect to theme of Ch. 9. xx)

Ex. 8.33 *CDs in the truncated exponential model.* Here we consider a model sometimes called the truncated exponential model. We start with its simplest form, with data $Y_1, \dots, Y_n$ i.i.d. from the density $\exp\{-(y - a)\}$ for $y \geq a$. The a is the unknown start point for the distribution.

(a) Show that the ML estimator is equal to $U_n = \min_{i \leq n} Y_i$, the smallest data point. Show that $n(U_n - a)$ has a unit exponential distribution. Build from this a natural CD for a .

(b) Construct a predictive CD for the next sample point Y_{n+1} . Illustrate by computing and displaying the confidence curve for the test sample point, after having observed the six data points 3.735, 3.338, 10.634, 3.839, 5.667, 5.808.

(c) Then consider the more realistic two-parameter version of the model, with density

$$f(y_i, a, b) = (1/b) \exp\{-(y_i - a)/b\} \quad \text{for } y_i \geq a,$$

with a being the unknown start-point and b a scale parameter. Show that the ML estimators become $\hat{a} = U_n$ and $\hat{b} = (1/n) \sum_{i=1}^n (Y_i - U_n)$, again with U_n being the smallest observation.

(d) Construct accurate CDs and confidence curves for a , for b , and for the next datapoint Y_{n+1} . If some of your formulae cannot be given very explicit mathematical forms, this is ok, as long as numerical solutions can be found via numerical integration or simulation. Give approximations for these CDs for large sample sizes n .

(e) Ignoring these large-sample approximations, compute and display confidence curves for a , b , Y_{n+1} with the simple $n = 6$ dataset above.

Ex. 8.34 *CD inference for the exponential rate, with censored data.* The lifelength distribution for a certain type of technical components is considered exponential, i.e. with density $\theta \exp(-\theta t)$ for $t > 0$, on a priori grounds. To arrive at a point estimate and a confidence curve for θ , the firm producing these components sets in motion the simple experiment where n such items are set to work, under controlled natural conditions. One cannot wait until all components have died out, however, and the firm needs to report what can be said about the lifelength distribution, via θ , a certain time t_0 after project start.

(a) With data of the form observed t_i for the N of the items which have died within t_0 , and the information $t_i > t_0$ for the $n - N$ which are still alive and well, show that the combined likelihood function may be expressed as

$$\theta^N \exp\left[-\theta \left\{ \sum_{t_i \leq t_0} t_i + (n - N)t_0 \right\}\right].$$

(b) Show that the ML estimator is

$$\hat{\theta} = N/R = N / \left\{ \sum_{t_i \leq t_0} t_i + (n - N)t_0 \right\}.$$

With increasing sample size, and fixed t_0 , find expressions for the probability limits of N/n and R/n , and show that $\hat{\theta}$ is consistent.

(c) Show in fact that there is a limiting normal distribution here, with $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \tau(t_0, \theta)^2)$, and attempt to find an explicit (though not necessarily quick and simple) formula for the limit variance.

(d) Explain why the construction $C_n(\theta) = \Pr_{\theta}(\hat{\theta} \geq \hat{\theta}_{\text{obs}})$ yields a CD, and also how it can be computed in practice.

(e) Suppose the experiment described involves $n = 20$ such items, and that the lifelengths for the $N = 11$ of these that conk out before the deadline of $t_0 = 2.00$ years are

0.528 0.743 0.869 1.180 0.602 0.133 0.327 1.115 0.117 0.208 1.808

Compute and display perhaps as many as three (exact or approximate) confidence curves for θ , for this little experiment: the one described in (c); one based on the normal

approximation to the distribution of the ML estimator; and a t-bootstrap based version. Comment on your findings.

Ex. 8.35 *Estimating n based on observing the first r .* Suppose $Y_1, \dots, Y_n$ are i.i.d., from some known distribution with density f and cumulative F , but that one only observes the first r order statistics, $Y_{(1)} < \dots < Y_{(r)}$. Can we estimate n ? Such nonstandard problems turn up in various context, from estimating the size of a vocabulary to the number of unseen species. In this exercise we consider the special case of the unit exponential distribution, where the Y_i can be seen as waiting times, so the question may be phrased as how long time do we need to wait, until we've seen all items, when we have used a certain time to observe the first r .

(a) Let then $Y_1, \dots, Y_n$ be i.i.d. from the unit exponential, and assume $Y_{(1)} < \dots < Y_{(r)}$ are observed, with unknown n . Observing these r first data points is equivalent to observing the spacings $D_1 = Y_{(1)}$, $D_2 = Y_{(2)} - Y_{(1)}$, up to $D_r = Y_{(r)} - Y_{(r-1)}$. Use Ex. 1.15 to show that the joint distribution of these r first spacings may be written

$$g_r(d_1, \dots, d_r) = n(n-1) \cdots (n-r+1) \exp[-\{n(d_1 + \dots + d_r) - d_2 - 2d_3 - \dots - (r-1)d_r\}],$$

and deduce from this that $Y_{(r)}$ is sufficient for n .

(b) With $F(x) = 1 - \exp(-x)$, show that $F(Y_{(r)})$ has a Beta distribution with parameters $(r, n - r + 1)$.

(c) Show that the optimal CD for n , based on having observed the smallest r datapoints, is

$$C_r(n) = \Pr_n(Y_{(r)} \leq Y_{(r),\text{obs}}) = \text{Be}(F(Y_{(r),\text{obs}}), r, n - r + 1).$$

(d) (xx an example or two. suppose $Y_{(r),\text{obs}} = 0.348$ with $r = 33$. estimate n . see nils com87a or thereabouts. give normal approximation. but these are not good for r/n close to zero or one. can we characterise ML estimator. xx)

Ex. 8.36 (xx another discrete model thing. xx) (xx to come. xx)

Prediction with confidence

Ex. 8.37 *What's the next data point?* The confidence distributions and curves worked with so far have been for focus parameters in various statistical models. The concepts carry over to unknown or not-yet-observed random variables, however. When Y_0 is to be predicted, based on data Y and a relevant model, we say that $C(y_0, y)$ is a prediction confidence distribution for Y_0 , provided (i) $C(y_0, y_{\text{obs}})$ is a c.d.f., for any dataset y_{obs} , and (ii) $C(Y_0, Y) \sim \text{unif}$. Analogous comments and definitions apply to the prediction confidence curve $\text{cc}(y_0, y_{\text{obs}})$, whether such a cc is constructed directly or via $\text{cc}(y_0, y_{\text{obs}}) = |1 - 2C(y_0, y_{\text{obs}})|$.

prediction CDs

(a) Suppose $Y_1, \dots, Y_n, Y_{n+1}$ are i.i.d. from the normal (ξ, σ^2) , with the first n observed but not yet the Y_{n+1} of tomorrow. Assume first that σ is known. From $Y_{n+1} - \bar{Y} \sim N(0, \sigma^2(1 + 1/n))$, show that $C(y_{n+1}, y_{\text{obs}}) = \Phi(\{n/(n+1)\}^{1/2}(y_{n+1} - \bar{y}_{\text{obs}})/\sigma)$ is a prediction CD for Y_{n+1} . When σ is not known, and needs to be estimated via the familiar empirical standard deviation, show that

$$C(y_0, \text{data}) = G_m(\{n/(n+1)\}^{1/2}(y - \bar{y}_{\text{obs}})/\hat{\sigma})$$

is a prediction CD, with G_m being the t c.d.f., here with $m = n - 1$ degrees of freedom. Explain that for n large, the prediction CD becomes close to the ideal $\Phi((y - \xi)/\sigma)$; one cannot do better than this. Tomorrow's randomness does not disappear, even when the parameters of the model are perfectly known.

“an eminent
representative
of the statistics
profession”

(b) In the period 2009-2024, there have been eight recipients of the *Sverdrup Prize*, awarded by the Norwegian Statistical Association, with ages at winning being 63, 68, 60, 68, 68, 68, 62, 59. Compute and plot the confidence curve $cc(\text{next})$ for the next winner's age. What does it take for the profession to be significantly surprised when awardee no. nine is announced, i.e. an age outside the 95 percent confidence level?

(c) Suppose $Y_1, \dots, Y_n$ have been observed, i.i.d. from $\text{Expo}(\theta)$, with ML estimator $\hat{\theta} = 1/\bar{Y}$. What will the next Y' be like, and more generally the next $T_m = Y'_1 + \dots + Y'_m$, when these new Y'_j are from the same distribution? Using $2\theta Y_i \sim \chi^2_2$, show that

$$\hat{\theta} T_m \sim \theta \frac{2n}{\chi^2_{2n}} \frac{m}{\theta} \frac{\chi^2_{2m}}{2m} \sim m F_{2n, 2m},$$

in terms for the F with degrees of freedom $(2n, 2m)$, and construct a prediction CD for T_m based on this. An illustration is given in Story #44.

(d) Consider $Y_1, \dots, Y_{n+1}$ being i.i.d. from the uniform on $[0, \theta]$. Find a prediction CD for tomorrow's Y_{n+1} , based on $M_n = \max_{i \leq n} Y_i$.

Ex. 8.38 *The next binomial.*

Ex. 8.39 *What is Y , for a new x ?* Consider the linear regression model, with $Y_i = a + b(x_i - \bar{x}) + \varepsilon_i$ for $i = 1, \dots, n$, with the ε_i being i.i.d. $N(0, \sigma^2)$. For a new x_0 , perhaps for an individual not part of the dataset, what will Y_0 be? From $Y_0 = a + b(x_0 - \bar{x}) + \varepsilon_0$, show that

$$t(Y_0, \text{data}) = \frac{Y_0 - \hat{a} - \hat{b}(x_0 - \bar{x})}{\hat{\sigma} \{1 + 1/n + (x_0 - \bar{x})^2/M_n\}^{1/2}} \sim t_{n-2},$$

and explain that $C(y_0, \text{data}) = G_{n-2}(t(y_0, \text{data}))$ becomes a prediction CD for Y_0 . (xx a bit more. point to illustration. easiest to predict for x_0 in the vicinity of $\bar{x}$. xx)

(a) For the general linear regression setup, with $Y_i = x_i^t \beta + \varepsilon_i$ for $i = 1, \dots, n$, for covariate vectors x_i of length p , show that

$$C(y_0, x_0, \text{data}) = G_m \left(\frac{y_0 - x_0^t \hat{\beta}}{\hat{\sigma} (1 + x_0^t \Sigma_n^{-1} x_0 / n)^{1/2}} \right)$$

is a prediction CD for Y_0 , with $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t$ and degrees of freedom $m = n - p$.

(b) (xx round off, point to some story. xx)

Ex. 8.40 *Prediction CD vs. Bayes prediction.* (xx well, to be cleaned. xx)

$$\bar{f}_B(y) = \int f(y, \theta) \pi(\theta | \text{data}), \quad \bar{f}(y) = \int f(y, \theta) dC(\theta | \text{data}),$$

with the appropriate $C(\theta | \text{data})$.

Risk functions for CDs

Ex. 8.41 *Risk functions for CDs.* This exercise looks into loss and risk functions for and hence comparisons between CDs. In decision theory lingo, the loss function, as per

Chs. 7 and 9, when $C(\theta, y)$ is a CD for θ based on data y , is

$$\text{lo}(\theta, C) = \mathbb{E}(\theta_{\text{cd}} - \theta)^2 = \int (\theta_{\text{cd}} - \theta)^2 dC(\theta_{\text{cd}}, y).$$

Here θ_{cd} denotes a random draw from the CD in question. The risk function, therefore, is

$$\text{risk}(\theta, C) = \mathbb{E}_\theta \text{lo}(\theta, C) = \mathbb{E}_\theta \int (\theta_{\text{cd}} - \theta)^2 dC(\theta_{\text{cd}}, Y).$$

There is randomness on two levels; first, Y is drawn from the model, at the θ studied, leading to the $C(\theta, Y)$, then, θ_{cd} is drawn from that CD. We may also use e.g. $\text{lo}(\theta, C) = \int |\theta_{\text{cd}} - \theta| dC(\theta, y)$ above, and yet other variations, though the squared error lends itself to the easiest interpretation and calculations.

(a) The CD is a probability distributio, given the data y , and hence has a mean $\theta^*(y)$ and a variance $v^*(y)$. Show that $\text{risk}(\theta, C) = \mathbb{E}_\theta \{(\theta^*(Y) - \theta)^2 + v^*(Y)\}$. Examine this, for the rather clean CD of the type $C(\theta, y) = \Phi((\theta - \hat{\theta})/\kappa)$, based on a clear estimator $\hat{\theta} \sim N(\theta, \kappa^2)$. Show that $\text{risk}(\theta, C) = 2\kappa^2$.

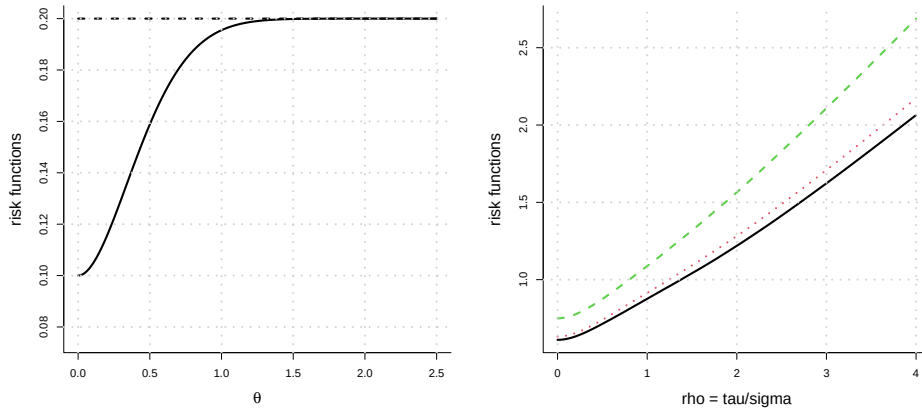


Figure 8.10: Left panel: for Ex. 8.41, risk function for the CD, in the setup with $Y_1, \dots, Y_n$ being i.i.d. from $N(\theta, 1)$, with $n = 10$, but with the restriction $\theta \geq 0$. It starts out at $1/n$ and then grows to the $2/n$ risk of the unrestricted case as θ grows. Right panel: for Ex. 8.42, for the variance component model, with $p = 4$ and $\sigma = 1$, risk functions $r(C, \tau)$ for three CDs for τ . The one based on $Z = \sum_{i=1}^p Y_i^2$ is best, closely followed by the one using $A = \sum_{i=1}^p |Y_i|$, whereas the one using the range $R = \max Y_i - \min Y_i$ does worse.

(b) Now study the simple model where $Y_1, \dots, Y_n$ are i.i.d. from $N(\theta, 1)$. Show that the natural CD based on the observed sample mean $\bar{y}_{\text{obs}} = n^{-1} \sum_{i=1}^n y_i$ is $C_n(\theta, y_{\text{obs}}) = \Phi(\sqrt{n}(\theta - \bar{y}_{\text{obs}}))$. Then prove that its risk function is $\text{risk}_n(C_n, \theta) = 2/n$.

(c) More generally, assume θ^* is some unbiased estimator of θ , with finite variance τ_n^2 , with the property that $\hat{\theta}^* - \theta$ has a distribution H_n symmetric around zero. Show that the associated CD becomes $C_n^*(\theta, y_{\text{obs}}) = H_n(\theta - \theta_{\text{obs}}^*)$, and show that its risk function becomes $2\tau_n^2$. The case of $\bar{Y}$ corresponds to $2/n$. Find the risk function for the case of the median based CD, with say $n = 10$, as for Figure 8.10, left panel.

(d) (xx fix this: Relate the above results to the optimality theorem for CDs, in certain situations, from CLP's Chapter 5. xx)

this CD is risk function optimal; see Schweder and Hjort (2016, Ch. 5)

(e) Now we change gears a bit, by putting the a priori assumption $\theta \geq 0$ on the table. Show that the ML estimator becomes $\hat{\theta} = \max(0, \bar{y})$, i.e. the sample mean truncated, if necessary, to zero. Argue that this leads to the natural CD

$$\tilde{C}_n(\theta, y) = \Phi(\sqrt{n}(\theta - \bar{y}_{\text{obs}})), \quad \text{for } \theta \geq 0,$$

in particular having a positive point-mass at zero.

(f) With θ_{cd} drawn from this CD, for given data, show that it may be expressed as $\max(0, \bar{y} + N/\sqrt{n})$, with N a standard normal. Show next that in the two-stage setup, with random data followed by θ_{cd} drawn from the $\tilde{C}_n$ CD, we have $\theta_{\text{cd}} - \theta = \max(0, \theta + (N + N')/\sqrt{n}) - \theta$, with N' another and independent standard normal. Use this to show that the risk_n($\tilde{C}_n, \theta$) can be expressed as

$$\begin{aligned} r_n &= \int [\{\max(0, \theta + (2/n)^{1/2}x)\} - \theta]^2 \phi(x) dx \\ &= \theta^2 \Phi(-(\tfrac{1}{2}n)^{1/2}\theta) + (2/n) \{ -(\tfrac{1}{2}n)^{1/2}\theta \phi((\tfrac{1}{2}n)^{1/2}\theta) + 1 - \Phi(-(\tfrac{1}{2}n)^{1/2}\theta) \}. \end{aligned}$$

Compute and display the risk functions for $\tilde{C}_n$ and C_n , for say $n = 10$, constructing a version of Figure 8.10, left panel. Comment on what we learn from this.

(g) (xx not fully sure about this one. xx) There are various other estimators and CDs worth considering in this $\theta \geq 0$ setting. To simplify matters, take $n = 1$, and consider the Bayes estimator $\hat{\theta}_B$, the conditional mean of $\theta | y$, with a flat prior on $(0, \infty)$. Show in fact that $\hat{\theta}_B = y + \phi(y)/\Phi(y)$, and verify that this is positive even when y is negative. Work out an expression for the naturally associated CD $C_B(\theta) = \Pr_{\theta}(\hat{\theta}_B \geq \hat{\theta}_{B,\text{obs}})$, and comment.

Ex. 8.42 Risk functions for three CDs in a variance components model. Consider the simple variance component model with independent observations $y_i \sim N(0, \sigma^2 + \tau^2)$ for $i = 1, \dots, p$, with σ known and τ the unknown parameter of interest; see Schweder and Hjort (2016, Example 4.1 and Exercise 5.8). The aim here is first to construct CDs based on (i) $Z = \sum_{i=1}^p y_i^2$, (ii) $A = \sum_{i=1}^p |y_i|$, and (iii) the range $R = \max y_i - \min y_i$; and then to compute and compare their risk functions. These are defined as

$$\text{risk}(C, \tau) = E_{\tau} |\tau_{\text{cd}} - \tau| = E_{\tau} \int |\tau_{\text{cd}} - \tau| dC(\tau_{\text{cd}}, Y),$$

with τ_{cd} a random draw from the $C(\tau, Y)$ distribution, and with Y itself denoting a dataset drawn from the distribution indexed by τ .

(a) Show that the natural CDs, based on Z , A , R respectively, are

$$\begin{aligned} C_Z(\tau, \text{data}) &= 1 - \Gamma_p(Z_{\text{obs}}/(\sigma^2 + \tau^2)), \\ C_A(\tau, \text{data}) &= 1 - G_p(A_{\text{obs}}/(\sigma^2 + \tau^2)^{1/2}), \\ C_R(\tau, \text{data}) &= 1 - H_p(R_{\text{obs}}/(\sigma^2 + \tau^2)^{1/2}). \end{aligned}$$

Here Γ_p is the cumulative distribution function of $Z_0 = \sum_{i=1}^p N_i^2$, with the N_i being i.i.d. and standard normal, which means $Z_0 \sim \chi_p^2$. Similarly, G_p and H_p are the cumulative distribution functions of $A_0 = \sum_{i=1}^p |N_i|$ and of $R_0 = \max N_i - \min N_i$, respectively.

(b) Show that a random draw τ_{cd} from the first of these, i.e. C_Z , for a given dataset, can be represented as $\tau_{\text{cd}} = (Z_{\text{obs}}/K - \sigma^2)_+^{1/2}$, where x_+ is notation for the truncated-to-zero

quantity $\max(x, 0)$, and where $K \sim \chi_p^2$. In the situation where data are random, from the model at position τ , deduce that

$$\tau_{\text{cd}} - \tau = \{(\sigma^2 + \tau^2)K_0/K - \sigma^2\}_+^{1/2} - \tau = \sigma[\{(1 + \rho^2)K_0/K - 1\}_+^{1/2} - \rho],$$

where $\rho = \tau/\sigma$, and K_0, K are two independent draws from the χ_p^2 . In other words, $F = K_0/K \sim F_{p,p}$, a F distribution with degrees of freedom (p, p) . Use this to compute the risk function $\text{risk}(C_Z, \tau)$, for $p = 4$ and $\sigma = 1$; this is the lowest of the three risk functions of Figure 8.10, right panel.

(c) Then consider the C_A option. Show that a random draw from an observed $C_A(\tau, \text{data})$ can be written $\tau_{\text{cd}} = \{(A_{\text{obs}}/A)^2 - \sigma^2\}_+^{1/2}$. Deduce that for random data behind the CD, we have the representation

$$\tau_{\text{cd}} - \tau = \{(\sigma^2 + \tau^2)(A_0/A)^2 - \sigma^2\}_+^{1/2} - \tau = \sigma[\{(1 + \rho^2)(A_0/A)^2 - 1\}_+^{1/2} - \rho],$$

with A and A_0 two independent draws from the G_p distribution. Use this to compute $\text{risk}(C_A, \tau)$. There is no simple expression for the density of A_0/A , so use simulation.

(d) Carry out similar analysis for the third CD, based on the range R . Construct a version of Figure 8.10, right panel.

(e) Use your program to explore the three risk functions for other values of p .

Meta-analysis, combining confidence curves

Ex. 8.43 *CD and cc for binomial probabilities.* Suppose y is observed from a binomial (n, θ) . The task is to construct a CD and a cc for θ .

(a) From the standard normal approximations for y , see Ex. 2.31, explain that

$$C_a(\theta, y) = \Phi\left(\frac{n\theta - y}{\{n\theta(1 - \theta)\}^{1/2}}\right) \quad \text{and} \quad C_b(\theta, y) = \Phi\left(\frac{\sqrt{n}(\theta - \hat{\theta})}{\{\hat{\theta}(1 - \hat{\theta})\}^{1/2}}\right),$$

are approximate CDs, with $\hat{\theta} = y/n$ the standard estimator for θ .

(b) The recipe of Ex. 8.7 does not quite work here since y has a discrete distribution. This invites the half-correction method

$$C(\theta, y) = \Pr_\theta(y > y_{\text{obs}}) + \frac{1}{2}\Pr_\theta(y = y_{\text{obs}}).$$

For say $n = 20$ and $y = 12$, compute and display this CD, along with (i) the same CD but without the half-correction, (ii) the two simple normal approximations above. Try other combinations of (n, y) , and demonstrate that they are approximately equal for moderate to large n .

(c) To investigate the basic CD property, take $n = 20$ and $\theta_{\text{true}} = 0.33$. Simulate a large number of $C(\theta_0, y)$, $C_a(\theta_0, y)$, $C_b(\theta_0, y)$, to check for their approximate uniform distribution. Try other values of (n, θ_0) , and summarise your findings.

Ex. 8.44 *CD and cc for comparing binomials.* In Ex. 8.43 we learn how to construct CDs for separate binomial parameters. Consider now the 2×2 table setup with two binomials, as with Ex. 4.31, say $y_0 \sim \text{binom}(m_0, p_0)$ and $y_1 \sim \text{binom}(m_1, p_1)$. How do we reach precise inference for the extent to which p_0 and p_1 differ?

(a) We first use the logistic transform $p_0 = H(\theta_0)$ and $p_1 = H(\theta_0 + \gamma)$, with $H(u) = \exp(u)/\{1 + \exp(u)\}$. Show that

$$\gamma = \log \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \log \frac{p_1}{1-p_1} - \log \frac{p_0}{1-p_0},$$

the log-odds difference. Write up the likelihood function for the observed (Y_0, Y_1) to deduce, via the optimal Recipe Six, that the best CD for γ takes the form

$$C(\gamma) = \Pr_\gamma(Y_1 > y_{1,\text{obs}} \mid Z = z_{\text{obs}}) + \frac{1}{2} \Pr_\gamma(Y_1 = y_{1,\text{obs}} \mid Z = z_{\text{obs}}),$$

this CD is used
in Stories #1
and #13

with $Z = Y_0 + Y_1$. The conditional distribution in question is the eccentric hypergeometric, found in Ex. 4.31.

(b) The following data table is from Normand (1999), and pertains to prophylactic use of lidocaine after a heart attack. The aim is to evaluate mortality from prophylactic use of lidocaine in acute myocardial infarction. We view the data here as pairs of binomials, with $y_{1,i} \sim \text{binom}(m_{i,1}, p_{1,i})$ and $y_{1,0} \sim \text{binom}(m_{i,0}, p_{1,0})$. Write the probabilities in logistic fashion, i.e. $p_{i,0} = H(\theta_{i,0})$ and $p_{i,1} = H(\theta_{i,0} + \gamma_i)$, with $H(u) = \exp(u)/\{1 + \exp(u)\}$. Show that

$$\gamma_i = H^{-1}(p_{i,1}) - H^{-1}(p_{0,i}) = \log \left(\frac{p_{i,1}}{1-p_{i,1}} \bigg/ \frac{p_{0,i}}{1-p_{0,i}} \right),$$

the log-odds difference. Construct and display the optimal CD for the γ_i , and also for the odds ratio $\rho_i = \exp(\gamma_i)$, for each of the six studies.

m1	m0	y1	y0
39	43	2	1
44	44	4	4
107	110	6	4
103	100	7	5
110	106	7	3
154	146	11	4

(c) Assume then that the log-odds parameter γ is the same, across studies, so that the six binomial data pairs relate to seven parameters. Find the optimal CD for this γ , and for the common odds ratio $\rho = \exp(\gamma)$. Translate the CDs to confidence curves, and display the six + one curves in a diagram. How would you conclude?

Ex. 8.45 Comparing Poisson parameters. Suppose $Y_0 \sim \text{Pois}(m_0\theta_0)$ and $Y_1 \sim \text{Pois}(m_1\theta_1)$. In what precise way is θ_1 different from θ_0 ?

(a) Writing $\gamma = \theta_1/\theta_0$, show that the likelihood is proportional to $\exp\{-\theta_0(m_0 + m_1\gamma)\} \theta_0^{y_0+y_1} \gamma^{y_1}$. Explain that the optimality recipe tells us inference should be made based on the distribution of $Y_1 \mid (Z = z)$, where $Z = Y_0 + Y_1$. Show that $Y_1 \mid (Z = z)$ has the binomial distribution $(z, m_1\gamma/(m_0 + m_1\gamma))$. Writing z for the observed $y_{0,\text{obs}} + y_{1,\text{obs}}$, show how this leads to the optimal CD

$$\begin{aligned} C(\gamma) &= \Pr_\gamma(Y_1 > y_{1,\text{obs}} \mid z) + \frac{1}{2} \Pr_\gamma(Y_1 = y_{1,\text{obs}} \mid z) \\ &= 1 - B_z(y_{1,\text{obs}}, m_1\gamma/(m_0 + m_1\gamma)) + \frac{1}{2} b_z(y_{1,\text{obs}}, m_1\gamma/(m_0 + m_1\gamma)). \end{aligned}$$

For an application, check with Story #12, for $y_0 = 1$, $y_1 = 7$, for the patient years m_0 , m_1 , in Aursnes et al. (2005).

(b) Assume now there are m independent Poisson pairs, perhaps from different but similar experiments, with $Y_{i,0}$ and $Y_{i,1}$ having parameters θ_i and $\theta_i\gamma$. With $Z_i = Y_{i,0} + Y_{i,1}$, explain that the above leads to the optimal CD for γ , based on pair i alone, is

$$C_i(\gamma) = 1 - B_{z_i}(y_{i,1}, \gamma/(\gamma+1)) + b_{z_i}(y_{i,1}, \gamma/(\gamma+1)).$$

Work also with the full log-likelihood, for the $m+1$ parameters, and derive the optimal CD for γ based in the full dataset:

$$C^*(\gamma) = \Pr_\gamma(S > s_{\text{obs}} | z_1, \dots, z_m) + \frac{1}{2} \Pr_\gamma(S > s_{\text{obs}} | z_1, \dots, z_m).$$

Here $S = \sum_{i=1}^m Y_{i,1}$. Explain that here, with no extra weighting between pairs, S is simply a binomial ($z_1 + \dots + z_m, \gamma/(\gamma+1)$), making it easy to compute the fusion CD.

Ex. 8.46 Basic meta-analysis. There is a very wide literature on combining information, with different names and labels, including meta-analysis, data fusion, etc. This exercise looks into some of the more basic versions, and where CDs will be helpful in later extensions below.

(a) Suppose $y_j \sim N(\mu, \sigma_j^2)$, for $j = 1, \dots, k$ independent sources, with the same focus parameter μ , and with variances taken to be known or well estimated. Consider the linear combination estimator $\hat{\mu} = \sum_{j=1}^k a_j y_j$. Show that it is unbiased, provided $\sum_{j=1}^k a_j = 1$, and find its variance. Show that the best choice, yielding minimal variance among the unbiased ones, is $a_j \propto 1/\sigma_j^2$, leading to

$$\hat{\mu} = \sum_{j=1}^k y_j / \sigma_j^2 \Big/ \sum_{j=1}^k 1/\sigma_j^2 = K^{-1} \sum_{j=1}^k y_j / \sigma_j^2, \quad \text{with } K = \sum_{j=1}^k 1/\sigma_j^2.$$

Show indeed that $\hat{\mu} \sim N(\mu, 1/K)$, with this minimal variance. Comment on what this leads to for the case where the σ_j are equal.

(b) In various settings there is a need to generalise the setting above to one where $y_j | \mu_j \sim N(\mu_j, \sigma_j^2)$, with these individual mean parameters not being equal, but having their own distribution, say $\mu_j \sim N(\mu_0, \tau^2)$. The task is then to reach inference for both the overall mean μ_0 and the spread τ among the μ_j . Show that $y_j \sim N(\mu_0, \sigma_j^2 + \tau^2)$, and that the log-likelihood function becomes

$$\ell(\mu_0, \tau) = -\frac{1}{2} \sum_{j=1}^k \left\{ \log(\sigma_j^2 + \tau^2) + \frac{(y_j - \mu_0)^2}{\sigma_j^2 + \tau^2} \right\}.$$

(c) Considering the spread parameter τ first, show that the profiled log-likelihood can be written

$$\ell_{\text{prof}}(\tau) = -\frac{1}{2} \sum_{j=1}^k \left[\log(\sigma_j^2 + \tau^2) + \frac{\{y_j - \hat{\mu}_0(\tau)\}^2}{\sigma_j^2 + \tau^2} \right], \quad \text{for } \tau \geq 0,$$

in which

$$\hat{\mu}_0(\tau) = \sum_{j=1}^k y_j / (\sigma_j^2 + \tau^2) \Big/ \sum_{j=1}^k 1/(\sigma_j^2 + \tau^2)$$

is the best linear combination estimator for μ_0 , for the fixed τ under inspection.

check Story
#13 for an
application,
with more
details

(d) Show, perhaps via Ex. 1.54, that

$$R(\tau) = \sum_{j=1}^k \frac{\{y_j - \hat{\mu}(\tau)\}^2}{\sigma_j^2 + \tau^2} \sim \chi_{k-1}^2,$$

and explain how this leads to the natural CD $C(\tau) = 1 - \Gamma_{k-1}(R(\tau))$. Examine its pointmass $C(0)$ at zero, and relate this to a test for homogeneity, i.e. that the $\mu_1, \dots, \mu_k$ have no spread.

(e) (xx fix this, needs to be clearer. xx) Consider this profiled log-likelihood as a function of $\gamma = \tau^2$, rather than of τ , and show that its derivative at zero is

$$D = \frac{1}{2} \sum_{j=1}^k \frac{1}{\sigma_j^2} \left\{ \frac{(y_j - \tilde{\mu}_0)^2}{\sigma_j^2} - 1 \right\}.$$

Here $\tilde{\mu}_0 = \hat{\mu}(0)$. A small D means that the y_j data have a low spread, and vice versa. Show that if $D \leq 0$, then $\hat{\tau}_{\text{ml}} = 0$, and if that $D > 0$, then $\hat{\tau}_{\text{ml}}$ is positive.

Ex. 8.47 *Combining CDs for the same parameter.* (xx a few exercises here. first for the same parameter, basic. then to CLP Ch 13 settings, then CDs to likelihood; then II-CC-FF. xx) Suppose that $C_1(\mu), \dots, C_k(\mu)$ are independent CDs for the same parameter μ , perhaps based on different sets of data. How can these be properly combined?

(a) Show that $N_j = \Phi^{-1}(C_j(\mu_{\text{true}}))$ is standard normal, at the true position in the parameter space underlying the C_j . With $w_1, \dots, w_k$ numbers such that $\sum_{j=1}^k w_j^2 = 1$, show that

$$\bar{C}(\mu) = \Phi\left(\sum_{j=1}^k w_j \Phi^{-1}(C_j(\mu))\right)$$

is a proper combination CD for μ . Variation: if one uses random weights $\hat{w}_j$, then $\sum_{j=1}^k w_j \Phi^{-1}(C_j(\mu))$ is typically not normal anymore, but its c.d.f. H might be easy enough to simulate. Show that this leads to the well-defined CD

$$\bar{C}(\mu) = H\left(\sum_{j=1}^k \hat{w}_j \Phi^{-1}(C_j(\mu))\right)$$

(b) (xx point back to Ex. 8.46. xx)

(c)

Ex. 8.48 *The problem of the Nile, via combo-CDs.* Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be independent and exponential, the X_i with parameter θ , the Y_i with parameter $1/\theta$. This is an instance where the natural sufficient stastic has dimension two, even though there is a single parameter.

(a) Show that the log-likelihoods, for the X_i and for the Y_i parts, become

$$\ell_1(\theta) = n(\log \theta - \theta \bar{X}), \quad \ell_2(\theta) = n(-\log \theta - \bar{Y}/\theta),$$

with separate ML estimators $\hat{\theta}_1 = 1/\bar{X}$ and $\hat{\theta}_2 = \bar{Y}$. Show further that we may represent these as $\hat{\theta}_1 \sim \theta A$, $\hat{\theta}_2 \sim \theta B$, where $A = 2n/K$ and $B = L/(2n)$, with K and L being χ_{2n}^2 .

(b) Show that the canonical CDs for θ , based on the two parts, are

$$C_1(\theta) = \Gamma_{2n}(2n\theta/\hat{\theta}_1), \quad C_2(\theta) = 1 - \Gamma_{2n}(2n\hat{\theta}_2/\theta).$$

the problem of
the Nile; check
with Ex. 9.36

The combination recipe hence leads to

$$\bar{C}(\theta) = \Phi((1/\sqrt{2})\{\Phi^{-1}(C_1(\theta)) + \Phi^{-1}(C_2(\theta))\}).$$

(c) Verify that the full log-likelihood becomes $\ell(\theta) = -n(\theta\bar{X} + \bar{Y}/\theta)$, and that the ML estimator is

$$\hat{\theta}_{\text{ml}} = (\bar{Y}/\bar{X})^{1/2} = (\hat{\theta}_1\hat{\theta}_2)^{1/2} \sim \theta F^{1/2},$$

where $F = A/B = L/K \sim F_{2n,2n}$. Explain that the CD associated with this ML estimator is $C_{\text{ml}}(\theta) = 1 - F_{2n,2n}(\hat{\theta}_{\text{ml}}^2/\theta^2)$.

(d) Run some simple experiments, where you for a few values of n simulate data and then plot the four CDs, and the four confidence curves (the two separate ones, the combination one, and the ML based one). You will find that the combo CD is very close to the ML based one.

Ex. 8.49 *From confidence to likelihood.* (xx to come. with illustrations. normal conversion. $\ell(\mu) = -\frac{1}{2}\Gamma_1^{-1}(\text{cc}_j(\mu))$. point to Ex. 5.22. xx) Consider a model with parameter vector θ and focus parameter $\mu = \mu(\theta)$. Above we have developed methodology effectively constructing a confidence curve from the log-likelihood function, via Wilks theorems and related tools. Here we address the topic of conversion; the task is to construct a log-likelihood function from the confidence distribution.

(a) Consider first the simple case of $\hat{\mu} \sim N(\mu, \sigma^2)$, with known σ . Explain that the log-likelihood is $-\frac{1}{2}(\mu - \hat{\mu})^2/\sigma^2$, with confidence curve $\text{cc}(\mu) = \Gamma_1((\mu - \hat{\mu})^2/\sigma^2)$. Show from this again that the *normal conversion*, $\ell_{\text{conv}}(\mu) = -\frac{1}{2}\Gamma_1^{-1}(\text{cc}(\mu))$, succeeds in precisely retrieving the log-likelihood function. normal conversion

(b) The key idea is now to carry out such normal conversion on any given confidence curve of relevance. From such a $\text{cc}_n(\mu)$, carry out the normal conversion above, with

$$\ell_{\text{conv},n}(\mu) = -\frac{1}{2}\Gamma_1^{-1}(\text{cc}_n(\mu)).$$

When a statistician is given this log-likelihood function, perhaps via summaries of other work for the focus parameter in question, show that she reaches the deviance function $D_{\text{conv},n}(\mu) = \Gamma_1^{-1}(\text{cc}_n(\mu))$. When the confidence curve is well calibrated, in the sense of having distribution $\text{cc}_n(\mu_0)$ close to a uniform, at the true value, show that $D_{\text{conv},n}(\mu_0)$ is close to a χ_1^2 . Also explain that the Wilks theorem recipe is consistent with this; the same confidence intervals are generated by recipes (i) the given $\text{cc}(\mu)$ and (ii) the conversion.

(c) (xx two examples, in situations where there is an exact cc and log-likelihood, to see that it works well, also with a bit of skewness. xx)

Ex. 8.50 *Combining three information sources.* The following is an illustration of the general principles for combining different information sources, beyond traditional meta-analysis (with the more general paradigm laid out in Ex. 8.51). Suppose one is interested in a particular parameter γ , the 0.75-quantile of a certain distribution. There are three independent sources of knowledge, related to this γ . Your statistical job is (i) to convert these sources of information into separate confidence curves, say $\text{cc}_1(\gamma), \text{cc}_2(\gamma), \text{cc}_3(\gamma)$, and (ii) then to combine these into a single fused $\text{cc}^*(\gamma)$.

The data sources can be summarised as follows. Source 1 is a classical normal sample, of size $n_1 = 20$, with unbiased estimates $\bar{y}_1 = 5.342$ and $\hat{\sigma}_1 = 2.179$, with

$\gamma = \xi_1 + 0.675\sigma_1$ the 0.75 quantile in the underlying normal distribution. Source 2 is a different sample, of size $n_2 = 12$, set up under different circumstances, and where normality cannot be trusted (though γ retains its interpretation as the 0.75 quantile of the underlying distribution in question). In fact the data are (in ordered fashion)
 1.159 2.453 3.320 4.160 4.168 5.287 5.343 5.826 6.084 6.632 7.150 12.044
 Here there is hence only ‘nonparametric information’ about γ . Source 3 stems from perhaps earlier and perhaps vaguely characterised data, expressed via an experienced Bayesian statistician, who after proper pondering says her prior for γ is a $N(7.50, 1.25^2)$.

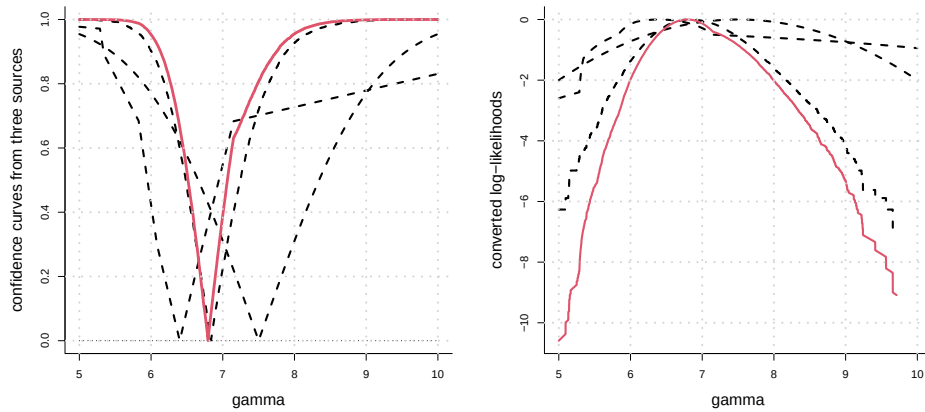


Figure 8.11: *Left panel: the three confidence curves for γ , the 0.75 quantile; the median confidence estimates are 6.40 (source 2), 6.84 (source 1), 7.50 (source 3). The full red curve is the fusion combined $cc^*(\gamma)$. Right panel: the three confidence curves converted to log-likelihoods, and then combined.*

(a) Transform the pieces of information above to a construction of a version of Figure 8.11, left panel, with the three confidence curves for γ . The most tricky of these is the nonparametric one, where you first might argue for the CD values $C_2(y_{2,(j)}) = 1 - \text{Be}(0.75, j, n_2 - j + 1)$ at the observed ordered data $y_{2,(j)}$, involving the relevant cumulative Beta distribution, and then use the convenient **approx** algorithm in R, perhaps as below. This is the Beta with interpolation method of Ex. 8.26.

```
gammaval <- seq(5,10,by=0.01)
jj2 <- 1:nn2
C2valshort <- 1 - pbeta(0.75, jj2,nn2-jj2+1)
plot(yy2,C2valshort,type="p")
C2plusval <- approx(yy2,C2valshort,method="linear",xout=gammaval)\$y
```

(b) Carry out normal conversion of the three confidence curves, to log-likelihoods functions. Then do the fusion: combine them, and transform back to the fused confidence curves $cc^*(\gamma)$. Do a version of Figure 8.11, right panel.

Ex. 8.51 II-CC-FF: Independent Inspection, Confidence Conversion, Focused Fusion. (xx to come. using Cunen and Hjort (2022). point to Bayesian updating being part of this, but allows user keeping only prior for the focus parameter. aim to demonstrate iiceff in Story #12. mention Ex. 5.22. xx)

Leftovers and further material

Ex. 8.52 *Private attributes.* (xx to be checked with care. xx) The probability ψ of cheating at exams might be hard to estimate, but a bit of randomisation might grant anonymity and yield valid estimates. Suppose there are three cards, two with the statement ‘I did cheat’, and ‘I did not cheat’ on the third. Students are asked to draw one of the three cards randomly and answer either *true* or *false* to the drawn statement, without revealing it.

(a) Show that the probability of *true* is $(1 + \psi)/3$. Assume a binomial model for the number of students answering *true*, and devise a CD for ψ .

(b) Assume 1000 students go through the simple post-exam exercise above (anonymously). Find and display CDs for ψ for the cases of respectively 300, 350, 400 out of the 1000 answered *true*.

Notes and pointers

[xx A few remarks. xx]

ToDo: (i) complete the optimality for CDs in expofamilies. (ii) give the argument that testing ought to have alternatives in mind, and CDs are useful for assessing this. calibrate that with Ch4.

point of course to [Schweder and Hjort \(2002\)](#) and [Schweder and Hjort \(2016\)](#), the prime source for CDs. there: rather more on optimality, risk functions, there is a Neyman–Pearson lemma for CDs, etc.

re Ex. 8.48: point to Problem of the Nile, with Fisher and more.

point to Brazzale Davison Reid, paper and book, and to CLP Ch 7, concerning the big topic of refining first-order asymptotics with various tools. also, Wilks often better than direct normal, see Story [#84](#).

we point to [Schweder and Hjort \(2016, Example 3.11\)](#), [Fisher \(1930\)](#), [Xie and Singh \(2013\)](#), [Hjort and Schweder \(2018\)](#), [Cunen and Hjort \(2022\)](#), [Singh et al. \(2005\)](#), ...

for the conversion, from confidence to likelihood, point both to CLP, to [Cunen and Hjort \(2022\)](#), but find also the Efron.

(xx now brief notes on solution, in CD for mean by standard deviation. xx) writing $d = x_b - x_a$, note first that $Y_b - Y_a \sim N(d^t \beta, 2\sigma^2)$, so $p = \Pr(Y_b > Y_a) = \Phi(\kappa/\sqrt{2})$, with $\kappa = d^t \beta / \sigma$. we wish a CD for this κ and hence for p . for this, $Y_b - Y_a - d^t \hat{\beta}$ is zero-mean normal with variance $(2 + \tau_n^2/n)\sigma^2$, where $\tau_n^2 = d^t \Sigma_n^{-1} d$, with $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t$, as per usual linear regression theory of Ch3. via this we have a CD for $Y_b - Y_a$, for large n close to simply $Y_b - Y_a \sim N(d^t \beta, 2\sigma^2)$. also, with $\hat{\kappa} = d^t \hat{\beta} / \hat{\sigma}$, we have

$$\sqrt{n} \hat{\kappa} / \tau_n = \frac{\sqrt{n} d^t \hat{\beta} / (\sigma \tau_n)}{\hat{\sigma} / \sigma} = \frac{N(\sqrt{n} \kappa / \tau_n, 1)}{(\chi_m^2 / m)^{1/2}}$$

this then leads to the clear CD

$$C(\kappa) = \Pr_{\kappa}(\sqrt{n} \hat{\kappa} \geq \sqrt{n} \kappa_{\text{obs}}) = 1 - G_m(\sqrt{n} \hat{\kappa}_{\text{obs}} / \tau_n, \sqrt{n} \kappa / \tau_n).$$

this should be done for New Haven, regarding $p = \Pr(Y_b > Y_a)$.

(xx repair. we point to Story [#60](#). xx) There’s a notable discrepancy between the frequentist Schweder-Hjort CD and the Bayesian posterior distribution associated with a flat prior on the $[0, \infty)$ interval, in cases where the y_{obs} is close to, or perhaps even to the left of, the boundary point.

I.9

Loss, risk, performance, optimality

Statistics is a mathematical formalisation of how to make good decisions under uncertainty. One source of uncertainty is that the true state of nature, say θ , is not known when we are to make our decisions, and since the utility or loss of a decision depends on θ , we need to be clear about how bad it is when our decision is off. This is the role of loss functions. Decisions are based on data, and it is not without consequence how we use the data: Some procedures for going from data to a decision are better than others. Therefore, it makes sense to see how a certain procedure performs on average. This is the role played by risk functions. Risk functions average out the data, but they still depend on θ , so the risk function of various decision procedures are often difficult to order. Some might be preferable for certain values of θ , while others might be better for other values of θ . This chapter introduces criteria that let us, nevertheless, say something about how good a decision procedure is. A decision procedure is said to be *admissible* if there is no other that does better, in terms of the risk function, whatever the truth or future may be. One says that a decision procedure is *minimax* if it is the best one in the most unfortunate situation. In proving that certain decision procedures are admissible or minimax, Bayesian thinking is an essential tool. This includes the concept of Bayes risk, where not only the data are averaged out, but also the various possible states of nature.

Key words: admissibility, Rao–Blackwell, loss functions, minimaxity, multiparameter estimation, UMVU estimators, risk functions

A statistical decision, as most decisions in life when you think about it, is a function whose domain is what we observe, and whose range are all possible decisions we can make in a given setting: I look out the window and see grey clouds, and choose to take my umbrella with me when I go out. I see that you are smiling and I think that you are happy. Formally, there is a probability space $(\Omega, \mathcal{F}, \text{Pr})$ in the background, and the data are realisations of a random element $X: \Omega \rightarrow \mathcal{X}$. A decision procedure is then a function $\hat{a}: \mathcal{X} \rightarrow \mathcal{A}$, where $\mathcal{A}$ is the collection of all possible decisions or actions we might take in a given setting, called the *action space*. How wise our choice of $\hat{a}$ is, or turns out to be, depends on the true distribution P_{θ_0} of X . This distribution belongs to a family $\{P_{\theta}: \theta \in \Theta\}$ of distributions on the measurable space $(\mathcal{X}, \mathcal{B})$. A *loss function* $L: \Theta \times \mathcal{A} \rightarrow [0, \infty)$ measures ‘how much’ we lose by choosing action a when the true state of nature is θ , and as the name indicates, the less loss the better.

To be concrete, consider the point estimation problem with data $X_1, \dots, X_n$ independent $N(\theta, 1)$, where θ is an unknown parameter to be estimated. A common and mathe-

matically tractable loss function is then the squared error loss function $L(\theta, a) = (a - \theta)^2$. For another classical example, suppose we wish to test $H_0: \theta \leq 0$ versus $H_A: \theta > 0$. The action space is $\mathcal{A} = \{\text{keep } H_0, \text{reject } H_0\}$, and a natural loss function is

$$L(\theta, a) = \begin{array}{c|cc} & \theta \leq 0 & \theta > 0 \\ \hline \text{keep } H_0 & 0 & L_2 \\ \text{reject } H_0 & L_1 & 0 \end{array}$$

With this loss function, we lose $L_1 > 0$ when rejecting a true null hypothesis, and L_2 if we fail to reject a null hypothesis that is false. In statistical jargon, failing to reject a null hypothesis is called a Type I error, while when we fail to reject a null hypothesis that is false, we commit a Type II error.

The general setup so far is essentially as for what we built with the Bayesian framework of Ch. 7, see Ex. 7.3, where one in addition to the elements above pushes in one more ingredient, that of a prior distribution for the θ . We do comment more below, and in various exercises, on Bayesian components and connections, but the main spirit of the present chapter is the classical or frequentist one, with no priors. Here different decision procedures are compared by the loss they incur for each value of θ , that is, by their *risk function*

$$R(\theta, \hat{a}) = E_{\theta} L(\theta, \hat{a}(X)) = \int_{\mathcal{X}} L(\theta, \hat{a}(x)) dP_{\theta}(x). \quad (9.1)$$

the risk
function

For each decision procedure $\hat{a}$, the risk function $R(\theta, \hat{a})$ is a function of θ , and the second argument $\hat{a}$ only indicates what procedure we are computing the risk of. For two decision procedures $\hat{a}_1$ and $\hat{a}_2$ their respective risk functions may cross, that is $R(\theta, \hat{a}_1) < R(\theta, \hat{a}_2)$ for some θ , while $R(\theta, \hat{a}_1) > R(\theta, \hat{a}_2)$ for other values of θ . This entails that comparison of risk functions only provides a partial ordering of decision procedures, and as such does not point clearly at a best decision procedure. What is clear, however, is that if $R(\theta, \hat{a}_1) \leq R(\theta, \hat{a}_2)$ for all θ , with strict inequality for at least one value of θ , then $\hat{a}_2$ should be discarded from the competition: We say that $\hat{a}_1$ *dominates* $\hat{a}_2$, and $\hat{a}_2$ is said to be *inadmissible*. A decision procedure that is not dominated by any other decision procedure is *admissible*. Clearly, inadmissible decisions procedures should be discarded, but admissibility in and of itself does not tell us much about a decision procedure.

admissibility

One fix to the problem of crossing risk functions is to limit the search of a best decision procedure to certain subclasses of decision procedures. In the search for a ‘best’ estimator, for example, it is common to restrict ones search to the class of *unbiased estimators*; see Ch. 3.

Another method by which to compare decision procedures is the the *minimax principle*. According to this principle, the estimator with the best performance in the worst possible scenario ought to be chosen. We say that a decision rule a^* is *minimax* if it minimises the maximum risk, that is, if $\sup_{\theta \in \Theta} R(\theta, a^*) \leq \sup_{\theta \in \Theta} R(\theta, \hat{a})$ for all decision procedures $\hat{a}$.

minimax

If you are a Bayesian and venture into the Ch. 7 terrain of constructing prior distributions $\pi(\theta)$ over the parameter space Θ , the problem of risk functions only being partially ordered can be circumvented. As studied in that chapter, what you are interested in then is the *Bayes risk* of the decision procedures you are comparing, that is

Bayes risk

$$BR(\hat{a}, \pi) = E_{\pi} R(\theta, \hat{a}) = \int_{\Theta} R(\theta, \hat{a}) \pi(\theta) d\theta.$$

For a given decision procedure $\hat{a}$ and prior π , the Bayes risk is a single number (in contrast to the risk function which is a function of the unknown truth), and under certain conditions that we will explore in Ch. 7 and will explore further below, there will be a unique decision procedure $\hat{a}$ that, for a given prior π , minimises $\text{BR}(\hat{a}, \pi)$. This decision procedure is the *Bayes solution*, see Ex. 7.3. As we shall see in this chapter, Bayes solutions are, despite the name, of high importance also for the frequentists.

Decisions and risk functions

Ex. 9.1 *Coin tossing.* To get a feeling for some of the basic challenges and concepts concerning the comparison of various estimators, we start out with the emblematic problem of estimating the probability of heads in $n = 10$ independent coin tosses. Let $Y_1, \dots, Y_n$ be independent Bernoulli random variables with expectation θ .

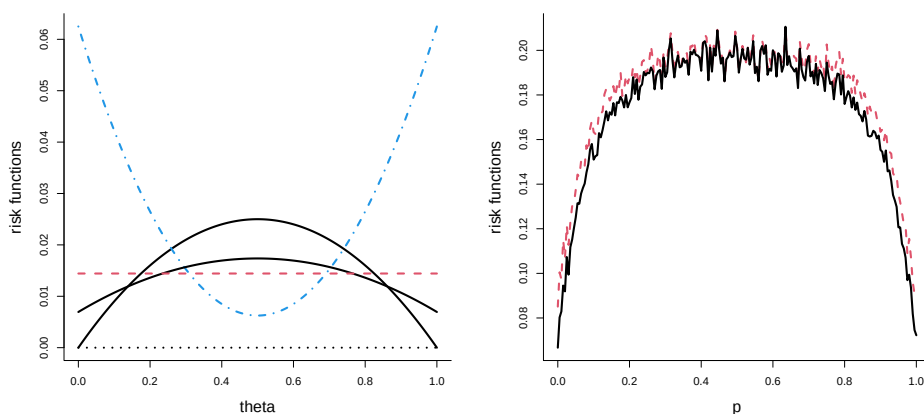


Figure 9.1: Left panel: for Ex. 9.1, risk functions for four estimators of $\theta = \text{Pr}(\text{heads})$: the ML (going down to zero at the ends); the Bayes estimator $(Y+1)/(n+2)$ (dotted line); the 50-50 and Y/n (going up at the ends); and the minimax estimator (flat). Right panel: for Ex. 9.2, risk functions for estimation methods A and B, for the case of $n = 100$ and distance 1 between the two normals in the mixture, based on 1000 simulations for each grid point. The ML method is just slightly better than the simpler moment matching method, for the given loss function.

(a) Sketch the risk function of the maximum likelihood (ML) estimator $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ under squared error loss $L(\theta, \bar{Y}_n) = (\delta - \bar{Y}_n)^2$, as for Figure 9.1. Recall that $n = 10$.

(b) Suppose we have some intuition about where on the unit interval the probability θ might be located, close to a value $0 < \theta_0 < 1$, say. One way in which such a prior hunch might be employed is by taking as our estimate a convex combination of the ML estimator and θ_0 , that is $\hat{a} = w\bar{Y}_n + (1-w)\theta_0$, for some user defined weight $w \in [0, 1]$. For $w = 1/2$ and $\theta_0 = 1/2$, sketch the risk function of this estimator, i.e. one of the four risk functions of Figure 9.1. Suppose that your task is to estimate the probability of giving birth to a boy. Which of the two above estimators do you prefer, the ML estimator, or $\hat{a}$ with $w = 1/2$ and $\theta_0 = 1/2$?

(c) Based on the risk functions you sketched in (a) and (b), we see that the estimators $\hat{a}$ and $\bar{Y}_n$ are difficult to compare, unless we have strong prior feelings about where the true value is. As mentioned in the introduction, a fix to this problem of comparison is to limit

our search to the class of estimators that are unbiased for what we are estimating. Look back at Ex. 5.7 and explain why, when the search is restricted to the class of unbiased estimators, the ML estimator is the clear winner.

(d) Consider the estimator from Ex. (b) with $\theta_0 = 1/2$. Find a weight $w = w(n)$ such that the risk function $R(\theta, \hat{a})$ is constant. For $n = 10$, plot the risk function of this estimator (that is, the flat risk of Figure 9.1). Suppose you have absolutely no idea whatsoever about where in the unit interval θ may be located. Which of your three estimators of θ do you prefer? In Ex. 9.12 we learn that this constant risk estimator is indeed minimax.

Ex. 9.2 *Estimating the mixture proportion.* Suppose i.i.d. observations $X_1, \dots, X_n$ come from a mixture of known densities f_0 and f_1 , i.e. from $f_p = (1-p)f_0 + pf_1$, but with unknown fraction p . The task is to estimate this p , yielding the estimated density as $\hat{f} = (1-\hat{p})f_0 + \hat{p}f_1$. On this occasion we evaluate the risk based on the loss function

$$L(p, \hat{p}) = \sqrt{n} \max |\hat{f}(x) - f_p(x)|,$$

as opposed to simpler ones like squared error. The $\sqrt{n}$ factor is mostly cosmetic, to make the risks stable in size for growing n . Below we compute and display risk functions for the case of $f_0 = N(0, 1)$ and $f_1 = N(d, 1)$, with the illustration depending on the value of d . For concreteness, take $d = 1$ and $n = 100$ below. Can data tell us the fraction of types 0 and 1, even with the densities are not far from each other?

(a) Estimator A is to match the sample mean $\bar{X}$ to the model mean. Show that this is solved by the equation $\bar{X} = (1-p)\xi_0 + p\xi_1$, with ξ_0 and ξ_1 the means under f_0 and f_1 . Estimator B is the ML one. Show that this is the maximiser of $\ell_n(p) = \sum_{i=1}^n \log((1-p)f_0(X_i) + pf_1(X_i))$, which may be found after a numerical search over a fine grid of p , or via another numerical optimiser.

(b) Now compute the two risk functions, via simulations; for each p , simulate a high number of $\hat{p}$ and hence a high number of losses, after which the risk is their average. Explain why this is an adequate risk finding procedure. Construct a version of Figure 9.1, right panel; for this loss function there is little difference in the risks. When running your code you may also read off the simpler mean squared error risks, i.e. $E_p(\hat{p} - p)^2$, though the point is here that we can handle much more complicated loss functions.

(c) While we're at it, find also the limiting distributions of $\sqrt{n}(\hat{p} - p)$, for the simple and the log-likelihood method, and compare their variances. One learns that the ML wins, with relatively speaking clearer amounts near the endpoints 0 and 1, compared to the risk results above. So the loss function matters, and with the one above the two estimation methods fare almost equally well.

(d) For any smooth increasing $h(x)$, show that $E_p h(X) = (1-p)h_0 + ph_1$, with h_0 and h_1 the mean values under f_0 and f_1 . Construct the relevant moment matching estimator for p . For such an $h(x)$, of your choice, compute the risk function and compare to the two of Figure 9.1, right panel.

Ex. 9.3 *How many cards in the hat?* There is an unknown number N of cards in a hat, numbered $1, \dots, N$. You draw $X_1, \dots, X_n$ cards, with $n = 5$, with replacement, and your task is to estimate N , with loss function $|\hat{N} - N|/N = |\hat{N}/N - 1|$.

(a) Show that X_i has mean $(N+1)/2$. Explain that the moment matching estimator is $\hat{N} = 2\bar{X} - 1$. Another estimator is $N^* = \max_{i \leq n} X_i$. Compute and display their risk

functions, say for $N = 10, \dots, 100$. To compute these you may resort to simulation, as with Ex. 9.2. One learns that N^* does better than $\hat{N}$, regardless of the correct N .

(b) Now attempt to be more clever than N^* , perhaps using $N^* + c_n$ for a suitable c_n . For $n = 5$, compute and display also this third risk function. Do you succeed in being uniformly more clever than N^* ?

(c) Now redo your analyses and risk calculations, when the n cards are drawn from the N -hat without replacement.

Ex. 9.4 *Decide on small, moderate, or big.* A crucial parameter θ is being assessed, in order to place it in categories small, $\theta \in B_1 = [0, 2]$; moderate, $\theta \in B_2 = (2, 4)$; or big, $\theta \in B_3 = [4, \infty)$. An action $a \in \{1, 2, 3\}$ is taken, based on data, corresponding to declaring θ to be small, moderate, or big. The potential losses incurred, i.e. $L(\theta, a)$, follow this table, with specified positive losses $\ell_{i,j}$ in cases where the action taken is not the correct one, i.e. when $j \neq i$:

	action 1	action 2	action 3
$\theta \in B_1$	0	$\ell_{1,2}$	$\ell_{1,3}$
$\theta \in B_2$	$\ell_{2,1}$	0	$\ell_{2,3}$
$\theta \in B_3$	$\ell_{3,1}$	$\ell_{3,2}$	0

(a) Assume for simplicity that all errors are equally unfortunate, with all $\ell_{i,j} = 1$ for $j \neq i$. Suppose that a single $Y \sim \text{Gam}(\theta, 1)$ is observed. Since its mean is θ it makes sense to take $\hat{a}(Y)$ equal to 1, 2, 3, according to which box Y lands in. Show that the risk function becomes

$$R(\theta) = \begin{cases} 1 - G(2, \theta) & \text{if } \theta \in B_1, \\ G(2, \theta) + 1 - G(4, \theta) & \text{if } \theta \in B_2, \\ G(4, \theta) & \text{if } \theta \in B_3, \end{cases}$$

writing $G(y, \theta)$ for the c.d.f. of the $\text{Gam}(\theta, 1)$. Compute and display.

(b) It might be more clever to take decisions 1, 2, 3 based on Y landing in $B_1^* = [0, 2 - c]$, $B_2^* = (2 - c, 4 + c)$, $B_3^* = [4 + c, \infty)$, for a suitably wise c . Compute and plot a few associated risk functions, to see whether this might be more clever than using $c = 0$. By tracking these risk functions numerically, find indeed the value of c which gives lowest possible max-risk in this class of procedures (the answer is at 0.22).

(c) Generalise the situation above to the case where one observes $\bar{Y}$, the sample mean of $Y_1, \dots, Y_m$ being i.i.d. from $\text{Gam}(\theta, 1)$. Compute and plot risk functions for e.g. $m = 10$.

UMVU estimators, Rao–Blackwell, Lehmann–Scheffé

Ex. 9.5 *Uniformly minimum variance unbiased estimators.* What is variably called a *best unbiased estimator*, the *uniformly minimum variance unbiased estimator*, or the *UMVU estimator*, is defined as follows: An estimator $\hat{a}$ is the UMVU estimator for $g(\theta)$ if it is unbiased for $g(\theta)$, and $\text{Var}_\theta \hat{a} \leq \text{Var}_\theta \tilde{a}$ for all θ , for any other unbiased estimator $\tilde{a}$. When it is feasible (see Ex. 5.6 and 5.7–5.8), the easiest way of establishing that an unbiased estimator is UMVU is to verify that it achieves the Cramér–Rao lower bound.

(a) Let us look at a few examples of the Cramér–Rao approach: (i) Suppose $Y \sim N(\theta, 1)$, and show that Y is UMVU for θ . (ii) Suppose Y has an exponential distribution with mean θ . Show that Y is UMVU for θ . (iii) Let $Y_i = \beta_0 + \beta_1 x_i + \sigma \varepsilon_i$ for $i = 1, \dots, n$,

where the covariates $x_1, \dots, x_n$ are fixed numbers, and the $\varepsilon_1, \dots, \varepsilon_n$ are independent standard normal random variables. Show that the least squares estimator for (β_0, β_1) is the UMVU estimator.

(b) Let $Y_1, \dots, Y_n$ be i.i.d. $N(\mu, \sigma^2)$. It is immediate from (a) that $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ is UMVU for μ . Show that the estimator $\hat{\sigma}_n^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2 / (n-1)$ is *not* UMVU for σ^2 . In Ex. 9.6 we will see that there is no unbiased estimator of σ^2 attaining the Cramér–Rao lower bound. emil, repair here

Ex. 9.6 *Cramér–Rao and Cauchy–Schwarz.* The proof of the Cramér–Rao inequality that we met in Ex. 5.7–5.8 is a clever application of the Cauchy–Schwarz inequality, introduced in Ex. 1.30.

(a) Let X and Y be two square integrable random variables with expectation zero. Show that $|E XY| = \{\text{Var}(X) \text{Var}(Y)\}^{1/2}$ if and only if X and Y are linearly related, i.e. with $Y = a + bX$ for some (a, b) .

(b) Explain why the Cramér–Rao inequality is an equality if and only if the estimator $\delta(y)$ and the score function are linearly related, that is

$$u(y, \theta) = \partial \log f(y, \theta) / \partial \theta = a(\theta) + b(\theta) \delta(y), \quad \text{for all } \theta,$$

for some functions $a(\theta)$ and $b(\theta)$. Solve the differential equation above for $f(y, \theta)$, and state what this entails for estimators and distributions when it comes to possible attainment of the Cramér–Rao lower bound.

(c) In Ex. 9.5(b) we saw that with $Y_1, \dots, Y_n$ i.i.d. $N(\mu, \sigma^2)$, the unbiased estimator $\hat{\sigma}_n^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)$ for σ^2 does not attain the Cramér–Rao lower bound. Use the result from (b) to argue that the Cramér–Rao lower bound may only be attained when μ is known.

Ex. 9.7 *Sufficiency and Rao–Blackwell.* Suppose that Y has a distribution from a family $\{P_\theta : \theta \in \Theta\}$ of distributions. Recall from Ex. 4.18 that $T = T(Y)$ is a sufficient statistic for this family if the conditional distribution of Y given T does not depend on θ . The Rao–Blackwell theorem says that any estimator can be improved upon by conditioning on a sufficient statistic. This is an important result, as it tells us that in our search for a best estimator, we need only consider those estimators that are functions of a sufficient statistic.

(a) Let $\tilde{a} = \tilde{a}(Y)$ be an unbiased estimator for $g(\theta)$, and suppose that $T = T(Y)$ is sufficient for θ . Consider the estimator given by $\hat{a} = E_\theta(\tilde{a} | T)$. First, explain why $\hat{a}$ is an estimator for $g(\theta)$; second, show that $\hat{a}$ is unbiased; and third, show that $\text{Var}_\theta \hat{a} \leq \text{Var}_\theta \tilde{a}$ for all θ . You have now proven the Rao–Blackwell theorem for unbiased estimators. In Ex. 9.11 we look at this theorem in a more general decision theoretic framework. Rao–Blackwell theorem

(b) Suppose that $Y_1, \dots, Y_n$ are i.i.d. uniforms on $[0, \theta]$, with $\theta > 0$ an unknown parameter. Show that the estimators

$$\tilde{\theta} = \frac{2}{n} \sum_{i=1}^n Y_i \quad \text{and} \quad \hat{\theta} = \frac{n+1}{n} \max_{i \leq n} Y_i$$

are both unbiased for θ . There are (at least) two ways of showing that $\hat{\theta}$ is a better estimator than $\tilde{\theta}$. One involves computing the variances of both estimators; the second uses the Rao–Blackwell theorem. For the Rao–Blackwell approach, use the results from Ex. 3.17 to establish that $E(\tilde{\theta} | \max_{i \leq n} Y_i) = \hat{\theta}$ almost surely. Try both methods.

(c) Let $Y_1, \dots, Y_n$ be i.i.d. $\text{Pois}(\lambda)$. We seek to estimate $\theta = \Pr(Y_1 = 0) = \exp(-\lambda)$. Find the ML estimator for θ , say $\hat{\theta}$, and show that $E\hat{\theta} = \theta\{1 + O(1/n)\}$, meaning the ML estimator for θ is biased. Next, find the Cramér–Rao lower bound for unbiased estimators of θ . Look back at Ex. 9.6 and consider whether this lower bound can be attained. Another estimation strategy is to estimate $\theta = \Pr(Y_1 = 0)$ by the share of zero counts, that is $\tilde{\theta} = n^{-1} \sum_{i=1}^n I\{Y_i = 0\}$. This estimator is clearly unbiased for θ , why is it not best unbiased? Finally, with the aid of the sufficient statistic $T(Y) = \sum_{i=1}^n Y_i$ we Rao–Blackwellise the estimator $\tilde{\theta}$. Derive an expression for this Rao–Blackwellised estimator, $\tilde{\theta}_{\text{rb}}$, say. Does $\tilde{\theta}_{\text{rb}}$ attain the Cramér–Rao lower bound?

Ex. 9.8 *Best unbiased, completeness, and Lehmann–Scheffé.* From the Rao–Blackwell theorem we know that any candidate for being an UMVU estimator must be a function of a sufficient statistic. In this exercise we first establish that in our search for the UMVU estimator, we are only looking for one estimator, i.e., the UMVU estimator is unique. Second, it is shown that an estimator is best unbiased if and only if it is uncorrelated with all unbiased estimators of zero. This yields a characterisation of the best unbiased estimators, albeit one of limited utility as it is, without further conditions, hard to describe all unbiased estimators of zero. Finally, completeness – a condition on the distribution of the data – is introduced, ensuring that the only unbiased estimator of zero is zero itself.

(a) Suppose that both $\tilde{a}$ and $\hat{a}$ are UMVU for $g(\theta)$. Define the estimator $\delta = \frac{1}{2}\tilde{a} + \frac{1}{2}\hat{a}$, and show that $\text{Var}_{\theta} \delta \leq \text{Var}_{\theta} \tilde{a}$. But since $\tilde{a}$ is an UMVU estimator and δ is unbiased (check it), it must be the case that $\text{Var}_{\theta} \delta = \text{Var}_{\theta} \tilde{a}$. Look back at the results in Ex. 9.6, and establish that if $\tilde{a}$ and $\hat{a}$ are both UMVU, then $\tilde{a} = \hat{a}$ almost surely, for all θ .

characterisation
of the UMVU
estimator

(b) Suppose that the estimator $\hat{a}$ is a function of a sufficient statistic and is unbiased for $g(\theta)$. How may we improve upon this estimator? Well, the family of estimators $\tilde{a}_b = \hat{a} + b\varepsilon$ as b ranges of the real numbers, and ε is some mean zero random variable, constitute a class of unbiased estimators. Show that if $\text{cov}_{\theta}(\hat{a}, \varepsilon) \neq 0$ for some θ , then b may be chosen so that $\text{Var}_{\theta} \tilde{a}_b < \text{Var}_{\theta} \hat{a}$ for some value(s) of θ , in which case $\hat{a}$ is not best unbiased. Prove the converse, namely that if $\hat{a}$ is unbiased and $\text{cov}_{\theta}(\hat{a}, \varepsilon) = 0$ for all θ and all mean zero random variables ε , then $\hat{a}$ is UMVU. In summary, an estimator is best unbiased if and only if it is uncorrelated with all unbiased estimators of zero.

completeness

(c) It is in general no easy task to show that an unbiased estimator, or more generally, a statistic $S = S(Y)$ say, is uncorrelated with all unbiased estimators of zero. Since the correlation between any random variable and zero is zero, the task would be much easier if we knew of the distribution of S that the only unbiased estimator of zero, is zero itself. That is, if for any measurable function h and for any θ , $E_{\theta} h(S) = 0$ implies that $h(S) = 0$ almost surely. Recall from Ex. 4.24 that a family of distributions with this property is called *complete*. Alternatively, we just say that the statistic $S(Y)$ is complete. Now, suppose that $T = T(Y)$ is sufficient and complete for θ . Let $\hat{a} = \hat{a}(T)$ be unbiased for $g(\theta)$. Prove the Lehmann–Scheffé theorem, that is, show that the estimator $\hat{a}$ is the unique UMVU estimator for $g(\theta)$.

Lehmann–
Scheffé theorem

(d) Return to Ex. 9.7(b), and show that the estimator $\hat{\theta}$ is the UMVU estimator.

(e) The completeness requirement in the Lehmann–Scheffé theorem was motivated by the characterisation in (b), saying that an estimator is UMVU if and only if it is uncorrelated with all unbiased estimators of zero. A perhaps more illuminating motivation comes from the fact, proven in Ex. 9.7(a), that a best estimator must be based on a sufficient

statistic. Intuitively, by getting rid of information irrelevant to the estimation problem at hand, we reduce the variance of our estimation procedure. Taking this intuition to its logical conclusion, we deduce that a best estimator must be based on a statistic achieving the maximum amount of data compression, while still retaining all the information in the data about the parameter we seek to estimate. In other words, a best estimator must be based on a minimal sufficient statistic. Recall from Ex. 4.22 that a statistic S is minimal sufficient if for any sufficient statistic T there exists a measurable function g so that $S = g(T)$. Suppose that δ is an unbiased estimator and form the estimators $\tilde{a} = E_\theta(\delta | T)$ and $\hat{a} = E_\theta(\delta | S)$, where T is sufficient and S is minimal sufficient. Show that then $\text{Var}_\theta \hat{a} \leq \text{Var}_\theta \tilde{a}$ for all θ . Next, suppose that T is sufficient *and* complete. Use the Lehmann–Scheffé theorem and (a) to conclude that $\hat{a} = \tilde{a}$ almost surely, for each θ . In view of this equality, it may not come as a surprise that if T is sufficient and complete, then T is minimal sufficient. A fact we will prove in (g).

(f) Here is a toy example illustrating some of the points made in (e). Let X_1 and X_2 be independent Bernoulli(θ) random variables and consider the estimator $\hat{\theta} = (X_1 + X_2)/2$ and the estimator $\tilde{\theta}$ defined by

$$\tilde{\theta}(x_1, x_2) = \begin{cases} 1, & (x_1, x_2) = (1, 1), \\ 2/3, & (x_1, x_2) = (1, 0), \\ 1/3, & (x_1, x_2) = (0, 1), \\ 0, & (x_1, x_2) = (0, 0). \end{cases}$$

Explain why both $\hat{\theta}$ and $\tilde{\theta}$ are sufficient for θ , and that they are both unbiased for θ . Show that the variance of $\tilde{\theta}$ exceeds the variance of $\hat{\theta}$ for all values of $\theta \in (0, 1)$, and explain how this relates to minimal sufficiency.

(g) The result quoted at the end of (e) is Bahadur's theorem: If T is sufficient and complete, then T is minimal sufficient. To prove this, let S be another sufficient statistic, and assume, with out loss of generality, that T and S are real valued. We must show that there is a function g such that $T = g(S)$. If so happens that $T = E_\theta(T | S)$, then we have found a function g : It is $g(s) = E_\theta(T | S = s)$, and g does not depend on θ since S is sufficient. Let us therefore prove that T equals $E_\theta(T | S)$ almost surely for all θ . To this end, assume that T has finite variance, and define $g(S) = E_\theta(T | S)$ and $h(T) = E_\theta\{g(S) | T\}$. Now, use the tower property of conditional expectation a couple of times and the fact that T is complete to show that $T = h(T)$ almost surely, for all θ . Next, combine the above with the variance decomposition formula to obtain

Bahadur's
theorem

emil, repair a
bit

$$\text{Var}_\theta g(S) = E_\theta \text{Var}_\theta(g(S) | T) + E_\theta \text{Var}_\theta(T | S) + \text{Var}_\theta g(S),$$

from which we conclude that $T = E_\theta(T | S)$ almost surely for all θ . To get rid of the finite variance assumption on T , replace T by $f(T) = 1/\{1 + \exp(-T)\}$ (which clearly has finite variance) throughout the proof, to conclude that $T = f^{-1}(g(S))$.

Ex. 9.9 *A uniform mean.* Let $Y_1, \dots, Y_n$ be i.i.d. $\text{unif}(a, b)$. We are to estimate the mean $\mu = (a + b)/2$.

(a) Show that $(\min_{i \leq n} Y_i, \max_{i \leq n} Y_i)$ is sufficient and complete.

(b) Propose an unbiased estimator for μ , and find its variance.

Ex. 9.10 *A weird unbiased estimator.* Limiting our search for a best estimator to the class of unbiased estimators lacks the decision theoretic foundations that the principle of

minimising expected loss enjoys. More on this in the Notes&Pointers. Sometimes, the search for unbiasedness might lead us astray. Let Y be a random variable with density

$$f(y, \lambda) = \frac{\lambda^y \exp(-\lambda)}{y! \{1 - \exp(-\lambda)\}}, \quad \text{for } y = 1, 2, \dots,$$

with $\lambda > 0$. This is a Poisson distribution truncated at zero, and the probability of being truncated is θ .

- (a) Show that $\hat{\theta} = (-1)^{Y+1}$ is the unique unbiased estimator for $\theta = \exp(-\lambda)$.
- (b) Find an expression for the risk function of $\hat{\theta}$.
- (c) Propose an estimator with uniformly smaller risk than $\hat{\theta}$.

Ex. 9.11 *More Rao–Blackwellisation.* In Ex. 1.30 we proved Jensen's inequality, saying that $g(\mathbb{E} X) \leq \mathbb{E} g(X)$ for convex functions g . Jensen's inequality extends to conditional expectations, as we show in Ex. C.11. In this exercise we consider loss functions $L(\theta, \delta)$ that are convex in δ for all θ . Think of your favourite loss function, and you will realise that this is a quite natural requirement.

- (a) Let $\tilde{\theta}$ be an estimator of θ , and suppose that $T = T(X)$ is sufficient for θ . If we Rao–Blackwellise $\tilde{\theta}$, that is, define $\hat{\theta} = \mathbb{E}_{\theta}(\tilde{\theta} | T)$ we obtain a new estimator (as shown in Ex. 9.7). Suppose that the loss function $L(\theta, a)$ is convex in a for all θ , and show that

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} L(\theta, \hat{\theta}) \leq \mathbb{E}_{\theta} L(\theta, \tilde{\theta}) = R(\theta, \tilde{\theta}), \quad \text{for all } \theta.$$

- (b) Under what conditions is the inequality in (a) a strict inequality for all θ ?
- (c) Suppose that $\mathbb{E}_{\theta} \hat{\theta}(T) = h(\theta)$ for all θ , and that T is complete. Show that, in the class of estimators $\{\delta: \mathbb{E}_{\theta} \delta = h(\theta), \text{ for all } \theta\}$, the estimator $\hat{a}$ is the unique estimator minimising the risk.
- (d) [xx Let $L(\theta, \delta) = (\delta - \theta)^2$ and specialise to UMVU estimator xx]
- (e) Suppose $L(\theta, a)$ is convex in a for each θ , and that a_{π} is the unique Bayes solution under the prior π . Show that a_{π} must be a function of a sufficient statistic.

Minimaxity

Ex. 9.12 *Tools for minimaxity.* An estimator is minimax if it minimises the risk in the worst possible scenario. That is, $\hat{a}$ is minimax if $\sup_{\theta} R(\theta, \hat{a}) \leq \sup_{\theta} R(\theta, \tilde{a})$ for all estimators $\tilde{a}$. In this exercise we develop some techniques for verifying that an estimator is minimax.

- (a) Consider any prior π and any decision procedure $\hat{a}$, with consequent risk function $R(\theta, \hat{a})$ and Bayes risk $\text{BR}(\hat{a}, \pi)$. Writing $R_{\max}(\hat{a}) = \sup_{\theta \in \Theta} R(\theta, \hat{a})$ for the maximum risk, show that $\text{BR}(\hat{a}, \pi) \leq R_{\max}(\hat{a})$, and that $\text{MBR}(\pi) \leq R_{\max}(\hat{a})$.
- (b) Let $\hat{a}_{\pi}$ be a Bayes solution with respect to the prior distribution π , and suppose that

$$\text{BR}(\hat{a}_{\pi}, \pi) = R_{\max}(\hat{a}_{\pi}), \tag{9.2}$$

that is, $\text{MBR}(\pi) = R_{\max}(\hat{a}_{\pi})$. Show that $\hat{a}_{\pi}$ is minimax.

- (c) Show that if $\hat{a}_{\pi}$ satisfies (9.2) and is the unique Bayes solution with respect to π , then $\hat{a}_{\pi}$ is the unique minimax procedure.
- (d) Show that if a Bayes solution has constant risk, then it is minimax.

(e) Show that if an estimator has constant risk and is admissible, it is minimax.

Ex. 9.13 *Minimax estimators for binomial and multinomial parameters.* In Ex. 7.8 and 7.15 we worked through certain Bayesian priors and their associated posteriors and posterior means, along with results for their risk functions. Those efforts now lead quickly to the additional quality marks for the derived estimators: they are minimax. The points below merely spell this out.

(a) First, consider $Y \sim \text{binom}(n, p)$, with p to be estimated with quadratic loss. Show that the Bayes estimator with the $\text{Beta}(\frac{1}{2}\sqrt{n}, \frac{1}{2}\sqrt{n})$ prior is $p^* = (\frac{1}{2} + Y/\sqrt{n})/(\sqrt{n} + 1)$, that its risk function is the constant $\frac{1}{4}/(\sqrt{n} + 1)^2$, and that it is minimax. Show also that it is fact admissible. For which interval of p is the minimax risk smaller than $R(p, Y/n)$ for the usual default estimator?

(b) Then consider a multinomial $(Y_1, \dots, Y_k)$, with parameters $(n, p_1, \dots, p_k)$, with sums $\sum_{j=1}^k Y_j = n$ and $\sum_{j=1}^k p_j = 1$, i.e. there are $k - 1$ free parameters for the k categories. With joint estimation of the probabilities, with loss function $\sum_{j=1}^k (a_j - p_j)^2$, start with the Dirichlet prior $\sqrt{n}(1/k, \dots, 1/k)$, show that the posterior mean estimator takes the form

$$p_j^* = \frac{1/k}{\sqrt{n} + 1} + \frac{\sqrt{n}}{\sqrt{n} + 1} \frac{Y_j}{n} \quad \text{for } j = 1, \dots, k.$$

and verify that its risk function is constant $(1 - 1/k)/(\sqrt{n} + 1)^2$. Explain that it is minimax and admissible. In which region of $(p_1, \dots, p_k)$ is the minimax estimator better than the ML estimator, i.e. $(Y_1/n, \dots, Y_k/n)$?

(c) Explain the implications of this minimax find when it comes to constructing histograms for continuous data, see Ex. 1.6 for pertinent details, say with equal binwidths.

Ex. 9.14 *Minimaxity and sequences of priors.* In Ex. 9.12(c) we assumed that the equality in (9.2) is attained. A prior distribution that succeeds in attaining this equality is, for natural reasons, called a *least favourable* prior distribution. If no such prior distribution exists, we cannot use the conclusion of that exercise to prove minimaxity. Consider independent $X_1, \dots, X_n | \theta$ from $N(\theta, 1)$. It seems reasonable that a least favourable prior for θ should spread its mass evenly out on the real line, that is

$$\int_a^{a+c} \pi(\theta) d\theta = \int_b^{b+c} \pi(\theta) d\theta, \quad \text{for all } a, b \in \mathbb{R} \text{ and } c > 0.$$

The measure $\pi(\theta)d\theta$ satisfying this equality is Lebesgue measure on $\mathbb{R}$, and is not a probability distribution, and so Ex. 9.12(c) is not immediately applicable. To fix this, the idea is to approximate an improper distribution (i.e., a measure not integrating to one) with proper ones. In the case of the normals, one may try $\theta \sim \pi_k(\theta)$, where $\pi_k(\theta)$ is the density of a uniform distribution over $[-k, k]$, then let k grow.

improper priors

(a) Let $\hat{a}$ be an estimator, $(\pi_k)_{k \geq 1}$ a sequence of prior distributions, and for each $k \geq 1$ let $\hat{a}_{\pi_k}$ be the Bayes solution for π_k . Suppose that

$$\lim_{k \rightarrow \infty} \text{BR}(\hat{a}_{\pi_k}, \pi_k) = R_{\max}(\hat{a}) < \infty.$$

Show that $\hat{a}$ is then minimax.

(b) Let $X_1, \dots, X_n$ be independent $N(\xi, \sigma^2)$. We are to estimate ξ under the squared error loss $L(\xi, \hat{\xi}) = (\hat{\xi} - \xi)^2$. Consider the sequence of priors $\mu \sim N(0, \tau_k)$ for $k = 1, 2, \dots$, and show that the estimator $\hat{\xi} = \bar{X}_n$ is minimax.

(c) Let $X_1, \dots, X_n$ be independent $\text{Pois}(\theta)$. We want to estimate θ under the weighted loss function $L(\theta, \hat{\theta}) = \theta^{-1}(\hat{\theta} - \theta)^2$. Use Gamma priors to show that $\hat{\theta} = \bar{X}_n$ is minimax.

Ex. 9.15 *A Poisson with parameter restriction.* (xx match this properly with the multi-poisson thing later. xx) Consider a single $Y \sim \text{Pois}(\theta)$, with the parameter to be estimated with weighted loss $(\delta - \theta)^2/\theta$.

(a) Explain first that the risk function for the obvious estimator Y is the constant 1. Then work with the prior $\theta \sim \text{Gam}(1, b)$, for b positive: (i) derive the formula $\hat{\theta}_B = y/(1+b)$ for the Bayes estimator, with some care required for the case of $y = 0$; then (ii) find that the minimum Bayes risk $\text{MBR} = 1/(b+1)$. Argue that Y consequently has the minimax property.

(b) To illustrate a more general modification, or aspect, concerned with restricted parameter regions, assume that it is known a priori that $\theta \geq \frac{1}{2}$. Consider the ‘we do not believe in zero’ estimator $\hat{\theta} = T(y)$, which is y if $y \geq 1$ but moves 0 to 1 in case $y = 0$. Find an explicit formula for its risk function $R(\theta, \hat{\theta})$, and plot it. Verify that it is minimax, like the untruncated Y , but that there is clear risk improvement over a sizeable interval: extra knowledge yields gain.

(c) Then, also for the sake of learning more about risk functions etc., go to the case of $n = 2$ Poissons Y_1 and Y_2 , still with the $\theta \geq \frac{1}{2}$ constraint on the parameter. Work out the risk functions for (i) $\hat{\theta}_A = \frac{1}{2}\{T(y_1) + T(y_2)\}$, truncating both y_1 and y_2 , in cases where one of these are zero; and (ii) $\hat{\theta}_B = T(\frac{1}{2}(y_1 + y_2))$, truncating only the average. Verify that both are better than the unrestricted Poisson mean $\frac{1}{2}(y_1 + y_2)$. Which procedures are best, in which intervals?

Admissibility and Bayes

Ex. 9.16 *Tools for admissibility.* If we have at hand an estimator $\hat{a}$, the most convenient way of showing that $\hat{a}$ is admissible is to show that it is the Bayes solution for some prior. In fact, it is *almost* true that an estimator is admissible if and only if it is Bayes. We will get to the cases where this implication fails, but as a rule of thumb it is pretty safe.

(a) Suppose that $X \sim f_\theta$, where $\theta \in \Theta = \{\theta_1, \dots, \theta_k\}$ for some finite $k \geq 2$. Consider the estimator $\hat{a}_\pi$ that is Bayes for the prior $\pi = \{\pi_1, \dots, \pi_k\}$, where π_j is the prior mass given to θ_j . Show that if $\pi_j > 0$ for $j = 1, \dots, k$, then $\hat{a}_\pi$ is admissible.

(b) Why does the conclusion in (a) fail if $\pi_j = 0$ for one or more j ?

(c) Show that if an estimator is unique minimax, it is admissible.

(d) Show that if a Bayes solution is *unique*, then it is admissible.

(e) To clarify what the uniqueness of the Bayes solution refers to, let's look at an example where there are several Bayes solutions. Let $X \sim \text{unif}(0, \theta)$ and suppose that we want to estimate θ under the squared error loss function $L(\theta, \delta) = (\delta - \theta)^2$. Suppose θ is given the prior distribution that is uniform on $(0, c)$. Find the posterior distribution $\theta | (X = x)$ and derive at least two different Bayes solutions (there are uncountably many). [xx comment on this exercise, more relevant in BNP, point to Nils' 1976-proof and the mistake made by Lehmann. Could also mention Lindley and his so-called Cromwell's rule xx].

(f) Recall that a parameter value θ is in the *support* of π (a probability density in our notation) if it is contained in the set $\{\theta \in \Theta: \pi(\theta) > 0\}$. Suppose that the support of the prior π is Θ , and that the risk function $R(\theta, \delta)$ is continuous in θ for all estimators

δ . Show that if $\hat{a}_\pi$ is Bayes with respect to π and have finite Bayes risk, then $\hat{a}_\pi$ is admissible.

Ex. 9.17 Generalised Bayes. (xx need a bit more nils-emil Ch6-Ch8 calibration; some of this might be in Ch6. xx) Suppose that in some experiment involving data from a normal distribution with expectation θ , you have no idea whatsoever about where on the real line θ might be located. A natural ‘prior’ is therefore $\pi(\theta) \propto 1$ (or just take it equal to one) that spreads the ‘probability’ mass uniformly over the real line. Now, $\pi(\theta) \propto 1$ corresponds to Lebesgue measure on the real line, and is not a probability measure because its integral is infinite. The fact that $\pi(\theta)$ does not integrate to one does not, however, stop us from using it to derive estimators using the Bayes theorem, or a stretched version thereof. Priors that are not probability distributions are called improper priors.

(a) Suppose $Y_1, \dots, Y_n$ are i.i.d. $N(\theta, \sigma^2)$, with known σ but unknown mean. Suppose θ is given the improper prior $\pi(\theta) = 1$. With $\bar{Y}$ as usual the sample mean, show that $\pi(\theta | y_1, \dots, y_n)$ is the normal density $N(\bar{y}, \sigma^2/n)$, noting that this is a real density, as soon as $n \geq 1$ datum is observed, even though the prior is improper. A generalised Bayes estimator is the estimator δ minimising the posterior expected loss $E\{L(\theta, \delta) | \text{data}\}$. With squared error loss, show that the generalised Bayes estimator for θ is $\bar{y}$.

(b) The estimator $\hat{\theta} = \bar{Y}$ you found in (a) is generalised Bayes; why is this not enough to conclude that it is admissible?

(c) (xx decide on balance Ch6 Ch8 here. xx) Let us generalise the above to normal data from $N(\theta, \sigma^2)$ with both parameters unknown. It is mathematically convenient to turn to the parametrisation $\lambda = 1/\sigma^2$. With (θ, λ) having the improper prior $\pi(\theta, \lambda) = 1/\lambda$, and writing $Q_0 = \sum_{i=1}^n (y_i - \bar{y})^2$, show that the posterior becomes

$$\begin{aligned} \pi_n(\theta, \lambda) &\propto \lambda^{n/2-1} \exp\{-\tfrac{1}{2}\lambda\{Q_0 + n(\theta - \bar{y})^2\}\} \\ &\propto \lambda^{1/2} \exp\{-\tfrac{1}{2}\lambda(\theta - \bar{y})^2\} \lambda^{(n-1)/2-1} \exp(-\tfrac{1}{2}\lambda Q_0). \end{aligned}$$

Translate this to saying (i) $(1/\sigma^2) | \text{data} \sim \text{Gam}(\frac{1}{2}(n-1), \frac{1}{2}Q_0)$; (ii) $\theta | (\text{data}, \sigma) \sim N(\bar{y}, \sigma^2/n)$. Show that the posterior mean of $1/\sigma^2$ is $1/\hat{\sigma}^2$ with the familiar empirical variance $Q_0/(n-1)$.

Ex. 9.18 The normal sample mean is admissible, in dimension 1. We tend to general sample size n below, but start wout considering the prototype setup with a single Y from the normal $(\theta, 1)$. The natural estimators $\bar{Y}$ itself. Here we show that it is admissible, under squared error loss, but the arguments used do not work for dimension $k \geq 2$.

(a) We have seen in Ex. 9.14 that $\bar{Y}$ is minimax, with constant risk function 1. Explain via Ex. 7.10 that it is not a Bayes estimator.

(b) We employ the prior $\pi_\tau \sim N(0, \tau^2)$ for θ , with increasing τ , resembling the flat prior under which we have seen that $\bar{Y}$ is generalised Bayes. Show that the Bayes estimator is $\hat{\theta}_\tau = \rho Y$, with $\rho = \tau^2/(1 + \tau^2)$, with minimum Bayes risk equal to this ρ . You may redo that argument leading from this to minimaxity of Y .

(c) Assume now that Y is not admissible. Explain that there must be a competitor $\tilde{\theta}$ with risk at most 1 and with $R(\theta, \tilde{\theta}) \leq 1 - \varepsilon$, for some small ε , for some interval $[a, b]$. Deduce from this that

$$\frac{\tau^2}{\tau^2 + 1} \leq \int R(\theta, \tilde{\theta}) \pi_\tau(\theta) d\theta \leq 1 - \varepsilon \int_{[a, b]} \phi(\theta/\tau) (1/\tau) d\theta,$$

and that

$$\varepsilon(1/\tau) \int_{[a,b]} \phi(\theta/\tau) d\theta \leq 1/(\tau^2 + 1) \quad \text{for all } \tau.$$

Show that this leads to a contradiction: Y is admissible.

(d) Generalise the above to any sample size; with $Y_1, \dots, Y_n$ being i.i.d. $N(\theta, \sigma^2)$, with given σ , then the sample mean $\bar{Y}$ is minimax and admissible.

(e) Consider the k -dimensional situation, with independent components $Y_i \sim N(\theta_i, 1)$ for $i = 1, \dots, k$, and with the vector of means to be estimated with loss $\sum_{i=1}^k (a_i - \theta_i)^2$. Using a sequence of independent priors $N(0, \tau^2)$, for $i = 1, \dots, k$, show that $Y = (Y_1, \dots, Y_k)$ is minimax, with constant rate k . Then try the argument above, and explain that it does not work, for $k \geq 2$. We learn indeed in Ex. 9.25 that Y is inadmissible, for $k \geq 3$; there are competitors which uniformly dominate Y .

Ex. 9.19 Blyth's method. We saw in Ex. 9.18 that admissibility may be established via contradiction. Here we formalise that argument, sometimes called Blyth's method. When it works, it works, often in a case by case modus; we saw for the case of a multidimensional normal mean that the Blyth method only worked for the one-dimensional case. In general, say that δ is a *limiting Bayes* estimator if there is a sequence $\{\pi_k\}_k$ of possibly improper priors such that the corresponding Bayes estimators δ_{π_k} converge almost surely to δ .

(a) Let $\hat{a}$ be an estimator. Suppose $\Theta \subset \mathbb{R}^p$ is open, and that $R(\theta, a)$ is continuous in θ for all a . Let $(\pi_k)_{k \geq 1}$ be a sequence of (possibly improper) prior distributions such that $\text{BR}(\hat{a}, \pi_k) < \infty$ for all k , and for any open set $\Theta_0 \subset \Theta$,

$$\frac{\text{BR}(\hat{a}, \pi_k) - \text{BR}(\hat{a}_{\pi_k}, \pi_k)}{\int_{\Theta_0} \pi_k(\theta) d\theta} \rightarrow 0, \quad \text{as } k \rightarrow \infty,$$

where $\hat{a}_{\pi_k}$ is the Bayes solution under π_k . Show that $\hat{a}$ is admissible. This is Blyth's method, which may be quoted, rather imprecisely, as saying that often limits of Bayes estimators are admissible.

but not always

(b) Check arguments from Ex. 9.18 again, to see these general arguments in action, leading to admissibility for normal means in dimension 1 but not in higher dimension.

(c) Consider $Y \sim \text{Pois}(\theta)$ with loss function $(a - \theta)^2/\theta$. Show that the estimator Y has constant risk. Use Gamma priors $(1, b)$, with growing b , to show that Y is minimax and admissible. Attempt to see how far the argument goes for the case of $k \geq 2$ Poisson variables, with loss function $\sum_{i=1}^k (a_i - \theta_i)^2/\theta_i$; see Ex. 9.30.

Ex. 9.20 One Poisson: Y is minimax and admissible. We have seen in Ex. 9.15 that with a single $Y \sim \text{Pois}(\theta)$, with loss function $(\delta - \theta)^2/\theta$, the simple estimator Y itself is minimax with constant risk 1. Here we deal with this one-only setup; matters change significantly in the multi-parameter setup, see Ex. 9.30.

(a) Attempt to use Blyth's method of Ex. 9.19 to prove that Y is also admissible, using the $\text{Gam}(1, b)$ prior with small b – and explain where that method fails to give a proof. Instead, use the Cramér–Rao variance bounds of Ex. 5.7, to show that if θ^* is a competitor to Y , with risk bounded everywhere by 1, then we must have

$$b(\theta)^2/\theta + \{1 + b'(\theta)\}^2 \leq 1 \quad \text{for all } \theta,$$

in terms of the bias function $E_\theta \theta^* = \theta + b(\theta)$. Show that $b(\theta) = 0$ is the only solution to this differential inequality, and conclude that there can be no competitor dominating Y .

(b) Generalise the above to the situation where $Y \sim \text{Pois}(w\theta)$, with a known exposure level w .

Ex. 9.21 *Bernoulli mean with weighted risk.* Let $X_1, \dots, X_n$ be i.i.d. Bernoulli(θ). We wish to estimate θ , and we are particularly interested in precise estimates of very small and very large values of θ . Therefore, we will work with the loss function

$$L(\theta, \delta) = (\delta - \theta)^2 / \{\theta(1 - \theta)\}.$$

(a) Compute the risk function of the maximum likelihood estimator. What's noticeable about this risk function?

(b) We now take a Bayesian point of view and give θ a Beta($a\theta'$, $a(1 - \theta')$) prior distribution. Compute the expectation and variance of this prior.

(c) With the prior introduced in (b), find the posterior distribution of θ . Derive also an expression for the Bayes solution δ_π .

(d) Tweak the parameters of the Beta prior distribution, so that the Bayes solution you found above equals the ML estimator from (a). What desirable properties does the ML estimator possess?

Ex. 9.22 *Estimating the standard deviation.* Suppose $X_1, \dots, X_n$ are i.i.d. from $N(0, \sigma^2)$. We are to estimate σ under the loss function $L(\sigma, \delta) = (\delta - \sigma)^2 / \sigma$.

(a) Find the ML estimator, say $\hat{\sigma}_{\text{ml}}$, and show that its risk function is

$$R(\sigma, \hat{\sigma}) = \sigma \{(n-1)/n + b_n^2 + (b_n - 1)^2\}, \quad \text{where} \quad b_n = \sqrt{\frac{2}{n}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

You may now use Stirling's formula $\Gamma(z) = (2\pi/z)^{1/2} (z/e)^z$ to show that $b_n \rightarrow 1$, and $R(\sigma, \hat{\sigma}) \rightarrow 2\sigma$ as $n \rightarrow \infty$, as we already knew from ML-theory (see Ex. xx in Chapter 5).

(b) Consider the prior distribution $\sigma \sim \pi(\sigma)$, whose density is

$$\pi(\sigma) \propto (1/\sigma)^{a+1} \exp(-b/\sigma^2).$$

with $a > 1$ and $b > 0$. Find the prior expectation of σ and of $1/\sigma$.

(c) Find the posterior distribution $\sigma | x_1, \dots, x_n$, and derive the Bayes solution under the loss function above.

(d) Show that the ML estimator is inadmissible by exhibiting an estimator, say δ^* , with uniformly smaller risk. *Hint:* Consider $\delta_\alpha = \alpha \hat{\sigma}_{\text{ml}}$.

(e) Is δ^* admissible? *Hint:* Use Blyth's method.

(f) (xx Nils not entirely convinced by details above; here some notes. xx) We have $\hat{\sigma}_{\text{ml}}^2 = n^{-1} \sum_{i=1}^n X_i^2$, so $\hat{\sigma}_{\text{ml}} = \sigma V_n^{1/2}$, with $V_n \sim \chi_n^2/n$. so

$$\text{E}(\hat{\sigma}_{\text{ml}} - \sigma)^2 = \sigma^2 \text{E}(V_n^{1/2} - 1)^2 = \sigma^2 2(1 - c_n),$$

where

$$c_n = \text{E} V_n^{1/2} = \frac{\sqrt{2}}{\sqrt{n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)}.$$

With $\sigma^* = a_n \hat{\sigma}_{\text{ml}} = a_n \sigma V_n^{1/2}$, we have

$$\text{E}(a_n \hat{\sigma}_{\text{ml}} - \sigma)^2 = \sigma^2 \text{E}(a_n V_n^{1/2} - 1)^2 = \sigma^2 (a_n^2 - 2a_n c_n + 1),$$

with the best value being $a_n = c_n$, for which this best risk is $\sigma^2(1 - c_n^2)$. we ought to have risk converging to $\frac{1}{2}\sigma/n$, not $2\sigma/n$? Is Emil aiming for Blyth working best for the $1/\sigma$ scaling?

Ex. 9.23 *Bayes and minimax normal estimation with the linex loss.* We worked out some basic properties of the linex loss function $\exp\{c(t - \theta)\} - 1 - c(t - \theta)$ in Ex. 7.21–7.22. Here we use the Bayesian machinery to find a minimax estimator for the normal mean.

(a) For the stylised case of a single $Y \sim N(\theta, 1)$, we saw in the exercises pointed to that $\hat{\theta} = Y - \frac{1}{2}c$ has constant risk $\frac{1}{2}c^2$, and that the Bayes estimator under the $N(0, \tau^2)$ prior is $\hat{\theta}_B = \rho Y - \frac{1}{2}c\rho^2$, with $\rho = \tau^2/(\tau^2 + 1)$. As part of the posterior mean loss minimisation, show that $\text{MBR}(N(0, \tau^2)) = \frac{1}{2}c^2\rho$. Use this to show that $\hat{\theta}$ is minimax. Show also, via Blyth's method, that it is in fact admissible.

(b) Generalise the above to the case of a full sample $X_1, \dots, X_n$ from $N(\theta, 1)$. Find the Bayes estimator and associated minimum Bayes risk, for the $N(0, \tau^2)$ prior, and prove that $\theta^* = \bar{X} - \frac{1}{2}c/n$ is minimax. What is its minimax risk?

(c) Find the distribution of $Z_n = \sqrt{n}(\theta^* - \theta)$, and comment on its limit, (i) when the loss-skewness parameter c is fixed, (ii) when $c = \sqrt{n}$.

(d) (xx perhaps another example with linex loss. and something where we see certain arguments lead to a choice of c . xx)

Ex. 9.24 *More on the linex loss.* For the linex loss, studied initially in Ex. 7.21 and then in Ex. 9.23 for the normal case, we now find out more. (xx pointer to approximation via $M(-c)$, the mgf for θ given data. we also look at $c = c_0\sqrt{n}$, a skewness remaining for growing n , avoiding simple symmetry there. xx)

(a) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the $\text{Pois}(\theta)$, with prior $\theta \sim \text{Gam}(a, b)$, as with Ex. 7.2. Show that the Bayes estimator with the linex loss is

$$\hat{\theta}_B = (1/c)(a + n\bar{y}) \log\{1 + c/(b + n)\}.$$

Verify that when $c \rightarrow 0$, we retrieve the posterior means of Ex. 7.2.

(b) (xx one more case with a clear formula. perhaps σ in normal. xx)

(c) As we know from Ex. 7.39, an approximation to the posterior distribution is $\theta | \text{data} \sim N(\hat{\theta}_{\text{ml}}, \hat{\sigma}^2/n)$, in terms of the ML estimate and estimated inverse Fisher information. Deduce that $M_n(-c) \doteq \exp(-c\hat{\theta}_{\text{ml}} + \frac{1}{2}c^2\hat{\sigma}^2/n)$, in the notation above, and that this leads to the approximation $\hat{\theta}_B = \hat{\theta}_{\text{ml}} - \frac{1}{2}c\hat{\sigma}^2/n$ for the Bayes estimator under linex loss. Show also that the posterior expected loss is approximately $\frac{1}{2}c^2\hat{\sigma}^2/n$.

(d) (xx an example where we can check the approximation with the exact Bayes estimator, e.g. with Poisson and gamma. xx)

Multiparameter estimation and the Stein phenomenon

For the following exercises, check also with partly related multiparameter exercises in Ch. 7, from the Bayesian construction point of view, as with Ex. 7.40, 7.41; here the emphasis is more one constructing estimators with good risk functions, dominance, inadmissibility.

Ex. 9.25 *The James–Stein estimator.* When estimating the price of apples in Oslo, the height of women in Bergen, and the unemployment rate in Trondheim, it is sometimes

advantageous to use information about apples in Oslo and women in Bergen to say something about the unemployment rate in Trondheim. The point is that when estimating an ensemble of unrelated quantities, we can sometimes do better by borrowing information across unrelated things. This phenomenon is known as Stein's paradox or the Stein effect; (i) it might be clever to combine these independent estimators, gaining precision for each case; and (ii) even in cases where not much is gained, it will never be worse than not-combining-at-all. In the present exercise we'll look at Stein's 1956 and 1961 result, a result that initiated a whole field of statistical research known by the labels of shrinkage, ensemble, small area estimation. So let $Y_i \sim N(\theta_i, 1)$ be independent for $i = 1, \dots, k$. We are to estimate $\theta_1, \dots, \theta_k$ under the combined loss function

$$L(\theta, a) = \sum_{i=1}^k (a_i - \theta_i)^2.$$

We actually studied several aspects of this situation in Ex. 7.40, from Bayesian and empirical Bayesian perspectives, and succeeded in finding good ensemble procedures. Presently the focus is more on risk functions and uniform dominance.

(a) The standard approach is to use Y_i as an estimator of θ_i . This Y_i is the ML estimator, it is admissible, it is the UMVU estimator, etc. For this standard estimator $Y = (Y_1, \dots, Y_k)$, show that its risk function is the constant k , and that Y indeed is minimax, a property it needs to share with many other schemes, as we learn below.

(b) For a single $Y \sim N(\theta, 1)$, show using integration by parts that under very mild conditions on the function $b(y)$, one has $E_\theta(Y - \theta)b(Y) = E_\theta b'(Y)$, where b' is the derivative of b .

(c) Let now $b(y) = (b_1(y), \dots, b_k(y))$. Generalise what you found in (b) to

$$E_\theta(Y_i - \theta_i)b_i(Y) = E_\theta b_{i,i}(Y), \quad \text{where } b_{i,i}(y) = \partial b_i(y)/\partial y_i.$$

(d) What you found in (b) and (c) is known as Stein's lemma. We shall use this to construct an estimator that uniformly dominates Y . Consider a general competitor to Y of the form $\delta(Y) = (\delta_1(Y), \dots, \delta_k(Y))$, with

$$\delta_i(Y) = Y_i - b_i(Y) \quad \text{for } i = 1, \dots, k.$$

Show that the difference in risk between Y and estimators of this form can be expressed as $R(\theta, \delta) - R(Y, \theta) = E_\theta D(Y)$, where

$$D(y) = \sum_{i=1}^k \{b_i(y)^2 - 2b_{i,i}(y)\}.$$

The clever thing about $R(\theta, \delta) = k + E_\theta D(Y)$ is that $D(y)$ does not depend on the unknown $\theta_1, \dots, \theta_k$. We can therefore try to find a data dependent function $b(y)$ such that $D(y) < 0$ for all y , and in such a case identify an estimator that uniformly dominates the standard estimator. This turns out to be possible for $k \geq 3$, but not for $k \leq 2$.

(e) Try $b_i(y) = ay_i/\|y\|^2$, with $\|y\|^2$ being the squared Euclidean norm $\sum_{i=1}^p y_i^2$, leading to the estimator $\delta(y) = y - b(y) = (1 - a/\|y\|^2)y$. With this choice of $b(y)$, show that

$$D(y) = \{a^2 - 2a(k-2)\}/\|y\|^2.$$

Show that this is negative for a range of a values, provided $k \geq 3$. Demonstrate that the optimal a is $a = k - 2$, corresponding to the estimator

$$\delta_{\text{stein}} = \{1 - (k - 2)/\|Y\|^2\} Y. \quad (9.3)$$

This estimator is known as the James–Stein estimator. Show that its risk can be expressed as

$$R(\theta, \delta_{\text{stein}}) = k - (k - 2)^2 \mathbb{E}_\theta 1/\|Y\|^2.$$

In particular, explain that the classic estimator $(Y_1, \dots, Y_k)$ for $(\theta_1, \dots, \theta_k)$ is inadmissible, for any $k \geq 3$. Show that the greatest reduction in risk from using δ_{stein} instead of Y takes place when $\theta_1 = \dots = \theta_k = 0$, and compute the risk $R(0, \delta_{\text{stein}})$ in this point.

(f) (xx more; link to Ex. 7.40. xx) We'll now make a connection to empirical Bayes procedures. Start with a prior that takes $\theta_1, \dots, \theta_k$ independent from $N(0, \tau^2)$. Show that the Bayes solution is $\delta^B = (\delta_1^B, \dots, \delta_k^B)$, with

$$\delta_i^B(Y) = \rho Y_i \text{ for } i = 1, \dots, k, \quad \text{where } \rho = \tau^2/(\tau^2 + 1).$$

The empirical Bayes approach consists in estimating hyperparameters from data. Hyperparameters are those parameters set by the statistician in a pure Bayesian approach. Show that the marginal distribution of $y_1, \dots, y_k$ is a product of $N(0, 1 + \tau^2)$ distributions. Find the ML estimator of α . Use the ML estimator to find an unbiased estimator, say $\tilde{\alpha}$, of α . The empirical Bayes estimator is then $\delta_{\text{EB}}(Y) = \tilde{\alpha}Y$. What's noticeable about this estimator?

(g) The zero position used actively above is really an arbitrary position in the parameter space, and we may indeed shrink the ML to any given θ_0 . With $k \geq 3$, show that arguments used in that exercise can be used to show that

$$\hat{\theta}_{\text{stein}} = \theta_0 + \{1 - (k - 2)/W\}(Y - \theta_0)$$

has risk function uniformly smaller than the k for the ML, where $W = \|Y - \theta_0\|^2 = \sum_{i=1}^k (Y_i - \theta_{0,i})^2$.

Ex. 9.26 *Estimating many correlations, or many standard deviations.* The methods of Ex. 9.25 have been developed in the clean stylised case of many normal means with independent $Y_i \sim N(\theta_i, 1)$. They may hence be applied also when the Y_i represent transformation-to-normality statistics. For illustration of this general point, consider empirical correlations $R_1, \dots, R_k$, estimates of $\rho_1, \dots, \rho_k$, for independent binormal situations. For simplicity we take the sample sizes to be the same, n , for each. How can Stein estimation help us better estimate the ensemble of ρ_i ? Below, let r_0 be any reasonable correlation value, say a prior guess for the average of the ρ_i ; it could of course be zero.

(a) Use Ex. 2.55 to transform the R_i to $Y_i = \sqrt{n}\hat{\zeta}_i$, with $\hat{\zeta}_i = \zeta(R_i)$ the zeta transform. With $\zeta_0 = \zeta(r_0)$, explain that $Y_i \approx_d N(\sqrt{n}\zeta_i, 1)$. Use Ex. 9.25 to show that the estimator

$$\hat{\zeta}_{\text{stein},i} = \zeta_0 + \{1 - (k - 2)/W\}(\hat{\zeta}_i - \zeta_0) \quad \text{for } i = 1, \dots, k$$

dominates the direct $(Y_1, \dots, Y_k)$ (modulo the normal approximation), with $W = \sum_{i=1}^k n(\hat{\zeta}_i - \zeta_0)^2$. Explain how to transform back to $\hat{\rho}_{\text{stein},i}$. Comment on the cases W small and W big.

(b) For a similar illustration of the general principle, consider empirical standard deviations $\hat{\sigma}_1, \dots, \hat{\sigma}_k$, for k different but perhaps comparable normal samples of size n . Let σ_0 be any reasonable value for the envisaged level of the σ_i . Demonstrate first that $(2n)^{1/2} \log \hat{\sigma}_i \approx N((2n)^{1/2} \log \sigma_i, 1)$. Give a Stein transform-and-back recipe for joint estimation of the σ_i .

Ex. 9.27 *Multiparameter Stein with a covariance matrix.* Suppose $Y \sim N_k(\theta, \Sigma)$ with a known covariance matrix Σ , and with the means to be estimated with Mahalanobis type loss function $(a - \theta)^t \Sigma^{-1} (a - \theta)$.

(a) Show that Y itself has constant risk function k . Using priors of the type $\theta \sim N_k(0, \tau^2 \Sigma)$, with growing τ , prove that Y is indeed minimax.

(b) Let $A = \Sigma^{-1/2}$, so $A^2 = \Sigma^{-1}$ and $A \Sigma A = I$. Using Ex. 9.25, explain that $Z = AY \sim N_k(\gamma, I)$, with $\gamma = A\theta$. For estimating γ with sum of squares loss, argue that

$$\hat{\gamma}_{\text{stein}} = \{1 - (k - 2)/W\}Z, \quad \text{with } W = \|Z\|^2 = Y^t \Sigma^{-1} Y,$$

beats Z uniformly. Transform to

$$\hat{\theta}_{\text{stein}} = A^{-1} \hat{\gamma}_{\text{stein}} = \{1 - (k - 2)/W\}Y,$$

spelling out that this estimator beats Y uniformly. An important special case is as follows: consider independent $Y_i \sim N(\theta_i, \sigma_i^2)$, with loss $\sum_{i=1}^n (a_i - \theta_i)/\sigma_i^2$. With θ_0 some overall value, show that the following estimator is uniformly better than Y :

an illustration:
in Story #20

$$\hat{\theta}_{\text{stein},i} = \theta_0 + \{1 - (k - 2)/W\}(Y_i - \theta_0), \quad \text{where } W = \sum_{i=1}^n (Y_i - \theta_0)^2 / \sigma_i^2.$$

(c) Suppose you need to estimate $k = 100$ pairs of binormal means (a_i, b_i) , based on (X_i, Y_i) with variances 1 and known correlation ρ , and with loss function

$$\begin{aligned} L &= \sum_{i=1}^k \begin{pmatrix} \hat{a}_i - a_i \\ \hat{b}_i - b_i \end{pmatrix}^t \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{a}_i - a_i \\ \hat{b}_i - b_i \end{pmatrix} \\ &= \sum_{i=1}^k \frac{1}{1 - \rho^2} \{(\hat{a}_i - a_i)^2 + (\hat{b}_i - b_i)^2 - 2\rho(\hat{a}_i - a_i)(\hat{b}_i - b_i)\}. \end{aligned}$$

Show that the standard estimator, using (X_i, Y_i) to estimate (a_i, b_i) , has constant risk $2k$. Explain that there is a better estimator for the ensemble of mean pairs, regardless of where the (a_i, b_i) might be in the parameter space:

$$\begin{pmatrix} \hat{a}_{i,\text{stein}} \\ \hat{b}_{i,\text{stein}} \end{pmatrix} = \{1 - (2k - 2)/W\} \begin{pmatrix} X_i \\ Y_i \end{pmatrix},$$

with

$$W = \sum_{i=1}^k \begin{pmatrix} X_i \\ Y_i \end{pmatrix}^t \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \sum_{i=1}^k (X_i^2 + Y_i^2 - 2\rho X_i Y_i) / (1 - \rho^2).$$

Generalise this properly, both to the case of having unequal correlations for the pairs, and when shrinking to some relevant (a_0, b_0) , as opposed to always shrinking to $(0, 0)$.

Ex. 9.28 *Stein type shrinkage to regression structure.* Consider again independent $Y_i \sim N(\theta_i, 1)$ for $i = 1, \dots, k$. The basic Stein estimator was worked with in Ex. 9.25, shrinking the Y_i to some predefined θ_0 , via $Y_i - (c/W_0)(Y_i - \theta_{0,i})$, with $W_0 = \sum_{i=1}^k (Y_i - \theta_{0,i})^2$, and with $c = k - 2$ being the optimal choice. When there is no clear such θ_0 , why not use the sample mean, or perhaps a regression line.

(a) Let therefore $\hat{\theta}_i = Y_i - (c/W)(Y_i - \bar{Y})$, with $W = \sum_{i=1}^k (Y_i - \bar{Y})^2$. Explain that this is within the framework of Ex. 9.25, now with $b_i(y) = (c/W)(y_i - \bar{y})$. With notation from that exercise, work out an expression for the $b_{i,i}(y)$ and find that $R(\theta, \hat{\theta}) = k + E_\theta D(Y)$, with $D(y)$ simplifying to $(1/W)\{c^2 - 2c(k - 3)\}$. Conclude from this that the estimator

$$\hat{\theta}_i = Y_i - \frac{k-3}{W}(Y_i - \bar{Y}) = Y_i - \frac{Y_i - \bar{Y}}{\hat{\tau}^2} = \left(1 - \frac{1}{\hat{\tau}^2}\right)Y_i + \frac{1}{\hat{\tau}^2}\bar{Y},$$

see Story #20
for an
application

in which we write $\hat{\tau}^2 = W/(k - 3)$, is uniformly better than the direct ML estimator, as long as $k \geq 4$. Find the minimum risk, the part of the parameter region where the θ_i are close.

(b) We may also attempt shrinking the datapoints to a regression line, or regression structure. Consider the model $Y_i = x_i^t \beta + \varepsilon_i$, with least squares estimator $\hat{\beta} = M^{-1} \sum_{i=1}^k x_i y_i$, with $M = X^t X$ of full rank p . Then work with

$$\hat{\theta}_i = (1 - c/W)y_i + (c/W)x_i^t \hat{\beta} = y_i - (c/W)(y_i - x_i^t \hat{\beta}) \quad \text{for } i = 1, \dots, k,$$

with $W = \sum_{i=1}^k (y_i - x_i^t \hat{\beta})^2$. Explain that this fits the general setup above, find

$$b_{i,i}(y) = (c/W^2)\{(1 - x_i^t M^{-1} x_i)W - 2(y_i - x_i^t \hat{\beta})^2\}, \quad D(y) = (1/W)\{c^2 - 2c(n - p - 2)\}.$$

So the Stein effect is at work as long as $k \geq p + 3$; explain that this means $k \geq 3$ for the original Stein; $k \geq 4$ when shrinking to common mean; $k \geq 5$ when shrinking to a regression line; etc. Show that the winning shrinkage estimator can be written

$$\hat{\theta}_i = (1 - 1/\hat{\tau}^2)Y_i + (1/\hat{\tau}^2)x_i^t \hat{\beta} = Y_i - (1/\hat{\tau}^2)(Y_i - x_i^t \hat{\beta}) \quad \text{for } i = 1, \dots, k,$$

in terms of $1/\hat{\tau}^2 = (m - p - 2)/W$, and examine its risk function.

(c) (xx then truncation argument. xx)

Ex. 9.29 *Good estimators can be combined.* The setting is as in Ex. 9.25 and 9.28, estimation of an ensemble $\theta_1, \dots, \theta_k$ of normal means, with loss function $\sum_{i=1}^k (\delta_i - \theta_i)^2$. We know that with $\hat{\theta}_i = Y_i - b_i(Y)$, we have $R(\theta, \hat{\theta}) = k + E_\theta D(Y)$, with $D(y) = \sum_{i=1}^k \{b_i(y)^2 - 2b_{i,i}(y)\}$.

(a) Consider two good schemes, with $\hat{\theta}_i^A = Y_i - b_i^A(Y)$ and $\theta_i^B = Y_i - b_i^B(Y)$, both dominating the ML estimator by having the respective $D^A(y)$ and $D^B(y)$ negative. Let us combine them, forming $\hat{\theta}_i = \frac{1}{2}(\hat{\theta}_i^A + \hat{\theta}_i^B)$, which then inherits the same structure, with $b_i = \frac{1}{2}(b_i^A + b_i^B)$. Show that the consequent $D(y)$ function can be written

$$D(y) = \frac{1}{2}D^A(y) + \frac{1}{2}D^B(y) - \frac{1}{4} \sum_{i=1}^k \{b_i^A(y) - b_i^B(y)\}^2.$$

Explain therefore that also the combo estimator is beating the ML uniformly.

(b) Generalise the above to the weighted combination $\hat{\theta} = \sum_{r=1}^m w_r \hat{\theta}_r$ with weights w_r summing to one: show that its D function can be expressed as

a portfolio may
be diversified

$$D(y) = \sum_{r=1}^m w_j D_j(y) - \sum_{r < s} w_r w_s \sum_{i=1}^k \{b_{r,i}(y) - b_{s,i}(y)\}^2.$$

Any portfolio of clever estimation schemes can hence be combined into a diversified scheme clever enough to beat the ML.

(c) To make the above concrete, suppose $\theta_1, \dots, \theta_k$ are to be estimated, and that there are associated covariate values $x_1, \dots, x_k$ which perhaps but not necessarily have any correlation with the θ_i . Consider estimators

$$\begin{aligned}\hat{\theta}_i^A &= Y_i - \{(k-2)/W_A\}(Y_i - \theta_0), & W_A &= \sum_{i=1}^k (Y_i - \theta_0)^2, \\ \hat{\theta}_i^B &= Y_i - \{(k-3)/W_B\}(Y_i - \bar{Y}), & W_B &= \sum_{i=1}^k (Y_i - \bar{Y})^2, \\ \hat{\theta}_i^C &= Y_i - \{(k-4)/W_C\}(Y_i - (\hat{a} + \hat{b}x_i)), & W_C &= \sum_{i=1}^k (Y_i - \hat{a} - \hat{b}x_i)^2,\end{aligned}$$

with $\hat{a} + \hat{b}x_i$ the estimated regression line (whether well-fitting to the data or not at all). Show that $w_a \hat{\theta}^A + w_b \hat{\theta}^B + w_c \hat{\theta}^C$ is also uniformly better than the ML, for any choices of weights w_a, w_b, w_c summing to one.

Ex. 9.30 Poisson means and inadmissibility of the ML estimator. To show that an estimator is inadmissible it suffices to showcase one estimator that is uniformly better. Let $Y_1, \dots, Y_k$ be independent Poisson with means $\theta_1, \dots, \theta_k$. We are to estimate the $\theta = (\theta_1, \dots, \theta_k)$ with the weighted loss function $L(\theta, \delta) = \sum_{i=1}^k (\delta_i - \theta_i)^2 / \theta_i$, where $\delta = (\delta_1, \dots, \delta_k)$. We demonstrate below that the usual estimator, with Y_i for θ_i , is minimax, but inadmissible; there are competitors doing uniformly better.

(a) Show first that the ML estimator in this problem is $\delta_{\text{ml}} = (Y_1, \dots, Y_k)$, i.e. using Y_i for θ_i . Show that its risk function is the constant k . Via gamma independent priors $\theta_i \sim \text{Gam}(1, b)$, find the Bayes estimator and the Bayes risk, and demonstrate that the ML estimator is minimax.

(b) Let $Z = \sum_{i=1}^k Y_i$ be the sum of these independent Poisson observations, write $\gamma = \sum_{i=1}^k \theta_i$ for the sum of the p Poisson means, and define $\pi_i = \theta_i / \gamma$ for $i = 1, \dots, k$. Show that $(Y_1, \dots, Y_k)$ given $Z = z$ is a multinomial $(z, \pi_1, \dots, \pi_k)$; this is valid for any $z \geq 1$. This establishes that $E(Y_i | Z) = Z\pi_i$ and $\text{Var}(Y_i | Z) = Z\pi_i(1 - \pi_i)$, for $i = 1, \dots, k$, with these formulae holding trivially also for $Z = 0$.

(c) Prove the following little lemma. If $X \sim \text{Pois}(\theta)$ and g is a function such that $g(0) = 0$, then $E_\theta g(X) / \theta = E_\theta g(X+1) / (X+1)$.

(d) Consider the estimator δ^* whose components are

$$\delta_j^* = \{1 - \phi(Z)\}Y_j \quad \text{for } j = 1, \dots, k.$$

The game to be played now (as with the James–Stein estimator worked with in Ex. 9.25) is to find an expression for the risk difference $R(\theta, \delta^*) - R(\theta, \delta_{\text{ml}})$ that is independent of the unknown parameters. Using the results from (b) and (c) show indeed that the risk difference $R(\theta, \delta^*) - R(\theta, \delta_{\text{ml}})$ can be expressed as $E_\gamma D(Z)$, with

$$D(z) = \{\phi(Z+1)^2 - 2\phi(Z+1)\}\{(Z+1) + (k-1)\} + 2\phi(Z)Z.$$

(e) Searching for a good ϕ , aiming for a negative $D(z)$, suppose $\phi(z)z$ is nondecreasing in z , and check what happens for $\phi(z) = c/(k-1+z)$. Show that there indeed is an interval of c succeeding in $D(z) < 0$ for all z , provided $k \geq 2$, with the best value being $c = k-1$. This is the Clevenson–Zidek estimator, uniformly better than the ML estimator:

$$\delta_j^*(y) = \{1 - (k-1)/(k-1+z)\}y_j \quad \text{for } j = 1, \dots, k.$$

(f) We have shown that δ_{CZ} uniformly dominates the ML estimator, however, we do not yet now whether or not there exists an estimator that dominates δ_{CZ} . Show that δ_{CZ} is admissible.

emil: are we asking readers to prove admissibility?, this is hard, see emil-nils paper

Ex. 9.31 *Multiparameter inference, III.* (xx to be calibrated with 9.30; could perhaps be pushed to Ch7. we do Bayes, empirical Bayes, full Bayes. xx) let $Y_i \sim \text{Pois}(\theta_i)$ for $i = 1, \dots, k$, aiming to estimate the ensemble of means, with loss $L = \sum_{i=1}^k (\hat{\theta}_i - \theta_i)^2 / \theta_i$. standard is Y_i , constant risk k . we do $\theta_i \sim \text{Gam}(a, b)$, $\hat{\theta}_i(a, b) = (y_i + a - 1)/(b + 1)$, i think. empirical Bayes yields almost Clevenson–Zidek. then throw in full Bayes, with prior for (a, b) too.

(a) The default estimator would be $\hat{\theta}_j = Y_j$, for $j = 1, \dots, k$. Show that its risk function is constant, equal to k .

(b) Consider then the prior which takes $\theta_1, \dots, \theta_k$ independent from a Gamma (a, b) . Show that θ_j | data is a $\text{Gam}(a_j + y_j, b + 1)$, find the Bayes estimator, and its risk function. (xx jotting down details here. include minimax. we find $\text{MBR}(a, b) = n/(b + 1)$, not depending on a , so Y is minimax. xx)

$$\hat{\theta}_j = \hat{\theta}_j(a, b) = 1/(E(1/\theta_j) | y_j) = \frac{a + y_j - 1}{b + 1}.$$

empirical Bayes: $E(Y_j | \theta_j) = \theta_j$, $\text{Var}(Y_j | \theta_j) = \theta_j$, leading to $E Y_j = a/b$, $\text{Var} Y_j = a/b + a/b^2$; can then use moment matching to find $\tilde{a}, \tilde{b}$. this leads to empirical Bayes estimator $\theta_j^* = \hat{\theta}_j(\tilde{a}, \tilde{b}) = (\tilde{a} + y_j - 1)/(\tilde{b} + 1)$. also:

$$\bar{f}(y_j, a, b) = \int_0^\infty \exp(-\theta) \theta^{y_j} / y_j! \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) d\theta = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + y_j)}{(b + 1)^{a + y_j}} \frac{1}{y_j!}$$

Full Bayes, a prior for (a, b) :

$$\pi_0(a, b) \prod_{j=1}^k g(\theta_j, a, b) f(y_j, \theta_j) = \pi_0(a, b) \prod_{j=1}^k g(\theta_j, a + y_j, b + 1) \bar{f}(y_j, a, b)$$

which leads to $\pi(a, b | \text{data}) \propto \pi_0(a, b) \prod_{j=1}^k \bar{f}(y_j, a, b)$. then mcmc for sampling from $(a, b, \theta_1, \dots, \theta_k)$ given data.

Ex. 9.32 *Multiparameter inference, IV.* Consider a setup with $Y_i \sim \text{unif}[0, \theta_i]$, for n independent experiments, with the ensemble of $\theta = (\theta_1, \dots, \theta_n)$ to be estimated.

(a) For a single $Y_i \sim \text{unif}[0, \theta_i]$, find the risk function $E_{\theta_i} (cY_i - \theta_i)^2$, show that the best value is $c_0 = 1.5$. With joint loss function $L = n^{-1} \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2$, show that risk function for the $1.5 Y_i$ estimator is $R_0(\theta) = \frac{1}{4} a_n$, with $a_n = n^{-1} \sum_{i=1}^n \theta_i^2$. For the below, write also $v_n = n^{-1} \sum_{i=1}^n (\theta_i - \bar{\theta})^2 = a_n - \bar{\theta}^2$.

(b) For a Bayesian and then empirical Bayesian construction, start with the prior $\theta_i \sim \text{Gam}(2, b)$, i.e. with density $b^2 \exp(-b\theta)$, first with b taken known. Show that $\theta_i | y_i$ has the distribution $y + \text{Expo}(b)$, with posterior mean $\hat{\theta}_i = y_i + 1/b$. Characterise the set of favourable $(\theta_1, \dots, \theta_n)$ where the risk function is better than $R_0(\theta)$.

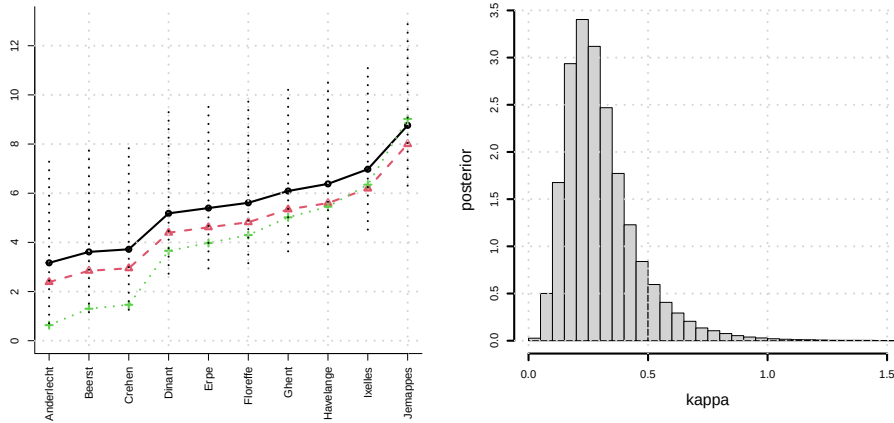


Figure 9.2: Left panel: empirical Bayes (full curve), full Bayes (slanted), direct estimates (dotted), for the ten Belgian laboratories, along with 80 percent credibility intervals displayed vertically. Right panel: posterior distribution for $\kappa/\bar{\theta}$.

(c) For the prior above, show that the marginal distribution of Y_i is the $\text{Expo}(b)$. Use this to motivate the empirical Bayes estimator $\theta_i^* = Y_i + \bar{Y}$. Work out the risk function can be written

$$R_{\text{eb}}(\theta) = n^{-1} \sum_{i=1}^n \left\{ \frac{1}{4} (\theta_i - \bar{\theta})^2 + \frac{\theta_i^2}{12} + \frac{na_n}{n^2} + \frac{2}{12} \theta_i^2/n \right\} = \frac{v_n}{4} + \frac{a_n}{12} + \frac{7}{6} \frac{a_n}{n}.$$

Argue in particular that for n moderate or large, the empirical Bayes will fare better than the direct $1.5 Y_i$, provided the parameters are such that $\kappa_n = n^{-1} \sum_{i=1}^n (\theta_i/\bar{\theta} - 1)^2 \leq 2$.

(d) For a full Bayesian analysis we need a prior for b with the marginal distribution for the Y_i , show that the Jeffreys prior is $\pi_{\text{jeff}}(b) \propto 1/b$. So the setup is now (i) $b \sim 1/b$; (ii) $\theta_i | b \sim \text{Gam}(2, b)$; (iii) $Y_i | \theta_i \sim \text{unif}[0, \theta_i]$. Show that the posterior for b is $\text{Gam}(n, n\bar{y})$.

(e) For an illustration, suppose now that there have been ten different experiments as described above, carried out in ten different Belgian laboratories, associated with unknown parameters θ_i , with the (sorted) data being 0.423, 0.868, 0.974, 2.436, 2.652, 2.865, 3.346, 3.635, 4.231, 6.016. Simulate say 10^5 realisations of $(b, \theta_1, \dots, \theta_n)$, from which you in particular can read off posterior means $\hat{\theta}_{i,B} = E(\theta_i | \text{data})$. Construct a version of Figure 9.2, left panel, with estimates direct, empirical Bayes, Bayes, along with 80 percent credibility intervals, i.e. using 0.10 and 0.90 quantiles for the posteriors. Read off also the distribution of κ_n , as for Figure 9.2, right panel. Why is this a strong indication that empirical Bayes fares far better than the $1.5 Y_i$ estimates?

Hypothesis testing, confidence intervals (again)

Ex. 9.33 *Testing a simple hypothesis.* (xx make connection to Bayes Ch6 and Neyman-Pearson Ch4. xx) Let Y have density $f(y, \theta)$ and consider the simple hypothesis $H_0: \theta = \theta_0$ versus the simple alternative $\theta = \theta_1$. The statistical tests ϕ , with $\phi(x) = 1$ meaning reject H_0 , and $\phi(x) = 0$ keep H_0 , are to be evaluated under the loss function

$$L(\theta_0, \phi) = \begin{cases} 0, & \text{if } \phi(x) = 0, \\ K_1, & \text{if } \phi(x) = 1, \end{cases} \quad L(\theta_1, \phi) = \begin{cases} K_2, & \text{if } \phi(x) = 0, \\ 0, & \text{if } \phi(x) = 1. \end{cases}$$

(a) Let $0 < \pi_0 < 1$ be your prior probability of H_0 being true. Derive an expression for the posterior expected loss, and show that the Bayes solution ϕ_π is of the likelihood ratio type

$$\phi_\pi(x) = \begin{cases} 1, & \text{if } f(y, \theta_1) > k_\pi f(y, \theta_0), \\ 0, & \text{if } f(y, \theta_1) < k_\pi f(y, \theta_0). \end{cases}$$

Find k_π and relate this quantity to the level of a test.

(b) Let now Y have the $N(\theta, 1)$ distribution, for given θ . We want to test $H_0: \theta = 0$ versus $\theta_1 = \frac{1}{2}$ using the Bayes solution when the prior is $\pi_0 = \frac{1}{2}$. Find K_1 and K_2 such that $E_{\theta_0} \phi_\pi(Y) = 0.05$.

(c) Show that any Bayesian test, with a prior giving weight to both the null- and the alternative hypothesis, is the most powerful test of its size. Hence the Bayesian method has somehow rediscovered the Neyman–Pearson Lemma of Ch. 4. *Hint:* Use what you know about Bayes solutions and admissibility.

Ex. 9.34 *Testing2.* (xx one more exercise here, on the Bayesian take on testing. a little bit on risk functions. xx)

Ex. 9.35 *Setting confidence intervals via loss and risk.* Suppose a confidence interval is needed for θ , and that this operation is associated with the loss function

$$l(\theta, a, b) = c(b - a) + I(\theta \notin [a, b]),$$

with c a positive constant balancing length vs. inclusion. Thus nothing is said a priori regarding the coverage level; this will rather be sorted out as a consequence of having selected the c .

(a) With $[\hat{a}, \hat{b}]$ such an interval, show that the risk function is

$$R(\theta, \hat{a}, \hat{b}) = c E_\theta (\hat{b} - \hat{a}) + \Pr_\theta(\theta \leq \hat{a}) + \Pr_\theta(\theta > \hat{b}).$$

(b) Consider the stylised case of $Y \sim N(\theta, 1)$, and assume $c < \phi(0) = 1/(2\pi)^{1/2}$. Find the risk function for the interval $[Y - d, Y + d]$, and show that the best choice is $d_0 = \phi^{-1}(c)$. Here we use $\phi^{-1}(c) = \{-2 \log((2\pi)^{1/2} c)\}^{1/2}$ for the positive solution to $\phi(d) = c$. Verify that we for $c = 0.058$ we have the classic $Y \pm 1.96$. Explain that the risk function is constant, equal to $2\{cd_0 + 1 - \Phi(d_0)\}$.

(c) The Bayesian take on such constructions is to give θ a prior. Assume the posterior density $g(\theta | y)$ is unimodal with a well-defined mode. Show that the expected posterior loss takes the form $c(b - a) + G(a | y) + 1 - G(b | y)$, with G the posterior c.d.f., and that the Bayes solution is to use $[a, b]$, with a and b the two solutions to $g(\theta | y) = c$, one to the left and one to the right of the mode.

(d) Back to the stylised case, use the prior $\theta \sim N(0, \tau^2)$. Verify, e.g. via Ex. 7.19, that $\theta | y \sim N(\rho y, \rho)$, with $\rho = \tau^2/(\tau^2 + 1)$. Show that the Bayes interval becomes $\rho Y_i \pm h$, with $h = \rho^{1/2} \phi^{-1}(\rho^{1/2} c)$. Compute also its associated minimum Bayes risk,

$$\text{MBR}(N(0, \tau^2)) = 2\{ch + 1 - \Phi(h/\rho^{1/2})\},$$

and use this to prove that the $Y \pm d_0$ is minimax.

Estimation in a few stylised prototype problems

Ex. 9.36 *The Problem of the Nile: performance, UMVU.* (xx round off, to be calibrated with Ex. 8.48. 2-dim sufficient in 1-dim model. ancillary. RB of ML, but with little effect. xx) Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. with density $f_\theta(x, y) = \exp(-\theta x - y/\theta)$ for positive (x, y) and θ .

called the
Problem of the
Nile

(a) Show that the sample averages can be represented as $\bar{X} = (1/\theta)A_n$ and $\bar{Y} = \theta B_n$, where A_n and B_n are independent with the distribution of $\chi_{2n}^2/(2n)$. From this show that

$$\hat{\theta} = (\bar{Y}/\bar{X})^{1/2} = \theta(B_n/A_n)^{1/2} = \theta F_{2n, 2n}^{1/2}.$$

Show that this is the ML estimator, a sensible 1-dimensional summary of the 1-parameter θ . Show that it is not sufficient, however. Show also that $(\hat{\theta}_n, V)$ is minimal sufficient, where $V = (\bar{X}\bar{Y})^{1/2}$, and that V is ancillary. (xx point back to Ch4. xx)

(b) Show first that $\bar{Y}/\bar{X}$ is complete. Then use Lehmann–Scheffé to show that any function of $\bar{Y}/\bar{X}$, with finite variance, is an UMVU for its mean. In particular, find the UMVU estimator for θ .

(c) Show that $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \frac{1}{2}\theta^2)$, and that this matches the limit for the Cramér–Rao lower bound. (xx perhaps exercise to be pushed to Ch9, with Rao–Blackwell. $\theta^* = E(\hat{\theta}|V)$ is better, but not by much. xx)

(d) In various examples, the ML estimator is a minimal sufficient statistic (xx spell this out for expofamily above xx). In the setup studied here it is not. One way of measuring the information loss in only knowing $\hat{\theta}$ compared to the full minimal sufficient statistic $(\hat{\theta}, V)$ is by way of the Fisher information. Let $g_{n,\theta}$ be the density of $\hat{\theta}$, and compute the difference $-E\{\partial^2/\partial\theta^2 \sum_{i=1}^n \log f_\theta(X_i, Y_i)\} + E\{\partial^2/\partial\theta^2 \sum_{i=1}^n \log g_{n,\theta}(\hat{\theta})\}$. (xx can also do KL? xx)

Ex. 9.37 *Estimating the squared bias.* Assume $Y \sim N(b, 1)$, and that $\phi = b^2$ is to be estimated. There are links from the present prototype exercise about squared biases to the FIC theory of model selection in Ch. 12.

(a) Go through the relevant motivation for the following three estimators for the squared bias. (i) Since Y estimates b , use $\hat{\phi}_1 = Y^2$. (ii) Explain that Y^2 has expected value $b^2 + 1$. This invites $\hat{\phi}_2 = Y^2 - 1$. Show also that this is the only unbiased estimator for b^2 . (iii) The trouble with the unbiased estimator for a nonnegative quantity is that it may be negative; what is the probability that this happens? Motivate the truncated estimator $\hat{\phi}_3 = \max(Y^2 - 1, 0)$.

(b) For these three estimators, find expressions for their risk functions $r_j(b) = E_b(\hat{\phi}_j - b^2)^2$, and plot them in a diagram. Comment on what you find.

(c) Compute also risk functions numerically under absolute loss, i.e. using $E_b|\hat{\phi}_j - b^2|$, where the better behaviour of $\hat{\phi}_3$ shows up more clearly.

Ex. 9.38 *Estimating a normalised distance, from a narrow to a wide model.* Consider a situation where $Y \sim N(\theta, 1)$ and where it is thought not unreasonable that θ is not far from zero. One may then consider estimating θ via a smooth mixture of the simplest estimator, 0, and the natural wider model estimator, Y . There are links from this stylised exercise to FIC and model averaging themes of Ch. 12, just as for Ex. 9.37.

(a) Explain that the estimator $\hat{\theta}(a) = aY$ has mean squared error $(1-a)^2\theta^2 + a^2$. Show that this is minimised for $a_0 = \rho = \theta^2/(\theta^2 + 1)$, with minimum mse equal to this same ρ .

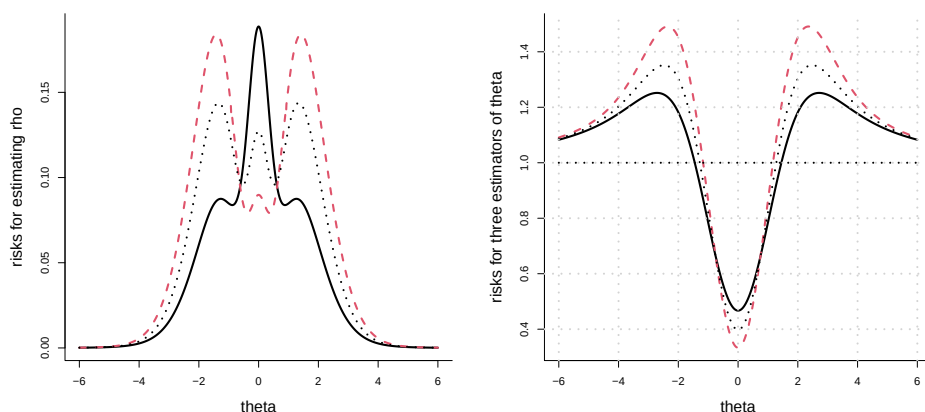


Figure 9.3: Left panel: risk functions for the three estimators of $\rho = \theta^2/(\theta^2 + 1)$ (the first full, the second slanted, the third dotted). Right panel: risk functions for the three consequent estimators $\hat{\theta}_j = \hat{\rho}_j Y$ of θ (the first, the second, the third in increasing order of their maxima).

(b) There is hence a need to estimate ρ well. Since Y estimates θ , a direct estimator is $\hat{\rho}_1 = Y^2/(Y^2 + 1)$. Since Y^2 has mean $1 + \theta^2$, however, argue that $\hat{\rho}_2 = \max(Y^2 - 1, 0)/Y^2$ might also work well. For good measure we include one more, between the two first, namely $\hat{\rho}_3 = \max(Y^2 - c, 0)/(Y^2 + c)$, with $c = \frac{1}{2}$. Compute and display the risk functions for these three estimators, under squared error loss, as with Figure 9.3, left panel.

(c) Yet another estimator to be contemplated is the Bayesian solution, starting with a flat prior for θ , often used in situations with an unknown normal mean. Show that $\theta | y \sim N(y, 1)$, and explain that the resulting estimator for ρ , under squared error, is

$$\hat{\rho}_4 = E(\rho | Y) = \int_0^\infty \frac{x}{x+1} g_1(x, Y^2) dx,$$

the mean of $K/(K+1)$, where $K \sim \chi_1^2(Y^2)$.

(d) Coming back to the intended use of estimating ρ , consider the three estimators of θ in the mean estimation problem,

$$\hat{\theta}_1 = \frac{Y^2}{Y^2 + 1} Y, \quad \hat{\theta}_2 = \frac{\max(Y^2 - 1, 0)}{Y^2} Y, \quad \hat{\theta}_3 = \frac{\max(Y^2 - \frac{1}{2}, 0)}{Y^2 + \frac{1}{2}} Y.$$

The second estimator may be written $(1 - 1/Y^2)Y$ for $|Y| \geq 1$ and zero for $|Y| \leq 1$; similarly the third estimator is $\{(Y^2 - \frac{1}{2})/(Y^2 + \frac{1}{2})\}Y$ for $Y^2 \geq \frac{1}{2}$ and zero for $Y^2 \leq \frac{1}{2}$. Compute and display the risk functions of these three estimators, under squared error loss, as with Figure 9.3, right panel.

(e) We learn in these two situations, estimating $\rho = \theta^2/(\theta^2 + 1)$ and then θ via $\hat{\rho}Y$, that there are no clear winners. Explain that the minimax principle may be applied, leading to $\hat{\rho}_1 Y$ being the best for the second competition. Explain further that to find a clear ranking, one solution is to apply a Bayesian prior, say a suitable $N(0, \kappa^2)$ for θ , here seen as a sensible weight function and not as a prior knowledge translated in to a prior, per se. With the standard normal as such a weight function, seeing values of θ outside $(-3, 3)$ as

relatively unimportant, for example, compute the Bayes risks and rank the estimators. Find also the genuine Bayes solutions, for this prior, when estimating ρ and θ .

Ex. 9.39 *Estimating a bounded mean.* With $Y \sim N(\theta, 1)$, the task is to estimate the mean under quadratic loss, when it is a priori known that $|\theta| \leq 1$.

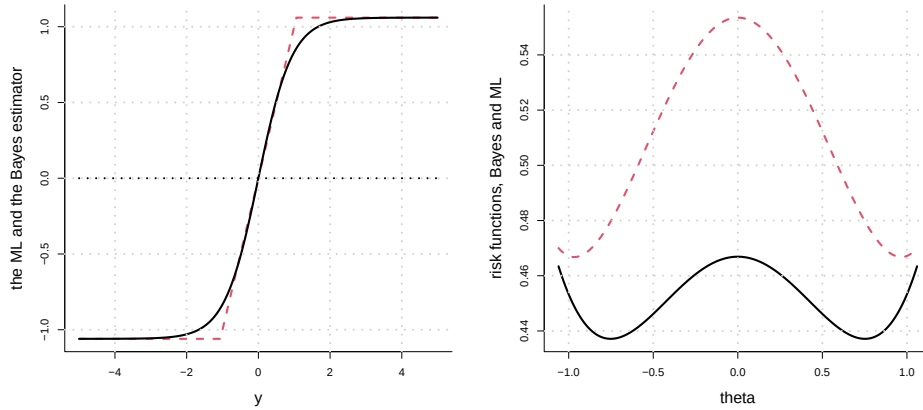


Figure 9.4: Left panel: the Bayes (full curve) and the ML estimators (slanted), as functions of y . Right panel: their risk functions (with full curve for Bayes and slanted for the ML), with minimum Bayes risk 0.4496 for the two-point Bayes prior equal to the risk value at these two points.

(a) Show that the ML estimator is Y truncated to $[-1, 1]$. Compute and find a workable formula for its risk function

$$R(\theta) = \int_{-1}^1 (y - \theta)^2 \phi(y - \theta) dy + \int_{-\infty}^{-1} (-1 - \theta)^2 \phi(y - \theta) dy + \int_1^{\infty} (1 - \theta)^2 \phi(y - \theta) dy.$$

(b) Then examine the Bayesian prior with θ taking values ± 1 with equal probabilities. What is the marginal density $\bar{f}(y)$ for Y ? Show that $\Pr(\theta = 1 | y)$ is $p(y) = \exp(2y)/\{1 + \exp(2y)\}$, and that this leads to the Bayes estimator

$$\hat{\theta}_B = E(\theta | y) = \frac{\exp(2y) - 1}{\exp(2y) + 1} = \frac{\sinh(Y)}{\cosh(Y)}.$$

Plot first the Bayes and ML estimators, as functions of y , as in Figure 9.4, left panel; then compute and display both risk functions, as in the right panel.

(c) Find the Bayes risk, perhaps in two ways: (i) via the risk function $R_B(\theta)$, averaged over the prior; (ii) via $\text{MBR} = E \text{Var}(\theta | Y) = 4 E p(Y) \{1 - p(Y)\}$. Show that $\hat{\theta}_B$ is a minimax estimator.

(d) (xx generalisations. could bother to do this. The computations above may be mimicked for the case where $\theta \in [-m, m]$ a priori, but we learn in the process that the minimax argument successfully used for $m = 1$ does not work, in cases where $m > m_0$, where $m_0 \approx 1.06$ (and indeed the Bayes estimator is then not minimax). Start out taking the θ prior to be symmetrically concentrated in $\{-m, m\}$. Show that the posterior probability of $\theta = m$ is equal to $p(y) = \exp(2my)/\{1 + \exp(2my)\}$, and that the posterior mean is $E(\theta | y) = m \tanh(my)$. Show that the minimum Bayes risk is equal to $E 4m^2 p(Y) \{1 - p(Y)\}$ and compute this, for values of m equal to say 0.90, 1.00, 1.10.

The point is now that this minimum Bayes risk is equal to the maximum risk for $m = 0.90, 1.00$, but not for $m = 1.10$. For values $m > m_0$, the risk function starts having three modes, at $-m, 0, m$, with the one at zero becoming bigger than at the endpoints.

Ex. 9.40 *Estimating a binormal mean with nonnegativity constraints.* (xx to be done and polished. xx) Let (X, Y) be binormal with mean (a, b) , unit variances, and known correlation ρ . The task is to estimate (a, b) , with the natural quadratic loss function

$$\begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix}^t \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix} = \frac{1}{1 - \rho^2} \{(\hat{a} - a)^2 + (\hat{b} - b)^2 - 2\rho(\hat{a} - a)(\hat{b} - b)\},$$

when it is known a priori that $a \geq 0, b \geq 0$. We omit the $1/(1 - \rho^2)$ constant in the loss and risk function questions below.

(a) A simple direct method is to truncate X and Y to zero, if negative, i.e. using $\tilde{a} = \max(X, 0)$ and $\tilde{b} = \max(Y, 0)$. Find formulae for $E_a \tilde{a}$, $\text{Var}_a \tilde{a}$, with of course similar results for $\tilde{b}$. Find the risk function

$$R_0(a, b) = E_{a,b} \{(\tilde{a} - a)^2 + (\tilde{b} - b)^2 - 2\rho(\tilde{a} - a)(\tilde{b} - b)\}.$$

(b) Then consider the ML estimators $(\hat{a}, \hat{b})$, maximising the log-likelihood under the given constraints. Show that this, for a given observed (x, y) , is equivalent to minimising

$$Q(a, b) = (a - x)^2 + (b - y)^2 - 2\rho(a - x)(b - y) \quad \text{with } a \geq 0, b \geq 0.$$

(c) Though $(\hat{a}, \hat{b})$ may be found via numerical routines, e.g. the ‘nlm in box’ algorithm `nlminb` in R, it is helpful to find explicit formulae for the ML. Going through the four situations, for the four quadrants into which (x, y) might fall, show this: (i) For $x > 0, y > 0$, we have $(\hat{a}, \hat{b}) = (x, y)$; (ii) for $x > 0, y < 0$, we have $\hat{b} = 0$ and $\hat{a} = \max(x - \rho y, 0)$; (iii) for $x < 0, y > 0$, we have $\hat{a} = 0$ and $\hat{b} = \max(y - \rho x, 0)$; (iv) for $x < 0, y < 0$, we have $\hat{a} = \max(x - \rho y, 0)$ and $\hat{b} = \max(y - \rho x, 0)$.

(d) There is now a well-defined risk function for the ML method,

$$R(a, b) = E_{a,b} \{(\hat{a} - a)^2 + (\hat{b} - b)^2 - 2\rho(\hat{a} - a)(\hat{b} - b)\},$$

though with no easy closed-form formula. Show that for $a \geq 2.50, b \geq 2.50$, say, the chance is so high for (X, Y) to land in the positive quadrant that both R_0 and R come close to $2(1 - \rho^2)$.

(e) Carry out some numerical computation and comparison of the two risk functions, via simulation of binormal (X, Y) , for given value of ρ , at grid positions (a, b) . (xx more. ML is often better, sometimes significantly so, but not uniformly. application to regression situation with constraints on coefficients. xx)

Ex. 9.41 *When a is smaller than b .* Consider independent X and Y being $N(a, 1)$ and $N(b, 1)$, with (a, b) to be estimated with loss function $L = (\hat{a} - a)^2 + (\hat{b} - b)^2$. The usual estimator for is (X, Y) , with risk 2. Suppose however that it is known a priori that $a \leq b$.

(a) Show that the ML $(\hat{a}, \hat{b})$ is

$$(X, Y) \text{ if } Y \geq X, \quad \left(\frac{1}{2}(X + Y), \frac{1}{2}(X + Y)\right) \text{ if } Y < X.$$

(b) To find workable expressions for the loss, and then its risk, introduce $U = (X + Y)/\sqrt{2}$ and $V = (Y - X)/\sqrt{2}$; explain that these are independent unit variance normals,

this is isotonic
repair; see
Ex. 8.6

with means $\alpha = (a + b)/\sqrt{2}$ and $\delta = (b - a)/\sqrt{2}$. Now do the algebra to show that

$$\begin{aligned} L &= \begin{cases} U^2 + V^2 + a^2 + b^2 - \sqrt{2}(a + b)U - \sqrt{2}(a - b)V & \text{if } V \geq 0 \text{ (no pooling),} \\ U^2 + a^2 + b^2 + \sqrt{2}(a + b)U & \text{if } V < 0 \text{ (pooling).} \end{cases} \\ &= U^2 + a^2 + b^2 - \sqrt{2}(a + b)U + \begin{cases} V^2 - 2\delta V & \text{if } V \geq 0 \text{ (no pooling),} \\ 0 & \text{if } V < 0 \text{ (pooling).} \end{cases} \end{aligned}$$

Take the means, including the shift from v to $w - \delta$ with $\int_{-\delta}^{\infty} (w^2 - \delta^2)\phi(w) dw$, to reach

$$R(a, b) = 1 + \delta^2 - (\delta^2 - 1)\Phi(\delta) - \delta\phi(\delta).$$

Plot this risk, and show in particular that it dominates the 2 risk for the plain (X, Y) . For $\delta = 0$ we have $3/2$, i.e. a 25 percent reduction; for δ big we're close to 2 again.

(c) Then be Bayesian, using a flat prior on $a \leq b$. Derive the posterior, and find that its mode is the ML.

Estimation of parametric densities

Ex. 9.42 *Unbiased estimation of a parametric density.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from a parametric density $f(y, \theta)$, like the normal or the Gamma or the Beta. How can we construct an unbiased estimator of the density function itself? For this exercise, assume there is a sufficient and complete statistic, say $T = (T_1, \dots, T_p)$.

(a) Let $\hat{\theta}$ be the ML estimator. Explain why it needs to be a function of $T_1, \dots, T_p$, and that the ML estimator of the density itself is $f(y, \hat{\theta})$.

(b) For a given window A (make it an interval, if the distribution is one-dimensional), a very simple estimator for $p(\theta) = \Pr(Y \in A) = \int_A f(y, \theta) dy$ is $\hat{p} = I(Y_1 \in A)$, using a single data point. Show that it is unbiased. This also invites the somewhat more sensible estimator $\bar{p} = n^{-1} \sum_{i=1}^n I(Y_i \in A)$, the binomial proportion of data points inside the A window. Show that it is unbiased and find a formula for its variance.

(c) Typically this estimator can be beaten, however. Consider indeed

$$p^* = E(\hat{p} | T) = \Pr(Y_1 \in A | T).$$

Explain why this is actually an estimator, i.e. that it does not depend on the parameter θ , that it is unbiased, and indeed the UMVU estimator for $p(\theta)$. Show furthermore that the construction $E(\bar{p} | T)$ leads to the very same p^* . In particular, there is a well-defined UMVU estimator for the c.d.f. $F(y, \theta)$, with this recipe being valid for any parametric distribution (again, provided there is a sufficient and complete statistic).

(d) Assume the situation is regular enough for the density $f_n(y | T)$ of a Y_i given T to exist. Explain why it does not depend on the parameter, and that

$$p^* = \int_A f_n(y | T) dy, \quad \text{for all windows } A.$$

Show that $f_n(y | T)$ is unbiased, and the UMVU for $f(y, \theta)$.

(e) Let $Y_1, \dots, Y_n$ be i.i.d. from the exponential density $\theta \exp(-\theta y)$. Show that $T = \sum_{i=1}^n Y_i$ is complete and sufficient, and that $(Y_1/T, \dots, Y_n/T)$ has a Dirichlet $(1, \dots, 1)$ distribution, with each Y_i/T having a Beta distribution. Show from this that the UMVU estimator for the density takes the form

$$f_n(y | T) = \text{be}(y/T, 1, n-1)/T = (n-1)(1-y/T)^{n-2} \quad \text{for } y \in [0, T].$$

Why does the fact that the estimator is zero, outside the $[0, T]$ window, not contradict its unbiasedness? Show also that $f_n(y|T)$ comes close to the ML estimator $f(y, \hat{\theta})$, for growing n . Give an example where the ML estimator is unbiased, but where $f(y, \hat{\theta})$ is not unbiased for $f(y, \theta)$.

(f) Above we managed to find a formula $f_n(y|T)$ for the unbiased estimator of the exponential density, via the fact that Y_1/T has a Beta distribution $(1, n-1)$. Suppose, for the sake of the argument, that you did not quite have the time to do the last part of the mathematics, deriving that formula. Explain and show via an illustration in your computer that the following would work. For the given dataset (perhaps with a small n , to see more interesting shapes), compute T , and draw a million realisations of Y'_1 given that values of T . Then form a very fine histogram of these Y'_1 . That fine histogram is a numerical approximation to the $f_n(y|T)$ you need. This argument is valid for any such parametric density estimation situation; as long as you can simulate Y'_1 given the sufficient and complete statistic, you have your unbiased density estimator.

(g) (xx may make a separate short exercise for the discrete case, umvu estimation of $f(y, \theta)$ in two cases. xx) (i) Suppose $Y_1, \dots, Y_n$ are i.i.d. $\text{Pois}(\theta)$. With $T = \sum_{i=1}^n Y_i$, show that $Y_1 | T \sim \text{binom}(T, 1/n)$, leading to the UMVU estimator $f_n(y|T) = \binom{T}{y} (1/n)^y (1 - 1/n)^{T-y}$. Show that this is close to $\exp(-\bar{Y})(\bar{Y})^y/y!$ for n growing. (ii) Consider $Y_1, \dots, Y_a$ being geometric with $(1-p)^y p = q^y p$ for $y = 0, 1, \dots$. Show that $Z = \sum_{i=1}^a Y_i$ is sufficient and complete. Use formulae from Ex. 1.28 to work out that the UMVU estimator of the point probability $q^y p$ is

$$f_a(y|Z) = \binom{Z-y+a-2}{a-2} / \binom{Z+a-1}{a-1} = \frac{(Z-y+a-2)!}{(Z-y)!(a-2)!} \frac{Z!(a-1)!}{(Z+a-1)!}$$

for $y = 0, 1, \dots, Z$. (xx note: not hypergeometric?; close to $(1-\hat{p})^y \hat{p}$, with $\hat{p} = 1/(1+\bar{y})$. xx)

Ex. 9.43 UMVU estimation of the uniform density and cumulative. In Ex. 9.42 we learned how to construct UMVU estimators for the cumulative and densities in parametric families, as long as there is a sufficient and complete statistic. This works also for the uniform case, but with certain extra components and nuances. The uniform density on $[0, \theta]$ is $1/\theta$ for $y \leq \theta$ and zero for $y > \theta$, and there is a subtle difference between estimating this density and estimating the parameter $1/\theta$.

(a) With $Y_1, \dots, Y_n$ i.i.d. from this uniform on $[0, \theta]$, show that $M_n = \max_{i \leq n} Y_i$ is sufficient and complete. Find the UMVU estimator for θ and for $1/\theta$.

(b) Show that the distribution of Y_1 given M_n is a mixture, equal to M_n with probability $1/n$ and uniform on $[0, M_n]$ with probability $1 - 1/n$. Explain that the c.d.f. is $F_n(y|M_n) = (1 - 1/n)(y/M_n)$ for $y < M_n$ and then 1 for $y \geq M_n$. Why is this the UMVU for the parametric c.d.f. $F(y, \theta)$?

(c) Then set up a clear formula for the UMVU estimator of the density, and verify that it is indeed unbiased.

(d) Suppose i.i.d. pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ are observed from the distribution uniform on the interior of the circle $\{(x, y): x^2 + y^2 \leq \theta^2\}$. Writing $R_i = \|(X_i, Y_i)\|$, show that the ML estimator is $\hat{\theta} = M_n = \max_{i \leq n} R_i$, find its distribution, and show that it is sufficient and complete. Then find the distribution of $(X_1, Y_1) | M_n$, and from this the UMVU estimators of the density and cumulative of R_i .

Ex. 9.44 *UMVU estimation of normal densities.* Suppose $Y_1, \dots, Y_n$ are i.i.d. $N(\xi, \sigma^2)$. How can we estimate the underlying density $f(y) = \sigma^{-1} \phi(\sigma^{-1}(y - \xi))$ unbiasedly? For this we use the general results of 9.42. (xx the following to be polished. xx)

(a) For a simpler separate question first, suppose $Y_1, \dots, Y_n$ are i.i.d. $N(\xi, 1)$. Here the sample average $\bar{Y}$ is sufficient and and complete, so the UMVU estimator is $f_n(y)$, the density of $Y_1 | \bar{Y}$. Writing $Y_1 = \bar{Y} + (Y_1 - \bar{Y})$, show that $Y_1 - \bar{Y}$ is independent of $\bar{Y}$. Explain that $f_n(y) = g_n(y - \bar{y})$, with g_n the density of $Y_1 - \bar{Y}$. Show next that g_n is zero-mean normal with variance $1 - 1/n$, leading to

$$f_n(y) = \left(\frac{n}{n-1}\right)^{1/2} \phi\left(\left(\frac{n}{n-1}\right)^{1/2}(y - \bar{y})\right).$$

It is of course close to the simpler ML estimator based density $\phi(y - \bar{y})$.

(b) Then consider a dataset Y_i from $N(\xi, \sigma^2)$, with sufficient and complete estimators $\bar{Y}$ and $\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)$. Write $Y_i = \bar{Y} + \hat{\sigma} W_i$, with $W_i = (Y_i - \bar{Y}) / \hat{\sigma}$. Show via the Basu lemma of Ex. 4.26 that W_i is independent of $(\bar{Y}, \hat{\sigma})$. Show from this that the UMVU estimator must be $f_n(y) = \hat{\sigma}^{-1} h_n(\hat{\sigma}^{-1}(y - \bar{y}))$, with $h_n(z)$ the density of W_1 . We find a formula for this density below. Explain nevertheless how that the numerical approximation trick via simulations, from Ex. 9.42(f), can be made to work here, without having the actual density formula.

(c) As an introductory exercise, before returning to the residuals, suppose $U_1, \dots, U_m$ are i.i.d. $N(0, 1)$, and consider $V_i = U_i / (U_1^2 + \dots + U_m^2)^{1/2}$. Working first with their squares, show that V_i^2 can be written $\chi_1^2 / (\chi_1^2 + \chi_{m-1}^2)$, which is a Beta; more broadly, show that $(V_1^2, \dots, V_m^2) \sim \text{Dir}(\frac{1}{2}, \dots, \frac{1}{2})$. Show next that the density of V_1 becomes

$$g_m(v) = \frac{\Gamma(\frac{1}{2}m)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}(m-1))} (1-v^2)^{(m-1)/2-1} \quad \text{on } (-1, 1).$$

Also relate what you find here to the following: with $U_1, \dots, U_m$ i.i.d. $N(0, 1)$, the distribution of $(U_1, \dots, U_m)$ given its length $\|U\| = r$ is a uniform on the sphere of radius r .

(d) Write $G = \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $V_i = (Y_i - \bar{Y}) / G^{1/2}$. From $\sum_{i=1}^n V_i^2 = 1$ note that $\sum_{i=1}^n W_i^2 \leq m$, with $m = n-1$ the degrees of freedom for $G \sim \chi_m^2$. Show that the distribution of V_i does not depend on the parameters, so for simplicity set (ξ, σ) to $(0, 1)$. With row vectors $a_1 = (1/\sqrt{n}, \dots, 1/\sqrt{n})$ and $a_2 = \{n/(n-1)\}^{1/2}(1 - 1/n, -1/n, \dots, -1/n)$, show that these are orthogonal and of length 1. Hence there is an orthonormal matrix A , with a_1, a_2 the two first rows. Transforming to $Z = AX$, with $Z_1, \dots, Z_n$ i.i.d. $N(0, 1)$ by Ex. 1.49, show that $Z_1 = \sqrt{n}\bar{Y}$ and $Z_2 = \{n/(n-1)\}^{1/2}(Y_1 - \bar{Y})$, and that this leads to

$$V_1 = \{(n-1)/n\}^{1/2} Z_2 / (Z_2^2 + \dots + Z_n^2)^{1/2}.$$

Derive a formula for V_1 based on this, using results from the previous point. Conclude that the UMVU estimator for the normal density is

$$f_n(y) = \frac{1}{\hat{\sigma}} \left(\frac{n}{n-1}\right)^{1/2} \frac{\Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}(n-2))} \left\{1 - \frac{n}{n-1} \left(\frac{y - \bar{y}}{\hat{\sigma}}\right)^2\right\}^{(n-2)/2-1}$$

(e) (xx can attempt multinormal density, would be nice. xx) $Y_i = \bar{Y} + \hat{\Sigma}^{1/2} W_i$, with $W_i = \hat{\Sigma}^{-1/2}(Y_i - \bar{Y})$. independent of the sufficient statistics, has a density $h_n(w)$. which implies

$$f_n(y) = |\hat{\Sigma}|^{-1/2} h_n(\hat{\Sigma}^{-1/2}(y - \bar{Y})).$$

would like to see this more clearly: let $U_1, \dots, U_m$ be i.i.d. $N_p(0, I)$, with $A_m = \sum_{i=1}^m U_i U_i^t$. Then we should understand the distribution of $V_i = U_i^t A_m^{-1} U_i$, with $\sum_{i=1}^m V_i = p$. the distribution of $A_m^{-1} U_1 U_1^t$ is a multivariate Beta type, related to Roy's largest root test, with 1 and $m - 1$ degrees of freedom, but careful with notation and terms. also keen on the root of V_i , i.e. $W_i = A_m^{-1/2} U_i$, with $W_i^t W_i = V_i$. it should have a density with a formula.

Leftovers and further material

Ex. 9.45 *Could your prior be wrong?* (xx ponder this. could be in Ch7, but here we can properly deal with the risk functions. xx) The following prototype setup illustrates also general issues, related to Bayesian methods when the priors used might not be adequate, Consider the classic simple normal-normal Bayesian setup, where $Y | \theta \sim N(\theta, 1)$ and $\theta \sim N(0, \sigma^2)$, where you ostensibly have carefully elicited a value of σ , based on the relevant context.

(a) From Ex. 7.19, make clear that the posterior is $N(\rho y, \rho)$ with $\rho = \sigma^2 / (1 + \sigma^2)$. For squared error loss, verify also that the risk function for the posterior mean estimator $\hat{\theta}_B = \rho Y$ is $R(\hat{\theta}_B, \theta) = \rho^2 + (1 - \rho)^2 \theta^2$. Sketch this loss function, compared with $R(Y, \theta) = 1$.

(b) You, being the Bayesian here, are indirectly betting that the prior is adequate. Show that, marginally, $Y \sim N(0, 1 + \sigma^2)$, so cases with $|y_{\text{obs}}| \geq c_0 = (1 + \sigma^2)^{1/2} z_0$, with e.g. $z_0 = 2.576$, the 0.995 normal quantile, should make you nervous. Another relevant perspective of risk is that you're hoping that the risk is smaller than the neutral frequentist value, i.e. 1. Verify that this corresponds to

$$\rho^2 + (1 - \rho)^2 \theta^2 \leq 1, \quad \text{i.e. } \theta^2 \leq (1 - \rho^2) / (1 - \rho)^2 = (1 + \rho) / (1 - \rho).$$

(c) Having observed y , what is the probability that things are not so fine with your prior, i.e. that $R(\hat{\theta}_B, \theta) > 1$? For the posterior, write $\theta | y \sim \rho y + \rho^{1/2} N = \rho^{1/2} (N + \rho^{1/2} y)$, with N a standard normal. Explain that we have $\theta^2 | y \sim \rho \chi_1^2(\rho y^2)$. From this, show

$$p(\sigma, y) = \Pr(\rho \chi_1^2(\rho y^2) \geq (1 + \rho) / (1 - \rho)) = 1 - \Gamma_1((1/\rho)(1 + \rho) / (1 - \rho), \rho y^2).$$

For $\sigma = 1$, with $\rho = \frac{1}{2}$, plot this function, to learn how big $|y|$ must be in order for you to accept that the chance is 90 percent that your risk function is in dire straits, i.e. associated with losing to the ML estimator Y . The answer takes the form $|y| \geq c = 4.813$.

(d) Taking this into account, work with the following empirical Bayes type approach. If $|y| \leq c$, keep ρy ; if $|y| > c$, revert to y . Compute the risk function for this modified estimator $\hat{\theta}^*$, numerically. Compare with the risk above for $\hat{\theta}_B$, and comment. Include in your risk comparisons also the similar strategy that takes the value c_0 above, instead of the c here.

(e) For a similar analysis, to learn more concerning 'doubting your own prior after having seen the data', consider the simple setup with $Y | p \sim \text{binom}(n, p)$, with n known. With the prior $p \sim \text{Beta}(3, 3)$, compute the risk function $R(\hat{p}_B, p)$ for the posterior mean estimator $\hat{p}_B = (Y + 3) / (n + 6)$. When is this risk bigger than the frequentist default risk $p(1 - p) / n$? For which values of y will the probability of bad news, i.e. p is outside the interval with Bayes being better than frequentist, be bigger than 0.90? Choose e.g. values 10, 20, 30 for n for your illustrations.

Ex. 9.46 *Not believing in zeros.* (xx this one leaves the room, i think, in view of Ex. 9.15. xx) Let $Y \sim \text{Pois}(\theta)$, with the parameter to be estimated using weighted loss $(\delta - \theta)^2/\theta$. The ML estimator Y itself has constant risk 1.

(a) Consider the estimator that doesn't believe in zeros, truncating such to 1, i.e. $\hat{\theta} = y$ if $y \geq 1$ and 1 if $y = 0$. Work with its risk function

$$\begin{aligned} R(\theta, \hat{\theta}) &= \frac{1}{\theta} \left[\sum_{y=1}^{\infty} (y - \theta)^2 f(y, \theta) + (1 - \theta)^2 \exp(-\theta) \right] \\ &= (1/\theta) [\theta - \theta^2 \exp(-\theta) + (1 - \theta)^2 \exp(-\theta)] \\ &= 1 + (1/\theta)(1 - 2\theta) \exp(-\theta). \end{aligned}$$

plot it, and show in particular that there is clear improvement on Y as long as $\theta \geq \frac{1}{2}$. If one is willing to bet on that, a priori, then the truncation gives a clear risk improvement.

(b) try the prior $\pi(\theta) \propto \exp(-b\theta)$ for $\theta \geq c$. the posterior is

$$\pi(\theta | y) \propto \theta^y \exp(-(b+1)\theta) \quad \text{for } \theta \geq c,$$

we need the Bayes solution:

$$1/E \{ (1/\theta) | y \} = \frac{\int_c^{\infty} \theta^y \exp(-(b+1)\theta) d\theta}{\int_c^{\infty} \theta^{y-1} \exp(-(b+1)\theta) d\theta}.$$

try the Blyth method, with $b \rightarrow 0$. would be fun to show admissibility.

Notes and pointers

[xx some notes and pointers here xx]

Stein: See [Stein \(1956\)](#); [James and Stein \(1961\)](#) for the original articles, and, for example [Efron and Morris \(1977\)](#) and [Stigler \(1990\)](#) for lucid presentations. [Clevenson and Zidek \(1975\)](#), [Stoltenberg and Hjort \(2020\)](#), more. point to Hudson1978, generalising Stein multiparameter things to exponential families.

(xx For estimating an ensemble of k normal means, Stein found his surprising inadmissibility result for the ML, for $k \geq 3$; see [Stein \(1956\)](#); [James and Stein \(1961\)](#). When asked how he invented this, he explained that he first proved admissibility for $k = 2$ and then 'for a long time' attempted to prove this also for $k = 3$, before realising that the answer was 'no'. for Ex. 9.28: point to [Lindley and Smith \(1972\)](#). xx)

Stein, personal communication, 1985

(xx Even Francis Galton could shrink. make an exercise based on insights from [Stigler \(1990\)](#), perhaps. xx)

(xx Dan Roy Annals. Lehmann, Blyth. Bhattacharyya bounds, generalising Cramer-Rao. BNP chapter for minimax for $\theta = E_F Y_i$, and admissibility. minimax for F . xx)

(xx multiparameter with eB and more : point to [Efron \(2012\)](#). xx)

(xx there is a literature on estimating bounded means. bickel, casella, more. xx)

Converging processes, Brownian motion, and bridges

Time has come for us to consider stochastic processes, as opposed to random variables and vectors. This necessitates formalising certain function spaces and the key aspects of the infinite-dimensional probability distributions over these. The spaces work with here are $C[0, 1]$, the set of all continuous functions $[0, 1] \rightarrow \mathbb{R}$, and the bigger $D[0, 1]$, of all right-continuous functions with left-hand limits. We study partial-sum processes and their limits in distribution, and reach the Donsker theorem, far more general than the CLT. Some partial-sum processes have non-continuous limits, like Gamma and Poisson processes. Key aspects involved in these undertakings are finite-dimensional convergence and tightness, for which criteria are found. Also studied are these process limits themselves, including Brownian motion, Brownian bridges, Ornstein–Uhlenbeck processes. Applications of high relevance for probability theory and statistics abound, from Kolmogorov–Smirnov type testing to using running t tests when hunting for change-points.

Key words: Brownian bridges, Brownian motion, Donsker theorem, empirical processes, function spaces, Kolmogorov–Smirnov, likelihood-max bridges, monitoring processes, Ornstein–Uhlenbeck process, running t tests, testing constancy, testing models, tightness

A key result, motivating the process view, tended to and applied several times in this chapter, is as follows. For $Y_1, Y_2, \dots$ being i.i.d., with mean ξ and variance σ^2 , form the partial sum process $X_n(t) = (1/\sqrt{n}) \sum_{i \leq [nt]} (Y_i - \xi)$, for $t \in [0, 1]$. The process starts at zero and has the value $\sqrt{n}(\bar{Y}_n - \xi)$ at $t = 1$, which we know tends to $N(0, \sigma^2)$ by the CLT of Ch. 2. Then *Donsker’s theorem* says that the full empirical process X_n tends in distribution to scaled Brownian motion $X = \sigma W$, in a well-defined space of functions $[0, 1] \rightarrow \mathbb{R}$. The $\{W(t) : t \in [0, 1]\}$ process in question has the basic property that increments over disjoint intervals are independent and normal, with $W(t) - W(s) \sim N(0, t - s)$.

Establishing such a $X_n \rightarrow_d \sigma W$ result, and many others, takes some work, from clear definitions and initial preparations to lemmas, criteria for securing other conditions, and theorems. Processes considered live in function spaces, as opposed to Euclidean vector spaces. Below we study $C[0, 1]$, the space of all continuous functions on $[0, 1]$, and then the bigger space $D[0, 1]$ of all $x : [0, 1] \rightarrow \mathbb{R}$ which are continuous to the right and with left-hand limits, i.e. $x(t-) = \lim_{\varepsilon \rightarrow 0} x(t - \varepsilon)$. The natural metric for the $C[0, 1]$ space is the uniform one, $d(x, y) = \max_t |x(t) - y(t)|$, and the more complicated Skorokhod metric $d^*(x, y)$ for $D[0, 1]$ is worked with in Ex. 10.19. Having equipped the function

the uniform
metric

spaces with metrics, the general convergence in distribution definition is now that

$$X_n \rightarrow_d X \quad \text{if } \mathbb{E}g(X_n) \rightarrow \mathbb{E}g(x) \quad \text{for all continuous bounded } g. \quad (10.1)$$

This is one of several equivalent criteria for convergence in distribution in finite-dimensional spaces, as we have seen in the Portmanteau theorem, see Ex. 2.18. It looks rather abstract, but in exercises below we work out different more practical criteria for process convergence.

With F_n the empirical c.d.f. associated with an i.i.d. sample $Y_1, \dots, Y_n$ from some distribution F , we shall work with the monitoring process $Z_n^0 = \sqrt{n}(F_n - F)$, along with various variations. To check whether data come from some given parametric family, with c.d.f. $F(t, \theta)$, we construct the more general $Z_n(t) = \sqrt{n}\{F_n(t) - F(t, \hat{\theta})\}$. Limit processes are relatives of certain Brownian bridges. We also build processes related to say $\hat{\mu}_R(t) - \hat{\mu}_L(t)$, with footscripts indicating data to the right and the left of datapoint $i = [nt]$. The limit processes may be used to check constancy of distributions over time, or to identify changepoints. We also study processes with independent increments, and use the apparatus to build e.g. Poisson and Gamma processes.

Building the appropriate machinery for $C[0, 1]$ and $D[0, 1]$ takes some heavy lifting, which in parts perhaps looks forbiddingly abstract. We are tempted to promise, though, that once the machinery is in place, various applications of the general theory are both widely and impressively useful, and do not require further deep mathematical work.

(xx then two sentences about technical exercises placed in an appendix, if we decide to do this. xx)

The space $C[0, 1]$, Brownian motion, Brownian bridge, Gaussian processes

Ex. 10.1 *The function space $C[0, 1]$.* Many of the more important stochastic processes live in the space $C[0, 1]$, the set of all continuous functions $x: [0, 1] \rightarrow \mathbb{R}$. There are also various situations where a random process $X_n = \{X_n(t): t \in [0, 1]\}$ may have perhaps small discontinuities, but where there is a clear limit process $X = \{X(t): t \in [0, 1]\}$ living in $C[0, 1]$ with probability one. We need a machinery for studying and handling the elements of $C[0, 1]$. the $C[0, 1]$ space of functions

(a) Show that each $x \in C[0, 1]$ is bounded and uniformly continuous. Define *the uniform metric* as the distance

$$d(x, y) = \sup_{t \in [0, 1]} |x(t) - y(t)|.$$

Show that d is a metric on $C[0, 1]$. From now on, the space $C[0, 1]$ refers to the metric space $(C[0, 1], d)$.

(b) Demonstrate that if $(x_n)_{n \geq 1}$ is a Cauchy sequence in $C[0, 1]$, i.e., for any $\varepsilon > 0$ there is an n_0 such that $d(x_n, x_m) < \varepsilon$ for all $n, m \geq n_0$, then $x_n \rightarrow x$ for a function x that is also an element of $C[0, 1]$. This shows that $C[0, 1]$ is *complete*, that is, all Cauchy sequences converge. $C[0, 1]$ is complete

(c) Go through the mechanics of the following simple example, where the point is that x_n tends to some x_0 pointwise, but not uniformly. Let x_0 be the zero function, and let $x_n(t)$ be zero on $[2/n, 1]$, and with a ‘tent’, from 0 to 1 on $[0, 1/n]$ and then from 1 to 0 on $[1/n, 2/n]$. Find $d(x_n, x_0)$. Make some useful variations.

(d) Plot your favourite function du jour $x_0: [0, 1] \rightarrow \mathbb{R}$, along with its 0.03-neighbourhood, $\{x: d(x_0, x) \leq 0.03\}$.

$C[0, 1]$ is
separable

(e) One sometimes works with other intervals than $[0, 1]$, and the basic definitions and properties go through also for say $C[a, b]$, with $0 < a < b$. Show that the function $x_n(t) = n(t^{1/n} - 1)$ converges to the function $x(t) = \log t$ in the space $C[0.01, 1]$. Show also that the function $x_n(t) = t^n$ converges to the zero function in the space $C[0, 0.99]$.

(f) An important property of $C[0, 1]$ is that it is a separable metric space, meaning that it contains a countable and dense subset of functions. For one such construction, for $k \geq 1$ let D_k be the family of piecewise linear functions $x_k(t)$ for $t \in [0, 1]$ where

$$x_k(t) = q_{j-1} + (q_j - q_{j-1}) \frac{t - (j-1)/k}{1/k}, \quad t \in [(j-1)/k, j/k], \quad j = 1, \dots, k,$$

and $q_0, q_1, \dots, q_k$ are rational. Show that $D_k \subset C[0, 1]$ for each $k \geq 1$; that $\cup_{k \geq 1} D_k$ is countable; and that for any $x \in C[0, 1]$ and any $\varepsilon > 0$ we can find $y \in \cup_{k \geq 1} D_k$ such that $d(x, y) < \varepsilon$, which shows that $\cup_{k \geq 1} D_k$ is dense in $C[0, 1]$.

Ex. 10.2 *Continuous functions on $C[0, 1]$.* As mentioned in the introduction, $X_n \rightarrow_d X$ for processes X_n and X in $C[0, 1]$ is formally defined via (10.1). We shall indeed derive more practical criteria in exercises to come, but we may already put this somewhat abstract definition to use.

(a) In Ex. 2.20, we proved the continuous mapping theorem for random vectors taking values in $\mathbb{R}^k$, and continuous functions $\mathbb{R}^k \rightarrow \mathbb{R}^m$. Verify that the proof of that theorem goes through for random processes X_n and X with values in $C[0, 1]$ and continuous functions $h: C[0, 1] \rightarrow \mathbb{R}^m$. That is, show that if $X_n \rightarrow_d X$, then $h(X_n) \rightarrow_d h(X)$, for each continuous function h on $C[0, 1]$. Show also that it is sufficient that h is continuous on a set A where $\Pr(X \in A) = 1$.

(b) For the functions $h_1(x) = \max_t x(t)$, $h_2(x) = \min_t x(t)$, $h_3(x) = \max_t |x(t)|$, $h_4(x) = \int_0^1 x(t) dt$, show that they are continuous.

(c) To anticipate a typical use of such results, suppose $X_n \rightarrow_d W$, with W being the Brownian motion process introduced in Ex. 10.3. Show that the range $R_n = \max_t X_n(t) - \min_t X_n(t)$ tends in distribution to the range $R = \max_t W(t) - \min_t W(t)$ of the Brownian motion. It is also to be noted that if somewhat different X_n processes have the same W limit, then also the different R_n will have the same R limit. Similar remarks apply to $\max_t |X_n(t)| \rightarrow_d \max_t |W(t)|$.

(d) Show that the projection function $\pi_t(x) = x(t)$ is continuous, for each t . More generally, show that the projection function $\pi_{t_1, \dots, t_k}: C[0, 1] \rightarrow \mathbb{R}^k$ which takes x to its values $(x(t_1), \dots, x(t_k))$, for any finite set $0 \leq t_1 < \dots < t_k \leq 1$, is continuous. Thus finite-dimensional convergence,

$$(X_n(t_1), \dots, X_n(t_k)) \rightarrow_d (X(t_1), \dots, X(t_k)) \quad \text{for each } 0 \leq t_1 < \dots < t_k \leq 1,$$

is necessary for process convergence, by the continuous mapping theorem above, but not sufficient, as we shall see below.

(e) Consider the function $h: C[0, 1] \rightarrow C[0, 1]$ which transforms $x(t)$ to $x(t) - tx(1)$; note that it starts and ends at zero, forming a bridge. Show that this h is continuous, and generalise.

(f) Explain that the integral operation $h(x) = \int_0^1 x(t) dt$ is continuous. More generally, with g any continuous function on the real line, show that $h(x) = \int_0^1 g(x(t)) dt$ is continuous.

(g) Let $h(x) = \int_0^1 I(x(t) > 0) dt$ be the total time $x(t)$ has spent above zero. Show that h is ‘almost continuous’ [xx this to be made precise xx].

Ex. 10.3 *Brownian motion.* Here we introduce the Brownian motion process, also called the Wiener process. They are zero-mean Gaussian processes with independent increments over disjoint intervals. We shall see in Ex. 10.16 that this process can be reached as a limit in distribution of random functions in $C[0, 1]$, and by completeness is then also inside, i.e. its paths, though highly wiggly, are continuous.

(a) Consider then $W = \{W(t) : t \geq 0\}$, a random process evolving over time t , with the following properties: (i) $W(0) = 0$; (ii) increments $W(t) - W(s)$ have normal $N(0, t - s)$ distributions, i.e. with variance equal to the length of the interval in question; (iii) increments over disjoint intervals are independent. Show that $\text{cov}\{W(s), W(t)\} = \min(s, t)$, and for the random triple $(W(s), W(t), W(u))$, with $s < t < u$, find the 3×3 covariance matrix. the Brownian motion process

(b) It is incidentally fruitful and useful to think of the W process via cumulative small movements, as in $W(t)$ being the sum of many tiny and independent $dW(s) = W(s + ds) - W(s) \sim N(0, ds)$. Express $W(1)$ as the sum of 1000 tiny such $W((i+1)/n) - W(i/n)$, and show that you get the right variance.

(c) Show that with $t_1 < \dots < t_k$, then by necessity the vector $(W(t_1), \dots, W(t_k))$ must be multinormal, and give its $k \times k$ covariance matrix.

(d) Define $I = \int_0^1 W(t) dt$, a random integral. Find its distribution, perhaps via Riemann approximations.

(e) Suppose a fellow reader of this book tries to define a process $W^* = \{W^*(t) : t \geq 0\}$ by putting up (i), (ii), (iii) above, but now with $\text{Var}\{W^*(t) - W^*(s)\} = (t - s)^{1/2}$, rather than $t - s$. Show that this would backfire solidly. – So it’s not enough to put up ‘something’ for the variance or covariance function, as it may lead astray; one needs to check for logical coherency, and each covariance matrix for finite-dimensional distributions need to be nonnegative definite. Show that with $\text{Var}\{W^*(t) - W^*(s)\} = (t - s)^\alpha$, only $\alpha = 1$ secures such coherence.

(f) Note that *the existence* of the Brownian motion process is not entirely obvious, and there is no easy way to put down a full joint probability density for the full thing. The pure existence of a stochastic process is hence a more delicate and complex matter than when working in one- and finite-dimensional cases. Brownian motion indeed exists, by all means which may be proved in many ways, including via the Donsker theorem worked through in Ex. 10.16 below. Argue in fact that if there is a well-defined sequence $X_n = \{X_n(t) : t \geq 0\}$ of elements in $C[0, 1]$, which converges in distribution to a limit X , then that X must be inside $C[0, 1]$. In the exercise pointed to we construct W in this fashion, verifying also that this limit behaves according to (i), (ii), (iii) given above.

Ex. 10.4 *The Brownian bridge.* Various Gaussian processes can be constructed on the basis of Brownian motion $W = \{W(t) : t \geq 0\}$.

(a) Consider the process $W^0(t) = W(t) - tW(1)$ for $t \in [0, 1]$. Show that W^0 is zero-mean normal, and covariance function $\min(s, t) - st$, i.e. $s(1 - t)$ for $s \leq t$. In particular, its variance is $t(1 - t)$. The W^0 is called the *Brownian bridge*. the Brownian bridge

(b) For two disjoint time windows, say $[s_1, s_2]$ and $[t_1, t_2]$, with $s_1 < s_2 < t_1 < t_2$, find the covariance and correlation between the two bridge increments $W^0(s_2) - W^0(s_1)$ and $W^0(t_2) - W^0(t_1)$.

- (c) Show that W^0 also can be characterised as W conditional on $W(1) = 0$.
 (d) For a variation, study Brownian motion conditional on $I = \int_0^1 W(s) ds = 0$. Find the mean and covariance structure.

Ex. 10.5 *The Ornstein–Uhlenbeck process.* With W the Brownian motion process, construct $X(t) = W(\exp(2\rho t))/\exp(\rho t)$, say for $t \in [-c, c]$.

(a) Show that X is Gaussian, with mean zero and unit variance, and with covariance function $K(s, t) = \text{cov}\{X(s), X(t)\} = \exp(-\rho|t - s|)$, with correlations decaying exponentially. This is *the Ornstein–Uhlenbeck process*.

(b) With $s < t < u$, set up the covariance matrix for $(X(s), X(t), X(u))$. From this work out the distribution of $X(u)$ given $X(s) = a$ and $X(t) = b$. You should find that this depends on the value at t but not the value at s .

(c) For t in the unit interval, consider $Y(t) = W^0(t)/\{t(1 - t)\}^{1/2}$, the scaled Brownian bridge. Find its covariance function, and demonstrate that it can be represented as a time-transformed Ornstein–Uhlenbeck process.

Ex. 10.6 *Extending and scaling the time axis.* Above we have worked with Brownian motion as a process W defined as a process $W(t)$ for $t \in [0, 1]$. It may however easily be defined also for longer stretches of time, directly, or via scaling.

(a) A full Brownian motion process $W = \{W(t) : t \geq 0\}$, on any time interval, is defined via the basic properties laid out in Ex. 10.3: it starts at zero, and its increments are independent, with $W(t_2) - W(t_1) \sim N(0, t_2 - t_1)$. If W is such a process, and $c > 0$, show that the process $W^*(t) = W(ct)/c^{1/2}$, defined for $t \in [0, 1]$, is a standard Brownian motion in $C[0, 1]$.

(b) Let $M_t = \max_{0 \leq s \leq t} W(s)$ be the maximum value obtained during the time interval $[0, t]$. Show that M_t must have the same distribution as $t^{1/2}M_1$. In particular, $E M_t = \gamma_0 t^{1/2}$ and $\text{Var } M_t = \gamma_1 t$, where γ_0 and γ_1 are the mean and variance of M_1 . Here we actually know the distribution; check with Ex. 10.17 to see that $M_1 \sim |N(0, 1)|$, and deduce that $\max_{a \leq t \leq b} W(t) \sim (b - a)^{1/2}|N(0, 1)|$.

(c) With W a Brownian motion process, define $W^*(t) = tW(1/t)$ for $t > 0$. Show that its variance is t , which also allows the definition $W^*(0) = 0$. Show that W^* is another Brownian motion. It somehow has the ability to time-transform itself.

Ex. 10.7 *Gaussian processes, and their simulation.* We say that a process $X = \{X(t) : t \in [0, 1]\}$ is normal, or Gaussian, if all its finite-dimensional distributions are normal. Such processes are extensively used in probability theory and statistics.

(a) Explain that X is Gaussian, as defined now, if and only if all linear combinations $\sum_{j=1}^k c_j X(t_j)$ are normal. Go through the details to explain that the Brownian motion process is normal.

(b) To define and describe a normal process X , show that it is sufficient to give (i) the mean function $\xi(t) = E X(t)$ and (ii) the covariance function $K(s, t) = \text{cov}\{X(s), X(t)\}$. So nothing more is required than these two functions. Describe the Brownian motion process via its mean and covariance functions.

(c) Which functions $K(s, t)$ are bona fide covariance functions? First, it needs to be symmetric, with $K(s, t) = K(t, s)$. Show next that K needs to be *nonnegative definite* in the sense that each matrix finite-dimensional matrix $\Sigma = (K(t_i, t_j))_{i,j=1,\dots,k}$ is nonnegative definite, i.e. $\sum_{i,j} a_i a_j K(t_i, t_j) \geq 0$ for all $a = (a_1, \dots, a_k)$.

(d) How can we simulate many paths, from a given Gaussian process Z ? Take for

simplicity Z to have mean function zero. If Z is known to be the limit of some Z_n that can be simulated, we may use a high n and then simulate. A version of this scheme is to start with a grid of $t_i = i/n$, for such a high n . Consider Z_n , the process with values $Z_n(t_i) = Z(t_i)$, completed by using linear interpolation. Explain that $Z_n \rightarrow_d Z$. Hence it is sufficient to find a scheme to simulate such Z_n .

(e) For a given and perhaps high n , and writing $t_i = i/n$, form the $n \times n$ covariance matrix Σ with elements $K(t_i, t_j)$. A recipe for simulating from such a $N_n(0, \Sigma)$ has actually been given in Ex. 1.47, working in principle in any dimension, as for the intended high n here. In essence, use $\Sigma^{1/2}N$, with $N = (N_1, \dots, N_n)^t$ an i.i.d. string of standard normals.

this can be
accomplished
even with
 1000×1000
matrices

(f) To illustrate this, consider the Brownian bridge again, with covariance function $\min(s, t) - st$. Use e.g. $n = 100$ in your computer, form the $n \times n$ matrix, carry out the spectral decomposition as above; then simulate and display say 1000 paths.

(g) The recipe above is generic in nature, and works in principle for any given covariance function $K(s, t)$. There is a subclass of such where the processes behave in a Markovian way, however, see Ex. 13.22. It is shown there that if $K(s, t)K(t, u) = K(s, u)K(t, t)$ for all triples $s < t < u$, then the conditional distribution of $Z(u)$ given $Z(s)$ and $Z(t)$ depends only on the latter; the process forgets its past. This leads to a better understanding of the process. Check that the covariance functions for Brownian motion, the Brownian bridge, the Ornstein–Uhlenbeck process are of this type, i.e. the processes are Markovian. Use this Markovian characterisation to simulate paths, from these three processes, in a sequential manner, which is both simpler and more rapid computationally than with the generic recipe above.

Ex. 10.8 *Correlated Brownian motions, and new motions from old ones.* Start out considering X and Y , independent Brownian motions.

(a) For a $\rho \in (-1, 1)$, define $Z(t) = \rho X(t) + (1 - \rho^2)^{1/2}Y(t)$. Show that Z is another Brownian motion, with constant correlation ρ between $X(t)$ and $Z(t)$. For $\rho = 0.90$, display simulated paths of $(X, Z) = \{(X(t), Z(t)) : t \in [0, 1]\}$. Then do this for $\rho = -0.90$, and comment on what you see.

(b) Make similar constructions also for Brownian bridges.

(c) With two independent Brownian motions W_1, W_2 , show that $(1/\sqrt{2})W_1 + (1/\sqrt{2})W_2$ is a new Brownian motion. Generalise this little finding, and give similar statements for linear combinations of independent Brownian bridges.

Technical details, Prokhorov's theorem, more: to be moved to an Appendix?

(xx as of 16 August 2026, nils thinks the following approx 4 exercises can go to an appendix; we decide at emil's tavle in early august. This concerns Ex. 10.9, 10.10, 10.11, 10.12. xx)

Ex. 10.9 *Finite-dimensional distributions and uniqueness.* When working with convergence of stochastic process we are dealing with several measurable spaces. To clarify, let X be a stochastic process with values in $C[0, 1]$. This means that in the background there is a probability space $(\Omega, \mathcal{F}, \Pr)$ on which the X is defined, and that

$$X: \Omega \rightarrow C[0, 1],$$

that is, when $\omega \in \Omega$ is fixed $t \mapsto X(t, \omega)$ is a continuous function with domain $[0, 1]$ and values in $\mathbb{R}$. The space $C[0, 1]$ is equipped with its Borel- σ -algebra $\mathcal{C}$, i.e., the smallest σ -algebra containing all the open sets, where open is defined via the metric $d(x, y) = \sup_t |x(t) - y(t)|$ introduced above. Thus $(C[0, 1], \mathcal{C})$ is a measurable space, and the distribution of X , namely $P(B) = \Pr(X \in B)$, is a probability measure on $(C[0, 1], \mathcal{C})$ (as shown in Ex. C.3(a)). The function $\pi_{t_1, \dots, t_k}$ introduced above maps elements of $C[0, 1]$ into $\mathbb{R}^k$,

$$\pi_{t_1, \dots, t_k}(x) = (x(t_1), \dots, x(t_k)).$$

Thus, for A is in the Borel- σ -algebra $\mathcal{B}(\mathbb{R}^k)$,

$$P\pi_{t_1, \dots, t_k}^{-1}(A) = P(\pi_{t_1, \dots, t_k}(x) \in A) = P(\{x \in C[0, 1] : \pi_{t_1, \dots, t_k}(x) \in A\}),$$

is a distribution on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. The distributions $P\pi_{t_1, \dots, t_k}^{-1}$ for various time points $0 \leq t_1 < \dots < t_k \leq 1$ are the *finite-dimensional distributions* of P , while $\pi_{t_1, \dots, t_k}^{-1}(A)$ for $A \in \mathcal{B}(\mathbb{R}^k)$ are referred to as the *finite-dimensional sets*. In this exercise we show that the finite-dimensional sets determine distributions on $C[0, 1]$. This means that if P and Q are two probability measures on $(C[0, 1], \mathcal{C})$ such that

$$P\pi_{t_1, \dots, t_k}^{-1} = Q\pi_{t_1, \dots, t_k}^{-1} \quad \text{for all } 0 \leq t_1 < \dots < t_k \leq 1,$$

then $P = Q$ on $(C, \mathcal{C})$, that is $P(B) = Q(B)$ for all $B \in \mathcal{C}$.

(a) Let Π be the collection of all finite-dimensional sets. In particular, for some sets $B_1 \in \mathcal{B}(\mathbb{R})$ and $B_2 \in \mathcal{B}(\mathbb{R}^2)$ and some points $0 \leq t_1 < s < t_2 \leq 1$ the sets $A_1 = \pi_s^{-1}(B_1)$ and $A_2 = \pi_{t_1, t_2}^{-1}(B_2)$ are in Π . Introduce the functions $g_1(a, b, c) = b$ and $g_2(a, b, c) = (a, c)$ and notice that $\pi_s(x) = g_1(\pi_{t_1, s, t_2}(x))$ and $\pi_{t_1, t_2}(x) = g_2(\pi_{t_1, s, t_2}(x))$. Show that $A_1 \cap A_2$ is in Π . Generalise and deduce that Π is a π -system, i.e., a collection of sets that is closed under finite intersections.

(b) Let P and Q be two probability measures on $(C[0, 1], \mathcal{C})$ such that their finite-dimensional distributions agree, that is $P(A) = Q(A)$ for all $A \in \Pi$. Since Π is a π -system, it will follow from Dynkin's lemma that $P = Q$ on $(C[0, 1], \mathcal{C})$ provided Π generates $\mathcal{C}$ (see Ex. B.5(d)). As a first step towards showing that $\sigma(\Pi) = \mathcal{C}$, show that $\sigma(\Pi) \subset \mathcal{C}$.

(c) To show the reverse inclusion, we start by showing that the smallest σ -algebra on $C[0, 1]$ containing all the open balls, say $\sigma(\text{open balls})$, is contained in $\sigma(\Pi)$. Show that a closed ball $\overline{B(x, \varepsilon)} = \{y \in C[0, 1] : d(x, y) \leq \varepsilon\}$ can be written

$$\overline{B(x, \varepsilon)} = \bigcap_{q \in \mathbb{Q} \cap [0, 1]} \{y : |x(q) - y(q)| \leq \varepsilon\} = \bigcap_{q \in \mathbb{Q} \cap [0, 1]} \pi_q^{-1}[x(q) - \varepsilon, x(q) + \varepsilon],$$

and explain why this implies that $\sigma(\Pi)$ contains all the closed balls, and therefore all the open balls, implying that $\sigma(\text{open balls}) \subset \sigma(\Pi)$.

(d) On separable metric spaces the Borel- σ -algebra and the σ -algebra generated by the open balls coincide, $\sigma(\text{open sets}) = \sigma(\text{open balls})$. Since an open ball is an open set, $\sigma(\text{open balls}) \subset \sigma(\text{open sets})$. For the reverse inclusion, let G be an open set and use the separability of the metric space to show that $G \in \sigma(\text{open balls})$. – Since $C[0, 1]$ is a separable metric space, the lemma just proved implies that $\mathcal{C} = \sigma(\Pi)$, so together with (a)–(c) we have shown that a distribution on $C[0, 1]$ is fully characterised by its finite-dimensional distributions.

Ex. 10.10 *Relatively compact families of distributions.* A family $\mathcal{P}$ of distributions is relatively compact if every sequence $(P_n)_{n \geq 1} \subset \mathcal{P}$ contains a subsequence $(P_{n_k})_{k \geq 1}$ such that $P_{n_k} \Rightarrow Q$ for some distribution Q . In our applications, $\mathcal{P}$ will typically be a countable family of distributions $(P_n)_{n \geq 1}$, and $(P_n)_{n \geq 1}$ is relatively compact if every subsequence of $(P_{n_k})_{k \geq 1}$ contains a subsequence $(P_{n_{k_j}})_{j \geq 1}$ converging weakly to some probability measure Q .

(a) Show that $P_n \Rightarrow P$ if and only if each subsequence $(P_{n_k})_{k \geq 1}$ contains a subsequence $(P_{n_{k_j}})_{j \geq 1}$ converging weakly to P . In other words, $P_n \Rightarrow P$ if and only if $(P_n)_{n \geq 1}$ is relatively compact and all the subsubsequences have the same limit P .

(b) Let $(P_n)_{n \geq 1}$ and P be distributions on $(C[0, 1], \mathcal{C})$. Assume that $(P_n)_{n \geq 1}$ is relatively compact and that all the finite-dimensional distributions of P_n converge to those of P , that is $P_n \pi_{t_1, \dots, t_k}^{-1} \Rightarrow P \pi_{t_1, \dots, t_k}^{-1}$ for all $0 \leq t_1 < \dots < t_k \leq 1$. Show that $P_n \Rightarrow P$.

(c) Let $(P_n)_{n \geq 1}$ and P be distributions on $(C[0, 1], \mathcal{C})$. Assume that $(P_n)_{n \geq 1}$ is relatively compact and that for all $0 \leq t_1 < \dots < t_k \leq 1$,

$$P_n \pi_{t_1, \dots, t_k}^{-1} \Rightarrow \mu_{t_1, \dots, t_k},$$

for a distribution $\mu_{t_1 \dots t_k}$ on $\mathbb{R}^k$, for example a k -dimensional multivariate normal distribution. Show that this implies the existence(!) of a probability measure P on $(C[0, 1], \mathcal{C})$ such that $P \pi_{t_1, \dots, t_k}^{-1} = \mu_{t_1, \dots, t_k}$ for all $0 \leq t_1 < \dots < t_k \leq 1$, and deduce from (b) that P_n converges weakly to P . It is in essence this result that we will use to prove the existence of Brownian motion (as introduced in Ex. 10.3).

(d) Let $Y_1, \dots, Y_n$ be i.i.d. random variables with mean zero and unit variance. Consider the process

$$X_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor tn \rfloor} Y_i + \frac{t - \lfloor tn \rfloor / n}{1/n} \frac{1}{\sqrt{n}} Y_{\lfloor tn \rfloor + 1}, \quad t \in [0, 1],$$

that is $X_n(t) = n^{-1/2} \sum_{i=1}^{k-1} Y_i + \sqrt{n}(t - (k-1)/n) Y_k$ when $t \in [(k-1)/n, k/n]$. Show that each sample path $t \mapsto X_n(t, \omega)$ is in $C[0, 1]$, and that

$$(X_n(t_1), \dots, X_n(t_k)) \rightarrow_d (W(t_1), \dots, W(t_k)) \sim N(0, \Sigma_{t_1 \dots t_k}),$$

for all $0 \leq t_1 < \dots < t_k \leq 1$, where $\Sigma_{t_1 \dots t_k}$ is the $k \times k$ matrix with $t_1, \dots, t_k$ along the diagonal and $\min(t_i, t_j)$ off the diagonal. In view of the above, we now see that to conclude that X_n converges weakly to a Brownian motion it ‘only’ remains to show that $(X_n)_{n \geq 1}$ is relatively compact. For that we need the notion of tightness and Prokhorov’s theorem.

Ex. 10.11 *Technical results needed for Prokhorov’s theorem.* In order to prove Prokhorov’s theorem we must be equipped with a few technical lemmata. These are lemmata of independent interest, and are here included (in their not most general version) in order to arrive at the Daniell extension theorem, which is sort of the little brother of Kolmogorov’s extension theorem (the latter sometimes also being credited to Daniell).

(a) Recall from Ex. 2.23 that if F is a c.d.f. on $\mathbb{R}$, $g(u) = \inf\{x: F(x) \geq u\}$ for $u \in (0, 1)$, and $U \sim \text{unif}(0, 1)$, then $g(U) \sim F$. In this exercise we lift this result to conditional distributions. Let $(\mathcal{Y}, \mathcal{A})$ be a measurable space and let $(\mathbb{R}, \mathcal{B})$ be the real line equipped with its Borel- σ -algebra. Consider the function $\nu: \mathcal{Y} \times \mathcal{B} \rightarrow [0, 1]$ that is such that $y \mapsto \nu(y, B)$ is $\mathcal{A}$ -measurable for each $B \in \mathcal{B}$, and $B \mapsto \nu(y, B)$ is a probability measure

on $(\mathbb{R}, \mathcal{B})$ for each $y \in \mathcal{Y}$. Such a function is called a *probability kernel* or a *regular conditional probability* (see Ex. C.8 for the existence of such, and more details). Show that if ν is as just described, then $g(y, u) = \inf\{x: \nu(y, (-\infty, x]) \geq u\}$ for $u \in (0, 1)$ is such that for each $y \in \mathcal{Y}$

$$g(y, U) \sim \nu(y, \cdot), \quad \text{where } U \sim \text{unif}(0, 1).$$

(b) Suppose that X and Y are identically distributed random elements with values in the measurable space $(\mathcal{Y}, \mathcal{A})$, and that Y' is a random variable. Under these conditions the function ν defined by

$$\nu(Y, B) = \Pr(Y' \in B | Y), \quad \text{for } B \in \mathcal{B}(\mathbb{R}),$$

is a probability kernel, in particular $B \mapsto \nu(y, B)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for each $y \in \mathcal{Y}$ (again, see Ex. C.8 for more details). With the function g as defined in (a) we then know that $g(y, U) \sim \nu(y, \cdot)$ for $U \sim \text{unif}(0, 1)$. Take U to be independent of X and show that for any bounded and continuous $f: \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$

$$\mathbb{E} f(X, g(X, U)) = \mathbb{E} f(Y, Y'),$$

which shows that $(X, g(X, U)) \sim (Y, Y')$ [xx where is the exercise showing that this indeed characterises a distribution? xx]

(c) Suppose that for each $k \geq 1$ we are given a probability measure μ_k on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ with the *projective property* that for each $k \geq 1$

$$\mu_{k+1}(B \times \mathbb{R}) = \mu_k(B), \quad \text{for all } B \in \mathcal{B}(\mathbb{R}^k).$$

Explain that there is a finite random sequence $(Y_1, \dots, Y_k)$ such that $(Y_1, \dots, Y_k) \sim \mu_k$, and show that $(Y_1, \dots, Y_j) \sim \mu_j$ for all $j < k$.

(d) The aim now is to show that if $(X_1, \dots, X_k) \sim \mu_k$, then we can construct X_{k+1} such that $(X_1, \dots, X_k, X_{k+1}) \sim \mu_{k+1}$, that is, we go upwards instead of backwards, the latter being what we did in (c). Combine the projective property of $(\mu_k)_{k \geq 1}$ with (b) to show that we can indeed go upwards in this fashion. Deduce Daniell's extension theorem, namely that if $(\mu_k)_{k \geq 1}$ has the projective property then there exists a random sequence $(X_1, X_2, \dots)$ such that $(X_1, \dots, X_k) \sim \mu_k$ for each $k \geq 1$.

Ex. 10.12 Prokhorov's theorem. In Ex. 2.26 and Ex. 2.46 we proved Prokhorov's theorem for random variables and random vectors. The setting is now a separable and complete metric space (C, d) equipped with its Borel- σ -algebra $\mathcal{C}$, and a sequence $(P_n)_{n \geq 1}$ of probability measures on $(C, \mathcal{C})$. The sequence $(P_n)_{n \geq 1}$ is tight if for any $\varepsilon > 0$ there exists a compact set $K \subset C$ such that

$$P_n(K) > 1 - \varepsilon, \quad \text{for all } n.$$

The measures P_n will typically be the distributions of a sequence of random elements $(X_n)_{n \geq 1}$, thus $P_n = \Pr X_n^{-1}$ where $\Pr$ is a probability measure on a background probability space. In such cases the above display reads $\Pr(X_n \in K) > 1 - \varepsilon$ for all n , and it is clear what we mean when we say that $(X_n)_{n \geq 1}$ is tight in C . Prokhorov's theorem says that when (C, d) is a separable and complete metric space, tightness and relative compactness are the same. In $\mathbb{R}^k$ we proved Prokhorov's theorem by way of Helly's theorem. In this exercise we sketch a proof of Prokhorov's theorem for Polish metric

spaces that uses Prokhorov's theorem for $\mathbb{R}^k$ by mapping processes in C over to $\mathbb{R}^k$. The strategy is as follows. Any separable metric space is homeomorphic to a subset A of $\mathbb{R}^\infty$, meaning that there exists a bijective (one-to-one and onto) function $h: C \rightarrow A \subset \mathbb{R}^\infty$ such that both h and h^{-1} are continuous (see Ex. A.8 in the appendix). This entails that the problem of showing relative compactness from tightness can be carried out in $\mathbb{R}^\infty$. But by way of the approximation lemma in Ex. 2.30, the problem can be further reduced from taking place in $\mathbb{R}^\infty$ to taking place in $\mathbb{R}^k$, and we are back to Prokhorov's theorem as we know it from Chapter 2. In (c)–(f) we prove that tightness implies relative compactness when $C = \mathbb{R}^\infty$ and d is the metric $d(x, y) = \sum_{k=1}^\infty \min(|x_k - y_k|, 1)/2^k$. In (g) we specialise to random sequences concentrated in a subset of $\mathbb{R}^\infty$,

(a) Verify that the definition of tightness given here agrees with the definition given for random vectors in Ex. 2.46.

(b) Show that tightness is preserved under continuous mappings. That is, if $(X_n)_{n \geq 1}$ is a tight sequence in $\mathcal{X}$, and $g: \mathcal{X} \rightarrow \mathcal{Y}$ is a continuous function, then $g(X_n)$ is tight in $\mathcal{Y}$.

(c) Let $\mathbb{R}^\infty$ be the space of all infinite sequences $x = (x_1, x_2, \dots)$ equipped with the metric $d(x, y) = \sum_{k=1}^\infty \min(|x_k - y_k|, 1)/2^k$. Verify that d is indeed a metric, and that $d(x_n, x) \rightarrow 0$ if and only if $x_{n,k} \rightarrow x_k$ for every k . Show that $(\mathbb{R}^\infty, d)$ is separable and complete.

(d) Let $(X_n)_{n \geq 1} = \{(X_{n,1}, X_{n,2}, \dots)\}_{n \geq 1}$ be a sequence of random sequences and $X = (X_1, X_2, \dots)$ a random sequence. Let $a_1, a_2, \dots$ be some fixed real numbers, and set $A_{n,k} = (X_{n,1}, \dots, X_{n,k}, a_{k+1}, a_{k+2}, \dots)$ and $A_k = (X_1, \dots, X_k, a_{k+1}, a_{k+2}, \dots)$ for each $k \geq 1$ and $n \geq 1$. Use these and the lemma of Ex. 2.30 to show that $X_n \rightarrow_d X$ if and only if we have finite-dimensional convergence in the sense that

$$(X_{n,1}, \dots, X_{n,k}) \rightarrow_d (X_1, \dots, X_k),$$

for each $k \geq 1$. This result reduces the problem of weak convergence in $\mathbb{R}^\infty$ to the more familiar problem of weak convergence in $\mathbb{R}^k$.

(e) Assume that $C = \mathbb{R}^\infty$ and that $(X_n)_{n \geq 1} = \{(X_{n,1}, X_{n,2}, \dots)\}_{n \geq 1}$ is a tight sequence of random sequences. Combine Prokhorov's theorem for $\mathbb{R}^k$ with a diagonal argument (see Ex. 2.25 for an example thereof) to extract a subsequence $(n_m)_{m \geq 1}$ such that

$$(X_{n_m,1}, \dots, X_{n_m,k}) \rightarrow_d \mu_k, \quad \text{for each } k \geq 1,$$

where μ_k is a probability measure on $\mathbb{R}^k$. The point here is that the subsequence $(n_m)_{m \geq 1}$ should work for all $k \geq 1$.

(f) Show that the family of limiting distributions $(\mu_k)_{k \geq 1}$ has the projective property (as defined in Ex. 10.11(c)). Combine this with (d) to show that for the subsequence $(n_m)_{m \geq 1}$ from (e) we have $(X_{n_m,1}, X_{n_m,2}, \dots) \rightarrow_d (X_1, X_2, \dots)$, where $(X_1, X_2, \dots)$ is a random sequence such that $(X_1, \dots, X_k) \sim \mu_k$ for each $k \geq 1$.

(g) Suppose that $(X_n)_{n \geq 1} = \{(X_{n,1}, X_{n,2}, \dots)\}_{n \geq 1}$ is tight in a subset A of $\mathbb{R}^\infty$. This means that (A, d) is a metric space with d is the same metric as on $\mathbb{R}^\infty$, and that for every $\varepsilon > 0$ there is a compact set $K \subset A$ such that $\Pr(X_n \in K) > 1 - \varepsilon$ for all n . Importantly, a compact set K in $\mathbb{R}^\infty$ is closed (see Ex. A.8). Show that there is a subsequence $(n_m)_{m \geq 1}$ such that $X_{n_m} \rightarrow_d X$ for some random sequence X that lies in A with probability one.

(h) Finally, let $(X_n)_{n \geq 1}$ be a random sequence in a separable metric space C . Use the fact that C is homeomorphic to a subset A of $\mathbb{R}^\infty$ to extract a subsequence $(n_m)_{m \geq 1}$

such that $X_{n_m} \rightarrow X$ for a random element X in C . We have now proven that tightness implies relative compactness in separable metric spaces.

(i) For the converse, namely that relative compactness implies tightness, recall that we assume that C is separable and complete. Let $D = \{y_1, y_2, \dots\}$ be a countable set that is dense in C , then the collection $\{B(y, r) : y \in D\}$ of open balls covers C . Let $G_k = B(y_1, r) \cup \dots \cup B(y_k, r)$ and show that

$$\lim_{k \rightarrow \infty} \inf_n \Pr(X_n \in G_k) = 1. \quad (10.2)$$

To do so, you may argue that if this is not the case, then there is a subsequence $(n_k)_{k \geq 1}$ such that $\Pr(X_{n_k} \in G_k) = a < 1$ for all $k \geq 1$. But by relative compactness this subsequence has a further subsequence $(n_{k_j})_{j \geq 1}$ such that $X_{n_{k_j}} \rightarrow_d X$ for some random element X , so by the Portmanteau theorem

$$\Pr(X \in G_m) \leq \liminf_{j \rightarrow \infty} \Pr(X_{n_{k_j}} \in G_m) \leq a < 1, \quad \text{for all } m \geq 1.$$

Explain that this leads to the absurdity $1 < 1$, and hence (10.2) must be true.

(j) Take $r = 1/m$ and let $G_k^m = B(y_1, 1/m) \cup \dots \cup B(y_k, 1/m)$. Argue that by (10.2) we can for any $\varepsilon > 0$ find $(k_m)_{m \geq 1}$ such that

$$\inf_n \Pr(X_n \in G_{k_m}^m) \geq 1 - \frac{\varepsilon}{2m}, \quad \text{for all } m \geq 1.$$

Let $A = \bigcap_{m \geq 1} G_{k_m}^m$ and explain that its closure of $\bar{A}$ is compact (see Ex. A.8). Deduce that $\Pr(X_n \in \bar{A}) \geq 1 - \varepsilon$ for all $n \geq 1$, meaning that $(X_n)_{n \geq 1}$ is tight. This concludes the proof of Prokhorov's theorem.

Tightness, Arzelà–Ascoli, partial sums, Donsker theorem

Ex. 10.13 *Compactness in $C[0, 1]$ and the Arzelà–Ascoli theorem.* In view of Prokhorov's theorem as well as Ex. 10.10(b) & (c) we see that it is crucial to obtain a characterisation of the compact sets in $C[0, 1]$. First, since $C[0, 1]$ is a metric space, compactness and sequential compactness is the same thing (see Ex. A.8). Recall that set K is *sequentially compact* if every sequence $(x_n)_{n \geq 1} \subset K$ has a subsequence $(x_{n_k})_{k \geq 1}$ converging to an element x that lies in K . A set A is *relatively compact* if every sequence $(x_n)_{n \geq 1} \subset A$ has a subsequence $(x_{n_k})_{k \geq 1}$ converging to some element x , but x need not be contained in A (this is in line with the definition in Ex. 10.10 as soon as we put a metric on the space of probability distributions). The closure $\bar{A}$ must then contain x , meaning that if A is relatively compact, then $\bar{A}$ is sequentially compact, and hence compact since we are dealing with metric spaces. In general, a set is relatively compact provided its closure is compact. The Arzelà–Ascoli theorem says that a set $A \subset C[0, 1]$ is relatively compact if and only if A is bounded and uniformly equicontinuous. That A is bounded means that

$$\sup_{t \in [0, 1]} |x(t)| < \infty, \quad \text{for all } x \in A,$$

while A is uniformly equicontinuous if for any $\varepsilon > 0$ we can find $\delta > 0$ such that

$$\sup_{|t-s| < \delta} |x(t) - x(s)| < \varepsilon, \quad \text{for all } x \in A.$$

As we have seen in Ex. 10.1, any function in $C[0, 1]$ is uniformly continuous, thus uniform equicontinuity is a uniformity of this uniform continuity over a given collection of functions.

(a) We start by showing that if $A \subset C[0, 1]$ is bounded and uniformly equicontinuous, then A is relatively compact. Let $(x_n)_{n \geq 1}$ be a sequence in A , and use the boundedness of A and a diagonal argument to extract a subsequence $(n_k)_{k \geq 1}$ such that

$$x_{n_k}(q) \rightarrow x(q), \quad \text{for all } q \in \mathbb{Q} \cap [0, 1],$$

where $x(q)$ is for now only the pointwise limit of $x_{n_k}(q)$, i.e., $x(t)$ is not defined for t outside of $\mathbb{Q} \cap [0, 1]$.

(b) The interval $[0, 1]$ is compact, implying that for any $\delta > 0$ we can cover $[0, 1]$ by finitely many intervals $(t_1 - \delta, t_1 + \delta), \dots, (t_k - \delta, t_k + \delta)$. Use this fact, the finding from (a), and the uniform equicontinuity of A to show $(x_{n_k})_{k \geq 1}$ is a Cauchy sequence. Point to Ex. 10.1(b) and conclude that A is relatively compact.

(c) Prove the converse, namely that if A is relatively compact it must be bounded and uniformly equicontinuous.

(d) A more succinct formulation of the Arzelà–Ascoli theorem is obtained by introducing the *modulus of continuity* on $C[0, 1]$,

modulus of
continuity

$$w(x, \delta) = \sup_{|t-s| < \delta} |x(t) - x(s)|,$$

for any function x and positive δ . Any function x on $[0, 1]$ can be written in terms of a telescoping sum as follows $x(t) = x(0) + \sum_{j=1}^k \{x(jt/k) - x((j-1)t/k)\}$. Use this to show that an equivalent formulation of the Arzelà–Ascoli theorem is that $A \subset C[0, 1]$ is relatively compact if and only if

$$\sup_{x \in A} |x(0)| < \infty, \quad \text{and} \quad \sup_{x \in A} w(x, \delta) \rightarrow 0 \text{ as } \delta \rightarrow 0.$$

This way of formulating the theorem is also more convenient for the proofs to come.

Ex. 10.14 *Tightness and the Arzelà–Ascoli theorem.* We learn from Ex. 10.10(b)&(c) and Ex. 10.12 that what we need for proving weak convergence in $C[0, 1]$ are finite-dimensional convergence, for which we have many tools from Ch. 2; and tightness. Here we go into the tightness.

(a) Use the Arzelà–Ascoli theorem of Ex. 10.13 to show the X_n sequence is tight provided (i) the sequence $X_n(0)$ in $\mathbb{R}$ is tight (bounded in probability); (ii) the equi-continuity condition, that for each positive ε , $\limsup \Pr(w(X_n, \delta) \geq \varepsilon) \rightarrow 0$ as $\delta \rightarrow 0$.

(b) So we need to control $w(X, \delta) = \sup_{|t-s| < \delta} |X(t) - X(s)|$, for various processes, and attempt to do this via subintervals. Show that with $0 = t_0 < t_1 < \dots < t_{k-1} < t_k = 1$, with $t_i - t_{i-1} \geq \delta$, then

$$\Pr(w(X, \delta) \geq 3\varepsilon) \leq \sum_{i=1}^k \Pr\left(\sup_{t_{i-1} \leq t \leq t_i} |X(t) - X(t_{i-1})| \geq \varepsilon\right).$$

In particular, for $\delta = 1/k$,

$$\Pr(w(X, 1/k) \geq 3\varepsilon) \leq \sum_{i=1}^k \Pr\left(\sup_{(i-1)/k \leq t \leq i/k} |X(t) - X((i-1)/k)| \geq \varepsilon\right).$$

Clearly, bounding such probabilities, for a sequence of X_n , depends in various ways on the probabilistic particularities of the sequence.

Ex. 10.15 *Tail inequalities for maxima of partial sums.* We learn from Ex. 10.14 that criteria for tightness must involve bounds for tail probabilities for random variables of the type $\sup_{t_1 \leq t \leq t_2} |X_n(t) - X_n(t_1)|$, and these again relate to partial sums of smaller increments $X_n(s + \varepsilon) - X_n(s)$. To set the stage for several useful inequalities of this type, start out with i.i.d. variables U_i , with mean zero and variance one, and their random walk partial sums $S_j = U_1 + \cdots + U_j$. Define $A_m = \max_{j \leq m} |S_j|/\sqrt{m}$, and put $p_m(a) = \Pr(A_m \geq a)$.

(a) Show that $p_m(a) \leq 1/a^2$; use the Kolmogorov inequality of Ex. 2.58. Note that this also generalises the Chebyshev inequality. For different variances, generalise to $p_m(a) \leq (1/a^2)(1/m) \sum_{j=1}^m \sigma_j^2$.

(b) The following leads to what in various cases is a stronger result. First verify that the event $\{A_m \geq a\}$ can be decomposed into $\cup_{i=1}^m E_i$, with $E_i = \{\max_{j < i} |S_j|/\sqrt{m} < a \leq |S_i|/\sqrt{m}\}$. For $c \in (0, a)$, show then that

$$p_m(a) \leq \Pr(|S_m|/\sqrt{m} \geq a - c) + \sum_{i=1}^{m-1} \Pr(E_i, |S_m|/\sqrt{m} < a - c),$$

with the sum further bounded by

$$\begin{aligned} \sum_{i=1}^{m-1} \Pr(E_i, |S_m - S_i|/\sqrt{m} \geq c) &\leq \Pr(|S_m|/\sqrt{m} \geq a - c) + \sum_{i=1}^{m-1} \Pr(E_i) \frac{m-i}{mc^2} \\ &\leq \Pr(|S_m|/\sqrt{m} \geq a - c) + (1/c^2)p_m(a). \end{aligned}$$

Show that this leads to $(1 - 1/c^2)p_m(a) \leq \Pr(|S_m|/\sqrt{m} \geq a - c)$. For $c = \sqrt{2}$, you find $p_m(a) \leq 2\Pr(|S_m|/\sqrt{m} \geq a - \sqrt{2})$, which we shall use below.

(c) The probability of $|S_m|/\sqrt{m} \geq a - \sqrt{2}$ can be bounded separately, either by Markov type inequalities, if e.g. the third moment is finite, or via the CLT fact that $\Pr(|S_m|/\sqrt{m} \geq a - \sqrt{2}) \rightarrow \Pr(N(0, 1) \geq a - \sqrt{2})$. This is less than e.g. $3/(a - \sqrt{2})^4$ for any $a > \sqrt{2}$. Show hence that $p_m(a) \leq 2 \cdot 3.01/(a - \sqrt{2})^4$ for all big enough m , even when we do not assume more than the finite variance.

(d) (xx we include useful variations, with different variances σ_j^2 and phrase results in terms of $s_m^2 = \sigma_1^2 + \cdots + \sigma_m^2$. the arguments are essentially the same as for the i.i.d. case. xx)

Ex. 10.16 *Partial-sum processes and the Donsker theorem.* Let $U_1, U_2, \dots$ be i.i.d., with mean zero and variance one (and we do not need to say anything more regarding their distribution). Then we know from the CLT that $n^{-1/2}(U_1 + \cdots + U_n) \rightarrow_d N(0, 1)$. But we also see that e.g. $n^{-1/2}(U_1 + \cdots + U_{[n/2]})$ must have a normal limit, and similarly with other partial sums. This exercise goes through the basics for the full partial-sum process, and leads to the fundamental and very useful Donsker theorem. Define indeed the random process $X_n = \{X_n(t) : t \geq 0\}$ via

$$X_n(t) = n^{-1/2} \sum_{i \leq [nt]} U_i = (U_1 + \cdots + U_{[nt]})/\sqrt{n} \quad \text{at values } t = i/n,$$

with linear interpolation for $t \in [i/n, (i+1)/n]$ to keep X_n inside $C[0, 1]$. We shall indeed prove *Donsker's theorem*, that $X_n \rightarrow_d W$, with W the Brownian motion process introduced in Ex. 10.3. The linear interpolation is essentially merely cosmetic, to keep X_n inside $C[0, 1]$. With $X_n^*(t)$ more naturally defined as $n^{-1/2} \sum_{i \leq [nt]} U_i$ for all t , without

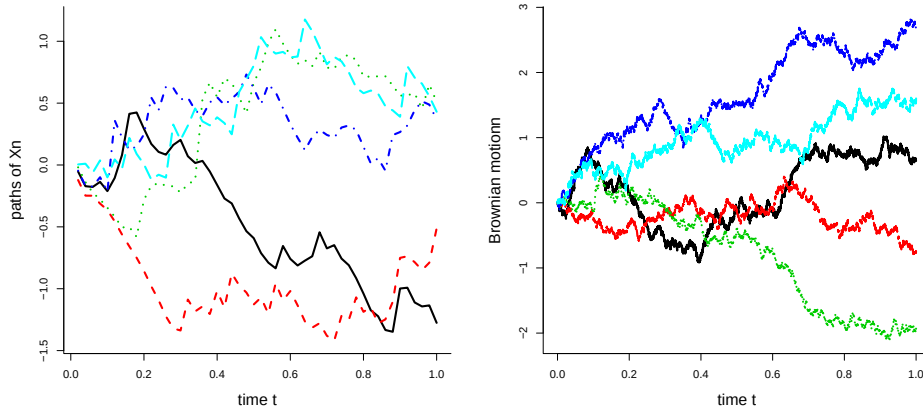


Figure 10.1: Simulated partial-sum processes X_n , as defined in Ex. 10.16, with $n = 50$ (left panel) and $n = 10^4$ (right panel), with i.i.d. components $U_i = V_i - 1$ and $V_i \sim \text{Expo}(1)$. For large n the eye can barely see the difference between these paths and real Brownian motion.

linear interpolation but with jumps of size $O_{\text{pr}}(1/\sqrt{n})$, we shall see in Ex. 10.21 that we still have $X_n^* \rightarrow_d W$, but then in the bigger natural space $D[0, 1]$.

(a) Use one or two of your favourite distributions for U_i , with mean zero and variance one, to generate some sample paths of X_n . We've done this for Figure 10.1, with $n = 50$ and $n = 10^4$, actually using $U_i = V_i - 1$, with the V_i being unit exponential; the intended point is to use something with a skewness, though that aspect disappears in the limit.

(b) Fix t positive. Show that $X_n(t)$ has mean zero and variance $[nt]/n$ converging to t . Show also that $X_n(t) \rightarrow_d N(0, t)$. In view of the Brownian motion process introduced in Ex. 10.3, we may write $X_n(t) \rightarrow_d W(t)$.

(c) Show that if $s < t < u < v$, then $(X_n(t) - X_n(s), X_n(v) - X_n(u))$ tends to $(W(t) - W(s), W(v) - W(u))$, two independent normal pieces with variances $t - s$ and $v - u$. More generally, show that there is finite-dimensional convergence in distribution; for $t_1 < \dots < t_k$, we have $(X_n(t_1), \dots, X_n(t_k)) \rightarrow_d (W(t_1), \dots, W(t_k))$.

(d) It remains to demonstrate that the X_n sequence is tight, where we shall use the characterisations and oscillation bounds from Ex. 10.14, along with probability tail inequalities from Ex. 10.15. Show that $\Pr(w(X_n, 1/k) \geq 3\varepsilon) \leq k \Pr(B_k \geq \varepsilon)$, where $B_k = \sup_{t \leq 1/k} |X_n(t)|$. We have $X_n(t) = S_j/\sqrt{n}$ for these $t = j/n \in [0, 1/k]$, up to $t_0 = j_0/n = 1/k$, so indexes at work are those with $j/n \leq 1/k$, or $j \leq n/k$. Show hence that $B_k = \max_{j \leq n/k} |S_j|/\sqrt{n} = A_{[n/k]}/k^{1/2}$, in terms of $A_{[n/k]} = \max_{j \leq n/k} |S_j|/([n/k])^{1/2}$, worked with in Ex. 10.15; the integer value modification $[n/k]$ is there merely to avoid saying that n/k needs to be an integer. With notation from that exercise, deduce that

$$k \Pr(B_k \geq \varepsilon) = k p_{[n/k]}(k^{1/2}\varepsilon) \leq 2k \Pr(|S_{[n/k]}|/([n/k])^{1/2} \geq k^{1/2}\varepsilon - \sqrt{2}),$$

Show that we for any given $\eta > 0$ have $k \Pr(B_k \geq \varepsilon) < \eta$ for all large enough k , and that this secures tightness, hence concluding the proof for the Donsker theorem.

(e) Explain that the $X_n \rightarrow_d W$ result also implies the existence of the Brownian motion, as a random element inside $C[0, 1]$, i.e. with continuous paths.

Ex. 10.17 *Applying the Donsker theorem.* In Ex. 10.16 we have established that the natural normalised partial-sum process X_n tends to Brownian motion W , and we may read off corollaries via the continuous functionals pointed to in Ex. 10.2, and yet others.

(a) A very simple continuous function is $h(x) = x(1)$, reading off the value of $x(t)$ at the endpoint $t = 1$. It is continuous. Hence $X_n(1) \rightarrow_d X(1)$. Go back to Donsker and deduce the CLT. In this light, the Donsker theorem is really a much bigger sibling of the CLT; the Donsker has a thousand corollaries, some simple, many complex, and one of these is the CLT.

(b) Consider $h(x) = \max_{0 \leq t \leq 1} x(t)$. Deduce that with $M_n = \max_{i \leq n} S_i$, where $S_i = U_1 + \cdots + U_i$, we have $M_n/\sqrt{n} = \max_{0 \leq t \leq 1} X_n(t) \rightarrow_d M = \max_{0 \leq t \leq 1} W(t)$. Note that this limit ensues regardless of the distribution of the i.i.d. sequence U_i , as long as these have mean zero and variance one. The limit is the same, with U_i being standard normal, or symmetric ± 1 variables, or of $U_i = V_i - 1$ with unit exponential V_i , etc.

(c) Often enough the limit $h(W)$ of $h(X_n)$ may be complicated, not admitting a clear explicit representation, in which case one may resort to simulations. In this particular case, with $M_n/\sqrt{n} \rightarrow_d M = \max_{0 \leq t \leq 1} W(t)$, an explicit solution can be found, however. Choose $U_i = \pm 1$ with probabilities $\frac{1}{2}, \frac{1}{2}$, so that the partial sums $S_i = U_1 + \cdots + U_i$ form a random walk, with each of the 2^n possible paths for S_n having the same probability $1/2^n$. Argue first that

$$\Pr(M_n \geq a) = \Pr(M_n \geq a, S_n > a) + \Pr(M_n \geq a, S_n < a) + \Pr(M_n \geq a, S_n = a),$$

and then show that the two first terms must be equal, whereas the third is small. Deduce from this that $\Pr(M_n/\sqrt{n} \geq u) = 2\Pr(S_n > \sqrt{n}u) + \varepsilon_n$, with $\varepsilon_n \rightarrow 0$, with resulting limit $\Pr(M \geq u) = 2\Pr(N(0, 1) \geq u) = 2\{1 - \Phi(u)\}$ for $u \geq 0$. Verify that this also means that $M \sim |N(0, 1)|$.

(d) Alice and Bob play a long game. Alice's winning in game i is U_i , with a distribution having mean zero and variance one; if it's positive, good for Alice, if it's negative, good for Bob. After i games, Alice's account has $U_1 + \cdots + U_i$, and Bob's account has minus this sum. They play a few thousand times. – How much of the time has Alice been in the lead? This is food for the Donsker theorem, since this concerns

$$T_n = (1/n) \sum_{i=1}^n I\{U_1 + \cdots + U_i > 0\} = m\{t \in [0, 1]: X_n(t) > 0\},$$

the Lebesgue measure of how much time the process has been above zero. This is a continuous functional, so $T_n \rightarrow_d T = m\{t \in [0, 1]: W(t) > 0\}$. The problem hence has a clear-cut general solution: Alice has been in the lead, over Bob, a portion T_n of the time, and T_n tends to T in distribution. So it's 'only' a matter of finding the distribution of T . This is non-trivial, but the solution is the $\text{Beta}(\frac{1}{2}, \frac{1}{2})$, intriguingly, with U-shaped density $(1/\pi)/\{t(1-t)\}^{1/2}$ on $(0, 1)$. Thus there's a high chance that *one of the two* has been leading for a very long time; the least likely outcome is that they've been leading about half the time.

Ex. 10.18 *Some functionals of Brownian motion and Brownian bridge.* We have constructed and properly defined the Brownian motion W and Brownian bridge W^0 , random continuous functions on $[0, 1]$. Various functionals, say $h_1(W)$ and $h_2(W^0)$, continuous functions $C[0, 1] \rightarrow \mathbb{R}$, are of sufficient interest to warrant examination of their distributions. For a given such $h(W)$, say, we may either attempt to work out its distribution

directly, or find it via the Donsker theorem and continuity, since $h(X_n) \rightarrow_d h(W)$ when $X_n \rightarrow_d W$; we have seen some such examples in Ex. 10.17. It is also typically possible and not complicated to simulate say 10^5 paths of W (or very close approximations thereof), and read off the distribution of $h(W)$ in that way.

(a) Consider $I = h(W) = \int_0^1 W(t) dt$. Show that $I_n \rightarrow_d I$, where $I_n = n^{-1} \sum_{i=1}^n W(i/n)$. Explain that I_n must be a normal $(0, \sigma_n^2)$; find the variance in question, and its limit. Conclude that $I \sim N(0, 1/3)$. Show similarly that $\int_0^1 W^0(t) dt$ has the $N(0, 1/12)$ distribution.

(b) Find mean and variance for the random integrals $A_1 = \int_0^1 W(t)^2 dt$, $A_2 = \int_0^1 W^0(t)^2 dt$, $A_3 = \int_0^1 W^0(t)^2 / \{t(1-t)\} dt$, $A_4 = \int_0^1 \exp(W(t)) dt$.

(c) With e.g. `sim = 100`, `N = 1000`, explain that the following simple R code produces `sim` simulated paths of Brownian motion $W(t)$ and Brownian bridge $W^0(t)$, at the fine grid of all points j/N :

```
tval = seq(1/N, 1, by=1/N)
Wsim = 0*tval %*% t(1:sim)
W0sim = 0*tval %*% t(1:sim)
for (ss in 1:sim)
{ Wsim[, ss] = cumsum(rnorm(N))/sqrt(N)
  W0sim[, ss] = Wsim[, ss] - tval*Wsim[N, ss] }
matplot(tval, W0sim, type="l", lwd=3)
```

This may now be followed by computing and plotting the required $h(W)$ or $h(W^0)$. With `Wmax = 0*(1:sim)`, `W0absmax = 0*(1:sim)`, and similar, verify that

```
for (ss in 1:sim)
{ Wmax[ss] = max(Wsim[, ss])
  W0absmax[ss] = max(abs(W0sim[, ss])) }
```

produces the intended simulated realisations of these functions. Carry out this for a suitably high number of `sim`, perhaps also with a high `N`, to simulate the distribution of $W_{\max}^0 = \max_{0 \leq t \leq 1} |W^0(t)|$ and $W_{\max} = \max_{0 \leq t \leq 1} |W(t)|$. Record the quantiles at levels 0.50, 0.90, 0.95, 0.975, 0.99; your answers should be close to (0.828, 1.224, 1.358, 1.480, 1.627) for the absolute bridge and (1.149, 1.960, 2.241, 2.807) for the absolute motion.

(d) We have $W^0(s) \sim N(0, s(1-s))$, so the variable $W^0(s)/\{s(1-s)\}^{1/2}$ is standard normal. Now consider

$$D = \max_{\varepsilon \leq s \leq 1-\varepsilon} |W^0(s)|/\{s(1-s)\}^{1/2},$$

for a fixed $\varepsilon \in (0, \frac{1}{2})$. Various naturally built test statistics have limit distributions related to this D , so it might be necessary to work out say the 0.90, 0.95, 0.99 quantiles of its distribution. There are no explicit formulae for this, so the apparently simplest method is to simulate a large number of W^0 paths (or close approximations to such) and then read off A values. Explain how this can be done. Find the 0.90, 0.95, 0.99 quantiles for intervals $[0.025, 0.975]$, $[0.05, 0.95]$, $[0.10, 0.90]$, displaying also the A distributions via fine histograms. (xx Answers, needed below, are xx.)

	0.50	0.90	0.95	0.99
eps = 0.025:	2.137	2.927	3.184	3.692
eps = 0.05:	2.051	2.858	3.121	3.649
eps = 0.10:	1.931	2.768	3.041	3.577

(e) In the long run, we found in Ex. 10.17, Alice leads over Bob in a proportion $T = m\{t \in [0, 1]: W(t) > 0\}$ of the time. Use simulations, as above, to construct a fine histogram for its distribution.

so probabilities
can be
computed,
without
knowing the
precise
theoretical
Beta($\frac{1}{2}, \frac{1}{2}$)
result

The space D , independent increments, Gamma and Poisson processes

the $D[0,1]$ space
of functions

Ex. 10.19 *Tightness and convergence in the function space $D[0, 1]$.* A long list of crucial processes have discontinuities, for which we need a bigger space to work with than the $C[0, 1]$. Let indeed $D[0, 1]$ consist of all functions $x: [0, 1] \rightarrow R$ which are right-continuous with left-hand limits. We need a metric for which two functions x and y are deemed close not merely if $|y(t) - x(t)|$ is small for all t , but also if $|\tilde{y}(t) - x(t)|$ can be small, for a slight time-transformation $\tilde{y}(t) = y(\lambda(t))$. The *Skorohod metric* is the customary metric for such purposes. Let Λ be the set of all strictly increasing, smooth transformations on $[0, 1]$ with $\lambda(0) = 0$ and $\lambda(1) = 1$. Writing $\text{id}(t) = t$ for the identity function and $d_u(x, y) = \sup_t |y(t) - x(t)|$ for the uniform metric, i.e. the metric used for our work with the $C[0, 1]$, let $d(x, y)$ be the smallest ε for which

the Skorohod
metric

$$d_u(\lambda, \text{id}) \leq \varepsilon \quad \text{and} \quad d_u(y \circ \lambda, x) \leq \varepsilon;$$

thus the latter is $\sup_t |y(\lambda(t)) - x(t)|$. It becomes beyond the scope of this book to go into all the mathematical details pertaining to the $D[0, 1]$ space with its d metric, but we list the salient points below. These can be viewed as either very hard problems for the most eager readers, or as a convenient summary list for results reached in the broader literature of stochastic processes and convergence in function spaces; see pointers in our Notes & Pointers. The crucial points for us, in this chapter and for a subset of our Statistical Stories, are

a classic
reference
remains
Billingsley
(1968)

- (i) that many important processes live in $D[0, 1]$;
- (ii) that various empirical processes X_n do converge to clear limits X ;
- (iii) that the criterion given in point (d) indeed guarantees tightness;
- (iv) that tightness of X_n plus finite-dimensional convergence $(X_n(t_1), \dots, X_n(t_k)) \rightarrow_d (X(t_1), \dots, X(t_k))$ for all finite sets $t_1 < \dots < t_k$ secures the desired process convergence $X_n \rightarrow_d X$.

In fact, this last condition may be relaxed a bit; it suffices to have such finite-dimensional convergence for all $t_1, \dots, t_k \in T$, where T contains 1, is dense, and consists of continuity points for X , i.e. points t for which $\Pr(X \text{ is discontinuous at } t) = 0$.

(a) First, the d is really a metric. We have $d(x_n, x) \rightarrow 0$ equivalent to there being a sequence of time machines for which $x_n(\lambda_n(t)) \rightarrow x(t)$ uniformly. The $D[0, 1]$ is complete and separable. Also, $C[0, 1]$ is a closed subset.

(b) For $T_0 \subset [0, 1]$, e.g. an interval, let $w(x, T_0)$ be the supremum of all differences $|x(t) - x(s)|$ with $s, t \in T_0$. A useful *D space modulus* is now

$$w''(x, \delta) = \sup \min\{|x(t) - x(s)|, |x(u) - x(t)|\},$$

with the supremum taken over all $s < t < u$ with $u - t \leq \delta$. Find $w''(x, \delta)$ for $x(t) = \exp(ct)$ and for $x(t) = \sin t$ and perhaps for a few more functions.

(c) Just as the Arzelà–Ascoli theorem is driving essential parts of the $C[0, 1]$ space methodology and applications, we need a corresponding but harder theorem characterising compact subsets of $(D[0, 1], d)$. The achievement is that K is compact if and only

if (i) $\sup_{x \in K} \sup_t |x(t)|$ is finite; (ii) $\sup_{x \in K} w(x, [0, \delta))$ and $\sup_{x \in K} w(x, [1 - \delta, 1])$ go both to zero as $\delta \rightarrow 0$; and (iii) $\sup_{x \in K} w''(x, \delta) \rightarrow 0$ as $\delta \rightarrow 0$. The crucial aspect to verify, for stochastic processes X_n , is typically related to point (iii) here, to ascertain that $w''(X_n, \delta)$ is small with high enough probability.

(d) Consider a sequence X_n of processes in $D[0, 1]$. Assume there are nondecreasing functions G_n converging to a continuous increasing function G on $[0, 1]$, such that, for some $\gamma > 0$ and $\alpha > 1$,

$$E |X_n(t) - X_n(s)|^\gamma |X_n(u) - X_n(t)|^\gamma \leq \{G_n(u) - G_n(s)\}^\alpha \quad (10.3)$$

holds for all $s < t < u$. Then X_n is tight.

(e) Consider a sequence of random elements X_n in $D[0, 1]$, with potential limit process X . Let T the set of t at which the projection function $\pi_t(x) = x(t)$ is continuous for X , i.e. the probability that X is continuous at position t is 1. If X_n is tight, typically via (10.3), then $X_n \rightarrow_d X$.

(f) Rather frequently tightness condition (10.3) holds and can be verified. There are however instances where X_n is tight, but with that condition failing. Let τ_n be uniform, and examine $X_n(t) = \sqrt{n} J_n I(t \geq \tau_n)$, where $J_n \sim \text{binom}(1, 1/n)$. So most of the time $X_n(t)$ is simply zero, but when jumps occur, they are big. Show that X_n is tight, but that (10.3) does not hold.

Ex. 10.20 *Criteria for existence of processes in $D[0, 1]$.* Suppose we set up a system of finite-dimensional distributions for a stochastic process $X = \{X(t) : t \in [0, 1]\}$ on the unit interval. These need to have or imply a minimum of regularity in order to ensure that yes, such an X process really exists. Let $\mu_{t_1, \dots, t_k}$ be the probability measure for the envisaged $(X(t_1), \dots, X(t_k))$. First, these need to be coherently defined, in the Kolmogorov sense: when integrating out the third element in $\mu_{s, t, u}$ we should get $\mu_{s, t}$, etc., and if we for say $s < t < u < v$ start out with the assigned distribution $\mu_{s, t, u, v}$, and then compute the distribution at (v, u, t, s) , we ought to get the specified $\mu_{v, u, t, s}$, etc. Suppose

$$\mu_{t, t+h} \{(u_1, u_2) : |u_1 - u_2| \geq \varepsilon\} \rightarrow 0 \quad \text{as } h \rightarrow 0, \quad (10.4)$$

and that we for all $t_1 \leq t \leq t_2$ have

$$\int |v - v_1|^\gamma |v_2 - v|^\gamma d\mu_{s, t, u}(v_1, v, v_2) \leq \{G(u) - G(s)\}^\alpha, \quad (10.5)$$

for some $\gamma > 0$ and $\alpha > 1$, where G is a nondecreasing and continuous function on $[0, 1]$. Then Theorem 15.7 in Billingsley (1968) secures the existence of an X in $D[0, 1]$ having these $\mu_{t_1, \dots, t_k}$ as its finite-dimensional distributions.

(a) Suppose that such an $X = \{X(t) : t \in [0, 1]\}$ really exists. Show that the left hand sides of (10.4) and (10.5) then can be written in the more familiar ways $\Pr(|X(t+h) - X(t)| \geq \varepsilon)$ and $E |X(t) - X(t_1)|^\gamma |X(t_2) - X(t)|^\gamma$. The point is that we in principle do not yet know that there is such an X in $D[0, 1]$, hence the somewhat pedantic notation.

(b) Specify finite-dimensional distributions by saying that for each t , $\mu_t \sim \text{Pois}(\lambda t)$, that for $s < t$ we have $\mu_{s, t}$ equal to the distribution of $(A, A + B)$ with $A \sim \text{Pois}(\lambda s)$ and $B \sim \text{Pois}(\lambda(t - s))$ independent, and so on, with independent Poisson increments for any $t_1 < \dots < t_k$. Show that conditions (10.4) and (10.5) hold, actually with $\gamma = 1$ and $\alpha = 2$. This ensures the existence of a Poisson process in $D[0, 1]$, with independent increments and $X(t) - X(s) \sim \text{Pois}(\lambda(t - s))$ for any $s < t$.

(c) Generalise the above to demonstrate the existence in $D[0, 1]$ of any time-inhomogeneous Poisson process, with mean function $E X(t) = \Lambda(t) = \int_0^t \lambda(s) ds$, where $\lambda(s)$ is any non-negative integrable intensity function. The standard Poisson process then corresponds to this intensity function being a constant λ .

(d) Use the criteria above to give another proof of the existence of the Brownian motion, as a process W in $D[0, 1]$. Here we already know that $\Pr(W \in C[0, 1]) = 1$.

(e) Show the existence of a *Gamma process* in $D[0, 1]$, defined via independent increments and $Z(t) - Z(s) \sim \text{Gam}(a(t-s), b)$, where a, b are given positive parameters. Show that its mean and variance are $(a/b)t$ and $(a/b^2)t$.

(f) Let $X = \{X(t) : t \in [0, 1]\}$ be a simple stationary Markov process with values in $\{0, 1\}$, starting in 0, and with transition probabilities $p_{a,b}(t) = \{X(s+t) = b \mid X(s) = a\}$ for $a, b \in \{0, 1\}$. We have $\sum_b p_{a,b}(t) = 1$ and $p_{a,b}(s+t) = \sum_c p_{a,c}(s)p_{c,b}(t)$ for all a, b and s, t . Show that X is really a random process in $D[0, 1]$. [xx perhaps a helpful hint here. xx] Generalise to any time-continuous Markov chain.

(g) xx make the point, via a simple exercise, that another way to prove existence in $D[0, 1]$ of some X is to derive it as a limit, $X_n \rightarrow_d X$, with $X_n \in D[0, 1]$, but this needs convergence criteria, to which we turn now. the point, from $\limsup P_n(F) \leq P(F)$, for any closed F , is that if all $X_n \in F$, then certainly also $X \in F$. xx

Ex. 10.21 Independent increment processes. Consider processes X_n in $D[0, 1]$ with independent increments; $X_n(u) - X_n(t)$ is independent of $X_n(t) - X_n(s)$ are independent, for any $s < t < u$, and similarly for any collection of disjoint intervals.

(a) Demonstrating tightness for independent increment processes becomes easier than for the general cases. Suppose there are nondecreasing functions G_n , tending to a continuous increasing G on $[0, 1]$, such that for some $\gamma > 0$

$$E |X_n(t) - X_n(s)|^\gamma \leq \{G_n(t) - G_n(s)\}^\alpha \quad \text{for all } s < t,$$

this time for some $\alpha > \frac{1}{2}$. Show that the X_n sequence is tight.

(b) Consider once more the Donsker theorem setup of Ex. 10.16, with $X_n(t) = n^{-1/2} \sum_{i \leq [nt]} U_i$, now defined without linear interpolation. Demonstrate tightness of X_n , again, verifying the simpler criterion $E |X_n(t) - X_n(s)|^2 = ([nt] - [ns])/n$.

(c) Construct a random process as follows. With $V_1, \dots, V_m$ i.i.d. $\text{Gam}(a/m, 1)$, let $Z_m(t) = \sum_{i \leq [mt]} V_i$, for simplicity on $[0, 1]$, though time t can run longer. Show that $Z_m(i/m) \sim \text{Gam}(ai/m, 1)$, and that $Z_m(t) - Z_m(s) \rightarrow_d \text{Gam}(a(t-s), 1)$ for each interval $[s, t]$. Generalise this to a statement with all finite-dimensional distributions. Argue thus that $Z_m \rightarrow_d Z$, in $D[0, 1]$, with a well-defined Gamma process limit, having independent Gamma distributed increments.

for generalised
gamma
processes see
Ex. 14.2

Ex. 10.22 Poisson processes. Poisson processes are fundamental in probability theory and statistics, on their own, also via extensions and further constructions. Here we establish existence and basic properties, via limits of simpler processes and the convergence machinery for $D[0, 1]$.

(a) For each given m , let $Y_{m,i}$ be i.i.d. $\text{Pois}(\lambda/m)$, and consider $Z_m(t) = \sum_{i/m \leq t} Y_{m,i}$ for $t \geq 0$, a well-defined random process in $D[0, 1]$. Show that Z_m has independent increments, that $Z_m(t) \sim \text{Pois}(\lambda[mt]/m)$, and that $Z_m(t) - Z_m(s) \rightarrow_d \text{Pois}(\lambda(t-s))$ for each cell $[s, t]$.

(b) From the simple expression $E \{Z_m(t) - Z_m(s)\} = \lambda([mt] - [ms])/m$, verify that Z_m

is tight, via criteria in Ex. 10.21. Deduce that there is convergence to a well-defined process $Z = \{Z(t) : t \geq 0\}$, with independent increments $Z(t) - Z(s) \sim \text{Pois}(\lambda(t-s))$.

(c) With an infinite sequence $Y_{m,i} \sim \text{Pois}(\lambda/m)$, we may define $Z_m(t) = \sum_{i/m \leq t} Y_{m,i}$ for all $t \geq 0$, i.e. beyond the $[0, 1]$ interval. Argue that the $Z_m \rightarrow_d Z$ convergence above holds for each $D[0, \tau]$, for any finite τ , and hence that $Z = \{Z(t) : t \geq 0\}$ may be seen as a well-defined Poisson process for the full half-line. For Z such a Poisson process, let T_1 be the time for the first event. Show that $T_1 \sim \text{Expo}(\lambda)$. More generally, let $T_1 < T_2 < \dots$ be the random time-points at which the Poisson events occur. Representing these via waiting times $W_i = T_i - T_{i-1}$, show that $W_1, W_2, \dots$ are i.i.d. $\text{Expo}(\lambda)$.

(d) Such processes may also be interpreted, constructed, and analysed via transition probabilities. Conditioning on what has happened up to time t , then considering what will happen over $[t, t+h]$, show that

$$\begin{aligned}\Pr(Z(t+h) = z \mid Z(t) = z) &= 1 - \lambda h + o(h), \\ \Pr(Z(t+h) = z+1 \mid Z(t) = z) &= \lambda h + o(h), \\ \Pr(Z(t+h) \geq z+2 \mid Z(t) = z) &= o(h).\end{aligned}$$

(e) Assume that Z is a nondecreasing process in $\{0, 1, \dots\}$ with the property above, for some λ . Show that the waiting time W_1 to the first event must be $\text{Expo}(\lambda)$, and use this to establish that Z must be a Poisson process. We learn that the infinitesimal description of the previous point fully characterises these processes.

(f) For t inside a wider time window $[0, T]$, show that $Z(t) \mid \{Z(T) = n\}$ is a binomial $(n, t/T)$. More generally, if $0 = t_0 < t_1 < \dots < t_k = T$, and N_j the number of events inside $(t_{j-1}, t_j]$, show that if $Z(T) = n$ is given, then $(N_1, \dots, N_k)$ is multinomial $(n, p_1, \dots, p_k)$, with $p_j = (t_j - t_{j-1})/T$.

(g) Given $Z(T) = n$, as in the previous points, with Poisson events occurring at time points $T_1 < \dots < T_n$, show that the distribution of $(T_1, \dots, T_n)$ is identical to that of $(U_{(1)}, \dots, U_{(n)})$, the order statistics for a uniformly distributed sample on $[0, T]$.

Ex. 10.23 *Nonhomogeneous Poisson processes and Bernoulli processes.* We know that many small probabilities give a Poisson, see e.g. Ex. 2.8. Here we exhibit process generalisations of such findings, leading also to nonhomogeneous Poisson processes.

(a) For $Y \sim \text{binom}(n, p)$, show that its Laplace transformation is $L_n(u) = \mathbb{E} \exp(-uY) = \{\exp(-u)p + 1 - p\}^n$. Use this to show that with n increasing and p decreasing in such a way that $np \rightarrow \lambda$, then $L_n(u)$ converges to $\exp[-\lambda\{1 - \exp(-u)\}]$, which is the Laplace transform of a $\text{Pois}(\lambda)$.

(b) Now study a Bernoulli event process, of the form $Z_m(t) = \sum_{i/m \leq t} B_{m,i}$, with independent Bernoulli components $B_{m,i} \sim \text{binom}(1, \lambda(i/m)(1/m))$. Here $\lambda(s)$ is a function, perhaps evolving over time, with cumulative intensity function $\Lambda(t) = \int_0^t \lambda(s) ds$. Show that Z_m has independent increments, and that Laplace transform converges,

$$\mathbb{E} \exp\{-uZ_m(t)\} = \prod_{i/m \leq t} [1 - \lambda(i/m)\{1 - \exp(-u)\}] \rightarrow \exp[-\Lambda(t)\{1 - \exp(-u)\}].$$

(c) Demonstrate also that the Z_m sequence is tight. So there is convergence of the fine-grid cumulative Bernoulli process, with small probabilities, to a full process with $Z(t) - Z(s) \sim \text{Pois}(\int_s^t \lambda(u) du)$.

Ex. 10.24 *Mapping a process to another process.* We've seen in Ex. 10.17 that if $X_n \rightarrow_d X$, in function spaces $C[0, 1]$ or $D[0, 1]$, and if g is continuous mapping to the real line,

check with
Story #77
where this is
used

then $g(X_n) \rightarrow_d g(X)$. We are also taking an interest in mappings from a function to another function.

(a) Assume a mapping $g: C[0, 1] \rightarrow C[0, 1]$ is continuous. Show that if $X_n \rightarrow_d X$ in $C[0, 1]$, then indeed the new random process $g(X_n)$ tends to the random limit process $g(X)$.

(b) Assume $X_n \rightarrow_d W$ in $C[0, 1]$, and consider the new random process $Y_n(t) = \exp\{X_n(t)\}$. Show that $Y_n \rightarrow_d Y$, with $Y(t) = \exp\{W(t)\}$. Note that each $Y(t)$ has a log-normal distribution. Find the mean and variance of $Y(t)$, and also the covariance between $Y(t_1)$ and $Y(t_2)$.

(c) Assume $X_n \rightarrow_d W$, and consider $Z_n(t) = X_n(t) - tX_n(1)$ for $t \in [0, 1]$. Show that the Z_n process converges in distribution to a well-defined Z zero-mean normal process, and find its covariance function.

The empirical distribution process and the Brownian bridge

the
e.c.d.f. process

Ex. 10.25 *The uniform empirical distribution process.* We have encountered the empirical distribution function in Ex. 3.9, including its use to form monitoring processes. Here we go further. Consider $U_1, U_2, \dots$ being i.i.d. from the uniform distribution. We study the empirical distribution process $G_n(t) = n^{-1} \sum_{i=1}^n I(U_i \leq t)$ for $t \in [0, 1]$. We have seen in the exercise pointed to that the process $Z_n(t) = \sqrt{n}\{G_n(t) - t\}$ has the property that $Z_n(t) \rightarrow_d N(0, t(1-t))$, for each t . Here we establish that there is a well-defined process limit, actually $Z_n \rightarrow_d W^0$, the Brownian bridge. Note indeed that Z_n starts and ends at zero.

(a) From the binomial CLT, show that $Z_n(t) \rightarrow_d N(0, t(1-t))$. Show also that for $s < t$, $(Z_n(s), Z_n(t)) \rightarrow_d (A, B)$, say, a binormal zero-mean with variances $s(1-s)$ and $t(1-t)$, and covariance $s(1-t)$.

(b) As an interlude, turning out to be relevant in a minute, study a zero-mean binormal (X, Y) , with variances σ_1^2 and σ_2^2 and correlation ρ . Show that $E X^2 Y^2 = \sigma_1^2 \sigma_2^2 (1 + 2\rho^2) \leq 3\sigma_1^2 \sigma_2^2$. If now $(\hat{p}, \hat{q}) = (A/n, B/n)$, in a setup where (A, B) is trinomial (n, p, q) , consider $K_n(p, q) = E |\sqrt{n}(\hat{p} - p)\sqrt{n}(\hat{q} - q)|^2$. Show that $K_n(p, q) \rightarrow K(p, q) = E X^2 Y^2$, where (X, Y) is binormal, with $\sigma^2 = p(1-p)$, $\sigma_2^2 = q(1-q)$, and covariance $-pq$. Deduce that $K(p, q) \leq 3pq(1-p)(1-q) \leq 3pq$, which is what we need below.

(c) Now back to G_n and the empirical process Z_n . There is actually full convergence in distribution here, $Z_n \rightarrow_d W^0$ in $D[0, 1]$, to the Brownian bridge. Show this – which means demonstrating (i) finite-dimensional convergence and (ii) tightness. For the latter, use the above trinomial work to show that

$$E |Z_n(t) - Z_n(s)|^2 |Z_n(u) - Z_n(t)|^2 \rightarrow K(t-s, u-t) \leq 3(t-s)(u-t) \leq 3(u-s)^2.$$

and appeal to (10.3) in Ex. 10.19.

Ex. 10.26 *The empirical distribution process.* In generalisation of what we worked with in Ex. 10.25, consider i.i.d. variables $X_1, \dots, X_n$ from some continuous and strictly monotone c.d.f. F , with empirical c.d.f. $F_n(t) = n^{-1} \sum_{i=1}^n I(X_i \leq t)$. We shall establish a limit process for $Z_n^*(t) = \sqrt{n}\{F_n(t) - F(t)\}$.

(a) Write $X_i = F^{-1}(U_i)$, with the U_i being i.i.d. and uniform. Show that $F_n(t) = G_n(F(t))$, in terms of the empirical c.d.f. for the U_i . From this deduce that $Z_n^*(t) = \sqrt{n}\{G_n(F(t)) - F(t)\} = Z_n(F(t))$, with Z_n the uniform empirical c.d.f. process of Ex. 10.25.

(b) Show from this that $Z_n^*(t) \rightarrow_d W^0(F(t))$, both pointwise and as a process; the limit is time-transformed Brownian motion.

(c) So how much can $F_n(t)$ deviate from its mean function $F(t)$? One answer is to apply the max functional. Start with

used in Story
#83

$$\begin{aligned} D_n &= \sqrt{n} \max_t |F_n(t) - F(t)| \\ &= \sqrt{n} \max_u |G_n(u) - u| = \sqrt{n} \max_{i \leq n} \{|G_n(i/n-) - i/n|, |G_n(i/n) - i/n|\}, \end{aligned}$$

the maximum distance from $F_n(t)$ to $F(t)$, also equal to the maximum distance from $G_n(t)$ to t . Then deduce

$$D_n = \max_{0 \leq t \leq 1} |Z_n(t)| \rightarrow_d D = \max_{0 \leq t \leq 1} |W^0(t)|.$$

The distribution of this limit D is sometimes called the Kolmogorov–Smirnov distribution, and due to its importance has been tabulated since pre-computer days. It can easily be simulated, without knowing any formulae for its c.d.f. or density. We record here that 0.8276, 1.2238, 1.3580, 1.6271 are its 0.50, 0.90, 0.95, 0.99 quantiles. So $\max_t |G_n(t) - t| \leq 1.358/\sqrt{n}$, with probability close to 0.95. If you have data and compute D_n , with D_n bigger than 1.358, you might be skeptical about the assumption that your data are uniform.

though such
exist in the
literature

(d) Pretend that we have lost the Kolmogorov–Smirnov tables, and that we don’t have the time to derive its (rather complicated) distribution. Simulate 10^5 paths in your computer, and read off the 0.95 point, which should then be close to the 1.358 number.

Ex. 10.27 *The Kolmogorov–Smirnov test and band.* Results and insights from Ex. 10.26 can be used for purposes of testing and confidence bands. Suppose you have i.i.d. data $Y_1, \dots, Y_n$ and need to test the hypothesis H_0 that their distribution F is equal to some given F_0 , like the standard normal. Consider the empirical c.d.f. $F_n(t) = (1/n) \sum_{i=1}^n I(y_i \leq t)$. Under F_0 , we have established above that $Z_n = \sqrt{n}(F_n - F_0) \rightarrow_d Z = W^0(F_0(\cdot))$.

(a) Consider the test for $F = F_0$ consisting in rejecting if $F_0(t)$ for some part of the t domain is outside the Kolmogorov–Smirnov band. Show that this test has level 0.05 for large n . Show also that it is equivalent to rejecting if $D_n = \sqrt{n} \max_t |F_n(t) - F_0(t)| > 1.358$.

the Kolomo-
gorov–Smirnov
test

(b) We have learned above that $\sqrt{n} \max_t |F_n(t) - F(t)|$ tends in distribution to $D = \max_t |W^0(F(t))| = \max_s |W^0(s)|$. Explain that this implies

$$\Pr\{F(t) \in F_n(t) \pm 1.358/\sqrt{n}, \text{ for all } t\} \rightarrow 0.95.$$

This is the Kolmogorov–Smirnov simultaneous confidence band.

(c) Simulate $n = 100$ points from a standard normal, and construct a version of Figure 10.2, with the empirical process $Z_n = \sqrt{n}(F_n - F_0)$, left panel, and the confidence band, right panel.

Ex. 10.28 *Weighting the Kolmogorov–Smirnov test and band.* Consider as in Ex. 10.27 i.i.d. data $Y_1, \dots, Y_n$ from some F , with F_n the empirical c.d.f.

(a) With $v(s)$ a nonnegative bounded function on the unit interval, with finitely many discontinuities, show that $Z_n^*(t) = \sqrt{n}v(F_n(t))\{F_n(t) - F(t)\}$ tends to $Z^*(t) = v(F(t))W^0(F(t))$. From this explain that

$$D_n^* = \sqrt{n} \max_t v(F_n(t)) |F_n(t) - F(t)| \quad \text{has } D_n^* \rightarrow_d D^* = \max v(s) |W^0(s)|,$$

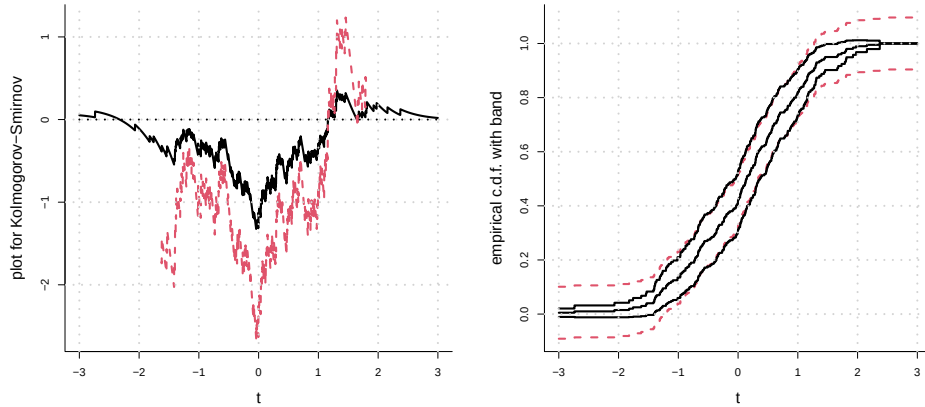


Figure 10.2: Left panel: plots of the Kolmogorov–Smirnov related processes $Z_n = \sqrt{n}(F_n - F)$ and its weighted version Z_n^* , as in Ex. 10.27 and 10.28, to check if 200 data points from the standard normal look nonnormal or not. Here the $\max_t |Z_n(t)|$ and $\max |Z_n^*(t)|$ values are 1.324 and 2.652, both lower than the respective upper 0.05 points, 1.358 and 3.121, of their limit distributions under normality. Right panel: simultaneous 95 percent confidence bands for F , one with constant width valid for all t , the other of width proportional to the standard deviation, valid in the central 90 percent data range.

with the same limit obtained if $v(F_n(t))$ is replaced by $v(F(t))$.

(b) This invites using weighted versions of the Kolmogorov–Smirnov test and band, with $v(t)$ chosen to reflect wished-for features, like giving more importance to some data regions than others. Consider in particular $v(s) = 1/\{s(1-s)\}^{1/2}$ on say $[\varepsilon, 1-\varepsilon]$, with ε small and positive, and $v(s) = 0$ outside that interval. Show that

$$D_n^* = \sqrt{n} \max_{\varepsilon \leq F(t) \leq 1-\varepsilon} \frac{|F_n(t) - F(t)|}{[F(t)\{1-F(t)\}]^{1/2}} \quad \text{has } D_n^* \rightarrow_d D^* = \max_{\varepsilon \leq s \leq 1-\varepsilon} \frac{|W^0(s)|}{\{s(1-s)\}^{1/2}}.$$

the absolute maximum of the naturally normalised Brownian bridge, to make the variance constant and equal to one. This is a process and a distribution we have met in Ex. 10.18, along with a few quantiles based on simulations, for e.g. the $[0.05, 0.95]$ interval. Explain how this may be used to test if the data come from some given $F = F_0$, say the standard normal.

(c) With z_ε the 0.95 quantile of the D^* distribution above, explain that

$$\Pr_F\{F(t) \in F_n(t) \pm z_\varepsilon [F_n(t)\{1-F_n(t)\}]^{1/2}/\sqrt{n}, \text{ for all } t \in A_n\} \rightarrow 0.95,$$

where A_n contains those t for which $\varepsilon \leq F_n(t) \leq 1-\varepsilon$. We have hence constructed a simultaneous confidence band, with approximate 95 percent level, proportional to the standard deviation of $F_n(t)$, as opposed to the fixed bandwidth construction of Ex. 10.27. We found such z_ε numbers in Ex. 10.18, so for $\varepsilon = 0.05$ we use 3.121, etc.

(d) Now redo the Figure 10.2 things, with this weighted version of the Kolmogorov–Smirnov.

Delta method for processes, order statistics processes

Ex. 10.29 Stochastic integrals. Working with empirical processes we sooner or later encounter stochastic integrals, of the type $A = \int_0^1 h(s) dW(s)$ and $B = \int_0^1 h(s) dW^0(s)$,

for Brownian motion and the Brownian bridge, with h some given function.

(a) In Riemann integral fashion, write $A'_m = \sum_{i=1}^m h(i/m)\{W(i/m) - W((i-1)/m)\}$, show that it is zero-mean normal with variance σ_m^2 converging to $\sigma^2 = \int_0^1 h(s)^2 ds$. Similarly, via $W^0(t) = W(t) - tW(1)$, write

$$\begin{aligned} B'_m &= \sum_{i=1}^m h(i/m)\{W^0(i/m) - W^0((i-1)/m)\} \\ &= \sum_{i=1}^m h(i/m)\{W(i/m) - W((i-1)/m)\} - (1/m) \sum_{i=1}^n h(i/m) W(1), \end{aligned}$$

and find that its distribution tends to $N(0, \tau^2)$, with $\tau^2 = \int_0^1 h(s)^2 ds - h_0^2$, with $h_0 = \int_0^1 h(s) ds$. Argue that this gives clear meanings to A and B .

(b) Consider $X_1, \dots, X_n$ i.i.d. from the uniform, i.e. with $F(t) = t$, for which we have established $Z_n = \sqrt{n}(F_n - F) \rightarrow_d W^0$. We then study $B_n = \int_0^1 h(s) dZ_n(s)$, which ought to converge in distribution to $B = \int_0^1 h(s) dW^0(s)$. Since the structure of this Z_n is simple, explain that $B_n = \sqrt{n}(\bar{h}_n - h_0)$, with $\bar{h}_n = n^{-1} \sum_{i=1}^n h(X_i)$. Verify the $B_n \rightarrow_d B$ claim from the CLT.

(c) The point is however to learn the $B_n \rightarrow_d B$ via process convergence, as this will be fruitful in more intricate situations. This is not immediate; explain that the functional taking $x \in C[0, 1]$ to $\int_0^1 h(x) dx(s)$ is not continuous; functions x and y may be uniformly close, but not their small increments. Show first that $B_n \rightarrow_d B$ for any ‘simple’ h , with constant values over a finite set of intervals. Then argue, perhaps via Ex. 2.30, that we also have $B_n \rightarrow_d B$ for any continuous h , via approximating simple functions.

(d) Suppose G_n estimates the function G and that the empirical process $Z_n = \sqrt{n}(G_n - G)$ tends to Z in $D[0, 1]$, with the limit process being zero-mean Gaussian. Explain first that $B_n = \int_0^1 h dZ_n$ tends to $B = \int_0^1 h dZ$ for each simple h . What is further required, to have $B_n \rightarrow_d B$ also for any continuous function h ?

Ex. 10.30 *The delta method for processes.* We have developed and used versions of the delta method multiple times in earlier chapters; in its simplest form, if $\sqrt{n}(X_n - a) \rightarrow_d V$, then $\sqrt{n}\{g(X_n) - g(a)\} \rightarrow_d g'(a)V$. Here are versions suited for empirical processes.

(a) Suppose $Z_n = \sqrt{n}(A_n - A) \rightarrow_d Z$, in $D[0, 1]$, with $A_n(t)$ and $A(t)$ being a process and a function, and with the limit Z being in $C[0, 1]$ a.s. Show that $\max_t |A_n(t) - A(t)| \rightarrow_{\text{pr}} 0$. Assume g is a function with a smooth and bounded derivative on $[0, 1]$. Show that the process Z_n^* , defined by $Z_n^*(t) = \sqrt{n}\{g(A_n(t)) - g(A(t))\}$, converges to Z^* , defined by $Z^*(t) = g'(A(t))Z(t)$. The same arguments should show that there is process convergence $Z_n^* \rightarrow_d Z^*$ in the function space $D[\varepsilon, 1 - \varepsilon]$ in cases where g has a smooth and bounded derivative only on the subinterval $[\varepsilon, 1 - \varepsilon]$.

(b) Consider any smooth function over finitely many indexes, i.e. the type $\hat{\theta} = T(A_n) = h(A_n(t_1), \dots, A_n(t_k))$, estimating $\theta = T(A) = h(A(t_1), \dots, A(t_k))$. Via finite-dimensional convergence, explain that we have large-sample control here, with

$$\sqrt{n}\{T(A_n) - T(A)\} \rightarrow_d c_1 Z(t_1) + \dots + c_k Z(t_k), \quad \text{with } c_j = \partial h(A) / \partial a_j.$$

With Z a Gaussian process, the limit is a normal, with variance determined by the partial derivatives and the covariance function.

(c) To illustrate the above, consider $X_1, \dots, X_n$ i.i.d., for which we do have the $\sqrt{n}(F_n - F) \rightarrow_d Z = W^0(F(\cdot))$ result, from Ex. 10.26. Now use $g(u) = 2 \arcsin(u^{1/2})$, and show

$$\sqrt{n}\{g(F_n(t)) - g(F(t))\} \rightarrow_d Z^*(t),$$

a Gaussian zero-mean process with constant variance 1. How can you use this to test that the data are standard normal?

(d) For the situation above, and with $t_0 < t_1 < \dots < t_k$, consider $D_{n,j} = F_n(t_j) - F_n(t_{j-1})$ for $j = 1, \dots, k$. Give the limit distribution for $(D_{n,1}, \dots, D_{n,k})$, properly normalised. The point here is partly that with basic knowledge of $\sqrt{n}(F_n - F)$, ‘everything else follows’ regarding large-sample approximations. This is made more broadly clear when we also enter the terrain of quantiles and inverses below.

Ex. 10.31 *The order statistics process.* Suppose $V_1, V_2, \dots$ are i.i.d. with mean 1 and variance 1, and construct from these the bridge process, starting and ending at zero,

$$Z_n(t) = \sqrt{n} \left(\sum_{i=1}^{[nt]} V_i / \sum_{i=1}^n V_i - t \right) \quad \text{for } t \in [0, 1].$$

(a) Studying first $A_n(t) = n^{-1/2} \sum_{i \leq [nt]} (V_i - 1)$, show via Donsker’s theorem that $B_n \rightarrow_d W^0$, the Brownian bridge, where $B_n(t) = A_n(t) - tA_n(1)$. Show that also $Z_n \rightarrow_d W^0$, via

$$Z_n(t) = \sqrt{n} \left\{ n^{-1} \sum_{i=1}^{[nt]} (V_i - 1) - tn^{-1} \sum_{i=1}^n (V_i - 1) + ([nt]/n - t) \right\} / n^{-1} \sum_{i=1}^n V_i.$$

(b) Consider $U_1, \dots, U_n$ being i.i.d. from the uniform distribution, with order statistics $U_{(1)} < \dots < U_{(n)}$. Use the representation of the $U_{([nq])}$ process from Ex. 3.17, in concert with results above, to demonstrate that the uniform order statistics process $O_n(q) = \sqrt{n}(U_{([nq])} - q)$ converges to the Brownian bridge.

(c) (xx to be done, linear combinations of order statistics, check the Stigler paper. we’re content with a simple case.xx) Consider $X_n = \int_0^1 h(s) dQ_n(s)$, which tends to $c = \int_0^1 h(s) ds$. or $X_n = \sum_{i=1}^n J(i/(n+1))U_{(i)}$, tending to $c = \int_0^1 J(s)s ds$. find limit for $\sqrt{n}(X_n - c)$.

(d) With observations $Y_1, \dots, Y_n$ from a smooth increasing c.d.f. F , with density f , consider the quantile process $R_n(q) = \sqrt{n}\{Q_n(q) - Q(q)\}$, with $Q_n(q) = Y_{([nq])}$ and $Q = F^{-1}$. Use Ex. 10.30 to show that $R_n \rightarrow_d R$, on each $D[\varepsilon, 1 - \varepsilon]$, where $R(q) = c(q)W^0(q)$, and $c(q) = (F^{-1})'(q) = 1/f(F^{-1}(q))$.

(e) (xx clean this. xx) Show that with i growing large with n such that $q_i = i/(n+1)$ is bounded away from 0 and 1, the event

$$Y_{(i)} \in F^{-1}(q_i) \pm 1.96 c(q_i) \{q_i(1 - q_i)\}^{1/2} / \sqrt{n}$$

will have probability tending to 0.95. Argue in particular from this that a plot of $(F^{-1}(q_i), Y_{(i)})$ should be approximately linear. This, along with variations, is called a q-q or quantile-quantile plot.

(f) In the situation just described, show with a positive weight function $w(t)$ that

$$H_n = \sqrt{n} \max_{\varepsilon \leq q \leq 1-\varepsilon} w(q) |Y_{([nq])} - F^{-1}(q)| \rightarrow_d H = \max_{\varepsilon \leq t \leq 1-\varepsilon} w(q)c(q) |W^0(q)|.$$

(g) (xx an application or illustration. both the QQ plot thing with pointwise band, and functionals of the type $V_n = \int_0^1 w(q) dR_n(q) \rightarrow_d V = \int_0^1 w(q) dR(q)$, almost linear combinations of order statistics. can check for normality via QQ plots. see nils com5* for birthweights for nonsmoking and smoking mothers. xx)

Ex. 10.32 Inverse processes. In Ex. 10.31 we found clean large-sample results for the quantile process, via a representation of order statistics. Consider in broader generality a setup with a non-decreasing $A_n(t)$ tending to some smooth increasing $A(t)$ on the unit interval, in such a fashion that $Z_n = \sqrt{n}(A_n - A) \rightarrow_d Z$, a zero-mean Gaussian process. What happens to their inverses?

(a) Let $A_n^{-1}(u) = \inf\{t: A_n(t) \geq u\}$, and similarly for $A^{-1}(u)$. We work through some details below to establish the general result that

$$Z_n^*(u) = \sqrt{n}\{A_n^{-1}(u) - A^{-1}(u)\} \rightarrow_d Z^*(u), \quad \text{with } Z^*(u) = -Z(t)/A'(t),$$

where $t = A^{-1}(u)$. Apply this result for the quantile process studied in Ex. 10.31, and derive the result there again: from $Z_n = \sqrt{n}(F_n - F) \rightarrow_d Z$ to $Z_n^* = \sqrt{n}(Q_n - Q) \rightarrow_d Z^*$, writing $Q_n = F_n^{-1}$ and $Q = F^{-1}$. The convergence holds for each space $D[\varepsilon, 1 - \varepsilon]$. Explain that the same limit ensues also for other versions $\tilde{F}_n$ than the strict empirical c.d.f. F_n , as long as $\sqrt{n}(\tilde{F}_n - F_n) \rightarrow_{\text{pr}} 0$.

(b) Why and how does the $-1/A'$ factor appear: convince yourself via Taylor expansion that at $t_0 = A^{-1}(u)$,

$$0 = A_n(A_n^{-1}(u)) - u = \{A(A_n^{-1}(u)) - A(t_0)\} + \{A_n(A_n^{-1}(u)) - A(A_n^{-1}(u))\}.$$

Then divide by $A'(t_0)$ and multiply with $\sqrt{n}$. Rearrange to get $Z_n^*(u) \approx_d -Z_n(t_0)/A'(t_0)$.

(c) Via somewhat different arguments, and in some more detail, fix the u , and write $t_0 = A^{-1}(u)$, i.e. $A(t_0) = u$. Show that $Z_n^*(u) \leq w$ is equivalent to $u \leq A_n(t_0 + w/\sqrt{n})$, and then to

$$\sqrt{n}\{u - A(t_0 + w/\sqrt{n})\} \leq Z_n(t_0 + w/\sqrt{n}).$$

Then argue that the l.h.s. is essentially $-A'(t_0)w$ and that the r.h.s. is close enough to $Z_n(t_0)$ to have the same limit. Explain that these details lead, again, to $Z_n^*(u) \approx_d -Z_n(t_0)/A'(t_0)$, and in the limit $Z_n^*(u) \rightarrow_d -Z(A^{-1}(u))/A'(A^{-1}(u))$.

(d) A useful exercise is to go through the steps here to give another proof than those used before (from the Scheffé lemma as in Ex. 3.20 and the order statistics representation of Ex. 3.21) of the empirical median result $\sqrt{n}(M_n - \mu) \rightarrow_d N(0, \frac{1}{4}/f(\mu)^2)$.

Ex. 10.33 The top ten percent do fifty percent of the total job. Statements are sometimes made along the lines of ‘the top ten percent in a workforce carry out fifty percent of the total job’ (with variations). There are settings where this can be seen to hold. Suppose efforts of i.i.d. workers can be seen as $X_1, \dots, X_n$ from some distribution on the halfline, with density f and c.d.f. F . Then consider the ratio $R_n = \sum_{i \geq [0.90n]} X_{(i)} / \sum_{i=1}^n X_i$, in terms of the order statistics $X_{(i)}$.

(a) First assume that the X_i are from the uniform distribution on the unit interval. Show that $R_n \rightarrow_{\text{pr}} 0.19$.

(b) For the general case, show that $n^{-1} \sum_{i \geq [0.90n]} X_{(i)} \rightarrow_{\text{pr}} \int_{0.90}^1 F^{-1}(u) du$, and that

$$R_n \rightarrow_{\text{pr}} \rho(0.90) = \frac{A(0.90)}{A(0)}, \quad \text{with } A(q) = \int_q^1 F^{-1}(u) du = \int_{F^{-1}(q)}^\infty x f(x) dx.$$

(c) Examine the Beta($a, 1$) distribution, with $F(x) = x^a$ and $F^{-1}(u) = u^{1/a}$. Find an expression for $\rho(0.90)$, and show that for large n , the top 10 percent in a sample amount to 50 percent of the total, for $a_0 = 0.179$. Carry out some simulations to illustrate this phenomenon.

(d) On any given day the efforts of the n i.i.d. workers might not correspond precisely to the 10 percent do 50 percent formula. Find the limit distribution of $\sqrt{n}\{R_n - \rho(0.90)\}$.

(e) Work through the following variation, with the output from i.i.d. workers following the Pareto c.d.f. $F(t) = 1 - 1/t^\theta$ for $t \geq 1$, see Ex. 1.16. Find the quantile function $F^{-1}(q) = \{1/(1 - q)\}^{1/\theta}$, followed by

$$\rho(q) = \frac{A(q)}{A(0)} = \frac{\int_q^1 \{1/(1-u)\}^{1/\theta} du}{\int_0^1 \{1/(1-u)\}^{1/\theta} du} = (1-q)^{1-1/\theta}$$

for $\theta > 1$. Check that $\theta = 1.161$ makes $\rho(0.80) = 0.80$. What happens in a society where the wages follow a Pareto with $\theta = 1.05$?

Ex. 10.34 *The correlation between mean and median.* Consider $X_1, \dots, X_n$ i.i.d. from a distribution with positive density f and c.d.f. F on the real line. With $\bar{X}_n$ the mean and M_n the median, and writing ξ and μ for the theoretical counterparts, we have seen in Ch. 2 that

$$\sqrt{n}(\bar{X}_n - \xi) \rightarrow_d A \sim N(0, \sigma^2), \quad \sqrt{n}(M_n - \mu) \rightarrow_d B \sim N(0, \tau^2), \quad \text{with } \tau = \frac{1}{2}/f(F^{-1}(\mu)),$$

but what is the correlation, under different assumptions?

(a) Start out assuming that the data are normal. Use the independence of $X_i - \bar{X}_n$ and $\bar{X}_n$ to explain that $M_n - \bar{X}_n$ and $\bar{X}_n$ are independent. Deduce from this that $\text{cov}(M_n, \bar{X}_n) = \sigma^2/n$. Writing τ_n^2/n for the variance of M_n , show that the correlation must be $\rho_n = \text{corr}(M_n, \bar{X}_n) = \sigma/\tau_n$. With the τ_n tending to the τ above, show that the limiting correlation is $(2/\pi)^{1/2} = 0.7979$. (xx footnote: we used $X_i - \bar{X}_n$ and $\bar{X}_n$ being independent. when does this hold? assume symmetry. use mgf, show $M(t) = M(t/2)^4$. but this implies normality. xx)

(b) (xx to fix. xx) $\sqrt{n}\{F_n^{-1}(s) - F^{-1}(s)\} \rightarrow_d q(s)W^0(s)$, with $q(s) = (F^{-1})'(s) = 1/f(F^{-1}(s))$. So $B = q(\frac{1}{2})W^0(\frac{1}{2})$, and $A = \int_0^1 q(s)W^0(s) ds = \int W^0(F(x)) dx$. Work with

$$E AB = q(\tfrac{1}{2}) \left[\int_{-\infty}^{\mu} F(x)(1 - \tfrac{1}{2}) dx + \int_{\mu}^{\infty} \tfrac{1}{2} \{1 - F(x)\} dx \right].$$

Under symmetry around zero, show

$$\rho = 2 \int_0^{\infty} x f(x) dx / \left\{ 2 \int_0^{\infty} x^2 f(x) dx \right\}^{1/2}.$$

can check how this behaves for t_ν , converging to 0.7979. couple all of this with criterion for when median is better than mean for estimating location. (xx notes, could go to solutions. point is to look at ρ_ν . xx)

$$I_\nu = \int_0^{\infty} x f_\nu(x) dx = \frac{\nu^{1/2}}{\pi^{1/2}} \frac{\Gamma(\frac{1}{2}\nu + \frac{1}{2})}{(\nu - 1)\Gamma(\frac{1}{2}\nu)}, \quad J_\nu = \int_0^{\infty} x^2 f_\nu(x) dx = \frac{1}{2} \frac{\nu}{\nu - 2}.$$

Ex. 10.35 *Nonparametric bootstrapping.* Suppose $X_1, \dots, X_n$ are i.i.d. from some F , and that inference is needed for some $\theta = T(F)$. We have established in Ex. 10.26 that

the Pareto
80-20 rule: 80%
consequences
from 20% of
what goes on

$Z_n = \sqrt{n}(F_n - F) \rightarrow_d Z = W^0(F(\cdot))$, and also, via Ex. 10.30 and 10.32, that ‘everything for large n flows from this’. Via $Z_n \rightarrow_d Z$ we have basically control over the limit distribution of $\sqrt{n}\{T(F_n) - T(F)\}$, for large classes of T , including all smooth functions of all finite collections of $F(t_j)$ and $F^{-1}(q_k)$.

(a) So what to do, for the given dataset, apart from computing $\hat{\theta} = T(F_n)$? If we can invent a sister process F_n^* to F_n , in our computers, with the same behaviour around F_n as has F_n around F , as data volume increases, we’re in business: argue that provided

$$\begin{aligned} Z_n &= \sqrt{n}(F_n - F) \\ Z_n^* &= \sqrt{n}(F_n^* - F_n) \mid \text{data} \end{aligned} \quad \text{have identical large-sample behaviour,} \quad (10.6)$$

then also

$$\begin{aligned} B_n &= \sqrt{n}\{T(F_n) - T(F)\} \\ B_n^* &= \sqrt{n}\{T(F_n^*) - T(F_n)\} \mid \text{data} \end{aligned} \quad \text{have identical large-sample behaviour,} \quad (10.7)$$

for all these sufficiently smooth functionals $\theta = T(F)$. Explain that with $[a_n, b_n]$ covering say 95 percent of B_n^* , then this interval also covers B_n with the same approximate probability, and that $\Pr(\hat{\theta} - b_n/\sqrt{n} \leq \theta \leq \hat{\theta} - a_n/\sqrt{n}) \doteq 0.95$.

lo & behold

(b) It remains to invent such a F_n^* from the data, for which $Z_n^* \rightarrow_d Z$ as data volume increases. One simple and almost canonical scheme is that of Efron’s classic bootstrap. Perhaps $B = 10^4$ times, generate datasets $X_1^*, \dots, X_n^*$ sampled i.i.d. from the observed $(x_1, \dots, x_n)$, and let $F_n^*(t) = (1/B) \sum_{b=1}^B I(X_i^* \leq t)$. Show that $F_n^*(t)$ has mean $F_n(t)$ and variance $F_n(t)\{1 - F_n(t)\}$, and that indeed the full process $Z_n^* = \sqrt{n}(F_n^* - F_n)$, approximately a $W^0(F_n(\cdot))$, tends to the same limit Z , as long as $F_n \rightarrow F$ uniformly. Explain that this is secured by Glivenko–Cantelli, with probability one. Argue, via the above, that the procedure works, with $\hat{\theta}^* - \hat{\theta}$ having nearly the same distribution as $\hat{\theta} - \theta$, where $\hat{\theta}^* = T(F_n^*)$ can be sampled in your computer, for the given dataset.

bootstrapping:
revolutionary,
in 1979

see Ex. 2.60

(c) To illustrate, get hold of the **egypt-data** from Story #50, with $n = 141$ lifetimes, and consider $\theta = EY/\text{med}(Y)$, the mean-median ratio. Compute $\hat{\theta} = 30.69/25.00 = 1.23$. Use bootstrapping to find a 95 percent interval for θ (this is nonparametric, without any parametric model used). Construct and plot also a full bootstrap based CD $C(\theta)$, and its accompanying confidence curve $cc(\theta)$, verifying the following steps:

$$C(\theta) = \Pr_{\theta}(\hat{\theta} \geq \hat{\theta}_{\text{obs}}) = \Pr(B_n \geq \sqrt{n}(\hat{\theta}_{\text{obs}} - \theta)) \doteq 1 - K_n(\sqrt{n}(\hat{\theta}_{\text{obs}} - \theta)),$$

where K_n is the c.d.f. for the bootstrapped B_n^* . Note this is fairly automatic, senza explicit large-sample calculations. Argue that there *is* limiting normality, say $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \tau^2)$, and the CD is close to $\Phi(\sqrt{n}(\theta - \hat{\theta})/\hat{\tau})$, but have been saved the trouble of finding a formula for and estimating τ . Your code should also work with small changes for other parameters $\theta(F)$ than this ratio.

Monitoring processes for models with estimated parameters

Ex. 10.36 *Monitoring processes and Kolmogorov–Smirnov with estimated parameters.* Suppose again that $Y_1, \dots, Y_n$ are i.i.d. from some c.d.f. F , which potentially belongs to some given parametric class, with density $f(y, \theta)$ and c.d.f. $F(\cdot, \theta)$. How can we test that modelling assumption? Assuming the model holds, with θ_0 the true parameter, let $\hat{\theta}$ be the ML estimator, for which we know Ch. 5 that $\hat{\theta} - \theta_0 = J^{-1}U_n + o_{\text{pr}}(n^{-1/2})$, under regularity conditions spelled out in Ex. 5.10–5.11. Here J is the Fisher information matrix at θ_0 and $U_n = n^{-1} \sum_{i=1}^n u(Y_i, \theta_0)$, with $u(y, \theta)$ the score function.

(a) Show that

$$F(t, \hat{\theta}) \doteq F(t, \theta_0) + H(t, \theta_0)^t (\hat{\theta} - \theta_0) \doteq F(t, \theta_0) + H(t, \theta_0)^t J^{-1} U_n,$$

in which $H(t, \theta_0) = \partial F(t, \theta_0) / \partial \theta = \int_{-\infty}^t f(y, \theta_0) u(y, \theta_0) dy$. The natural monitoring process to examine is $Z_n(t) = \sqrt{n} \{F_n(t) - F(t, \hat{\theta})\}$. With $Z_{n,0}(t)$ the simpler $\sqrt{n} \{F_n(t) - F(t, \theta_0)\}$, explain that

$$Z_n(t) = Z_{n,0}(t) - H(t, \theta_0)^t J^{-1} \sqrt{n} U_n + o_{\text{pr}}(1).$$

Supplement the $Z_{n,0}$ results of Ex. 10.26 with the broader statement $(Z_{n,0}(t), U_n) \rightarrow_d (Z_0(t), U)$, in terms of joint convergence, where $Z_0(t) = W^0(F(t, \theta_0))$ is time-transformed Brownian bridge, $U \sim N_p(0, J)$, and

$$\text{cov}(Z_0(t), U) = \text{cov}\{I(Y_i \leq t), u(Y, \theta_0)\} = H(t, \theta_0).$$

Show from this that there is indeed process convergence

$$Z_n(t) \rightarrow_d Z(t) = Z_0(t) - H(t, \theta_0)^t J^{-1} U.$$

Show furthermore that the covariance function for the limit process is

$$\text{cov}\{Z(t_1), Z(t_2)\} = F(t_1, \theta_0)\{1 - F(t_2, \theta_0)\} - H(t_1, \theta_0)^t J^{-1} H(t_2, \theta_0)$$

for $t_1 \leq t_2$.

(b) For processes X_n, Y_n in $D[0, 1]$, suppose $Y_n \rightarrow_d Y$ and that $X_n(t) \rightarrow_{\text{pr}} x(t)$, uniformly in probability, a deterministic and continuous limit function. Show that there is process convergence $Y_n(X_n(t)) \rightarrow_d Y(x(t))$.

(c) We now carry out a convenient time transformation, bringing processes to $[0, 1]$. Transforming by $F^{-1}(v, \hat{\theta})$, show that

$$A_n(v) = Z_n(F^{-1}(v, \hat{\theta})) = \sqrt{n} \{G_n(v) - v\},$$

in terms of the empirical c.d.f. G_n of the $\hat{v}_i = F(y_i, \hat{\theta})$. Show that

$$A_n(v) \rightarrow_d A(v) = Z(F^{-1}(v, \theta_0)) = W^0(v) - H_0(v)^t J^{-1} U,$$

writing $H_0(v) = H(F^{-1}(v, \theta_0))$, and that this transformed limit process has covariance function

$$\text{cov}\{A(v_1), A(v_2)\} = v_1(1 - v_2) - H_0(v_1)^t J^{-1} H_0(v_2) \quad \text{for } v_1 \leq v_2.$$

Convince yourself that $\max_t |Z_n(t)| = \max_v |A_n(v)|$, so monitoring and analysis may be carried on the scale of A_n on $[0, 1]$.

(d) For an illustration, consider testing that data come from the $N(\xi, 1)$, for some ξ . In the notation above, show that $u(y, \xi) = y - \xi$, that $J = 1$, and that

$$Z_n(t) = \sqrt{n} \{F_n(t) - \Phi(t - \hat{\xi})\} \rightarrow_d Z(t) = Z_0(t) + \phi(t - \xi_0)U,$$

in which U is standard normal and $\text{cov}\{Z_0(t), U\} = -\phi(t - \xi_0)$.

Ex. 10.37 *Monitoring processes and tests for normality.* We now apply the general methods and results of Ex. 10.36 to the important case of testing whether data follow a normal distribution, i.e. with both parameters unknown.

(a) For the two-parameter $N(\xi, \sigma^2)$ model, show that $Z_n(t) = \sqrt{n}\{F_n(t) - \Phi((y - \hat{\xi})/\hat{\sigma})\} \rightarrow_d Z(t)$, with the limit process being

$$Z(t) = W^0(\Phi(s)) + \phi(s)U_1 + \phi(s)sU_2, \quad \text{writing } s = (t - \xi_0)/\sigma_0,$$

in which U_1, U_2 are independent normals with variances 1 and 2, and explain how these are connected to $W^0(\Phi(s))$. Deduce that

$$D_n = \sqrt{n} \max_t |F_n(t) - \Phi((t - \hat{\xi})/\hat{\sigma})| \rightarrow D = \max_s |Z^*(s)|,$$

where $Z^*(s)$ has a distribution independent of the parameters. Show indeed that with this notation, for $s_1 \leq s_2$,

$$\text{cov}\{Z^*(s_1), Z^*(s_2)\} = \Phi(s_1)\{1 - \Phi(s_2)\} - \phi(s_1)\phi(s_2) - \frac{1}{2}\phi(s_1)\phi(s_2)s_1s_2.$$

(b) We need upper quantiles of the D distribution, for testing normality with this type of Kolmogorov–Smirnov with estimated parameters. Explain that there are two ways of simulating close approximations to this D ; (i) to simulate from the actual D_n , for a high n , and (ii) to ignore the Z_n processes but simulate via the limit process. For the latter, show that the limit process $A(v) = Z^*(\Phi^{-1}(v))$ on $[0, 1]$ has covariance structure

$$\text{cov}\{A(v_1), A(v_2)\} = v_1(1 - v_2) - \phi(\Phi^{-1}(v_1))\phi(\Phi^{-1}(v_2))\{1 + \frac{1}{2}\Phi^{-1}(v_1)\Phi^{-1}(v_2)\}.$$

To simulate from the A we may use the general recipe of Ex. ??, for a fine grid of v in $[0, 1]$. Carry out versions of both schemes, to arrive at numbers close to the following. Here are quantiles at levels 0.50, 0.90, 0.95, 0.99 found from having simulated $2 \cdot 10^4$ times, with a fine grid 0, 0.001, ..., 0.999, 1 for the Gaussian limit process.

method (i) from Z_n					method (ii) from limit process with grid				
0.50	0.90	0.95	0.99		0.50	0.90	0.95	0.99	
0.822	1.202	1.360	1.613		0.805	1.205	1.341	1.614	with known parameters
0.632	0.854	0.949	1.107		0.638	0.872	0.949	1.101	xi unknown, sigma known
0.601	0.822	0.885	1.044		0.600	0.812	0.888	1.039	with both parameters estimated

(c) You may check Story #22 to see such normal monitoring processes Z_n , checking for normality of birthweight distributions. (xx to be done and polished. xx)

(d) The basic process result, that Z_n tends to a well-defined zero-mean Gaussian process Z , was used above to identify the limit distribution of the Kolmogorov–Smirnov type test statistic $\max_t |Z_n(t)|$. It may however also be used to form other tests. To illustrate this, consider (xx do this, with simulations, for absolute integral or CvM xx).

Ex. 10.38 *Monitoring processes and tests for exponentiality.* For another application of the general methodology of Ex. 10.36, consider the exponential model with c.d.f. $F(t, \theta) = F_0(\theta t)$, writing $F_0(t) = 1 - \exp(-t)$ for $t \geq 0$.

(a) With i.i.d. data $Y_1, \dots, Y_n$, show that the ML estimator is $\hat{\theta} = 1/\bar{Y}$ and that the Fisher information is $1/\theta^2$. Show that the results above lead to

$$Z_n(t) = \sqrt{n}\{F_n(t) - F_0(\hat{\theta}t)\} \rightarrow_d Z_0(t) + \exp(-\theta_0 t)\theta_0 t U_0,$$

where U_0 is standard normal, with $\text{cov}\{Z_0(t), \theta_0 U_0\} = -\exp(-\theta_0 t)t$. Write down the variance of the limit process. It is convenient here to transform this further, to the monitoring process

$$Z_n^*(t) = Z_n(t/\hat{\theta}) = \sqrt{n}\{G_n(t) - F_0(t)\}, \quad \text{with } G_n(t) = F_n(t/\hat{\theta}) = n^{-1} \sum_{i=1}^n I(\hat{\theta}Y_i \leq t),$$

which has a limit process free of the θ_0 parameter. Show indeed that $Z_n^*(t) \rightarrow_d Z^*(t) = Z(t/\theta_0)$, with

$$\text{cov}\{Z^*(t_1), Z^*(t_2)\} = F_0(t_1)\{1 - F_0(t_2)\} - t_1 t_2 \exp(-t_1 - t_2) \quad \text{for } t_1 \leq t_2.$$

With $D_n^* = \sqrt{n} \max_t |G_n(t) - F_0(t)|$, explain that $D_n^* \rightarrow_d D^* = \max_t |Z^*(t)|$.

(b) To carry out such a test in practice, one needs quantiles from the limit distribution D . As with the normality tests of Ex. 10.37 one may use method (i), choosing a high n and simulating the Z_n process, or method (ii), ignoring sampling exponentials and instead simulate from an approximation to the Gaussian limit process, for this purpose again using the general recipe of Ex. 10.7. Carry out such simulations, to find 0.50, 0.90, 0.95, 0.99 quantiles of the D^* distribution, as those given below. For that little table we have used $n = 1000$ for method (i). Note that D^* , associated with Z^* , is stochastically smaller than the D from the easier Kolmogorov–Smirnov with known parameter.

method (i) from Z_n				method (ii) from limit process with grid				
0.50	0.90	0.95	0.99	0.50	0.90	0.95	0.99	
0.822	1.202	1.359	1.612	0.811	1.201	1.336	1.610	for given parameter
0.695	0.980	1.075	1.296	0.686	0.977	1.075	1.289	with estimated parameter

Ex. 10.39 *Batting averages are bimodal, not unimodal.* Access the `batting-data` on the book site to get hold of $n = 140$ batting averages $y_1, \dots, y_n$, from US season-level Major League Baseball player data, with relatively small range $[0.213, 0.312]$.

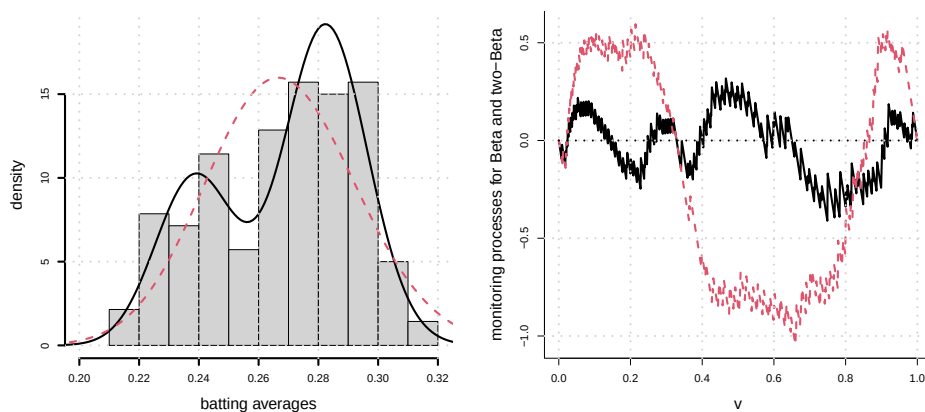


Figure 10.3: Analysis of 140 batting averages. Left panel: histogram, with fitted densities for the Beta and two-Beta models. Right panel: monitoring processes A_1 and A_2 for the two models; the one-Beta model does not pass muster.

high commotion
for staring at
data with
standard
deviation 0.025

(a) Fit the data to a $\text{Beta}(a, b)$ distribution, via likelihood. Check that the fitted density has mean and standard deviation $(0.268, 0.025)$. Compute also the monitoring process $A_1(v) = \sqrt{n}\{G_{n,1}(v) - v\}$, with the empirical c.d.f. for $\hat{v}_{i,1} = \text{Be}(y_i, \hat{a}, \hat{b})$, see Figure 10.3, right panel. Compute its $K_1 = \max |A_1(v)| = 1.047$. Via simulations of datasets from the estimated Beta, find the associated p-value, which is 0.012.

(b) Proceed then to a mixture of two Betas, with $f = (1-p)\text{Beta}(a_1, b_1) + p\text{Beta}(a_2, b_2)$. Fit this bigger model to the 140 datapoints; you should find $\hat{p} = 0.659$, with means and standard deviations being $(0.239, 0.013)$ and $(0.283, 0.014)$ for the ‘not so very good’

and ‘very good’ portions of batters. Compute and plot the monitoring process $A_2(v) = \sqrt{n}\{G_{n,2}(v) - v\}$ also for this bigger model, completing the right panel plot of Figure 10.3, right panel. Make also a version of the figure’s left panel. Comment on what is learned here.

(c) Use simulated datasets from the estimated Beta-mixture to plot such simulated monitoring plots $A_2^*(v)$ alongside the observed $A_2(v)$. Check that the value $K_2 = \max |A_2(v)| = 0.411$ is fully acceptable, i.e. the model appears adequate.

Ex. 10.40 *Monitoring processes for general regression models.* Consider a general regression setting for pairs (x_i, y_i) , with continuous outcomes Y_i following distributions with densities $f(y, x_i, \theta)$ and c.d.f.s $F(y, x_i, \theta)$. We do assume the model is correct, with θ_0 denoting the true parameter. Let $\hat{\theta}$ be the ML estimator. For moderate to large n , $\hat{v}_i = F(y_i, x_i, \hat{\theta})$ will be close to the real $v_i = F(y_i, x_i, \theta_0)$, and these again will be i.i.d. uniform under model conditions. Below let F_n and G_n the empirical c.d.f.s for these v_i and $\hat{v}_i$, respectively. The aim is to understand the behaviour of

$$V_n(t) = \sqrt{n}\{F_n(v) - v\} \quad \text{and} \quad W_n(t) = \sqrt{n}\{G_n(v) - v\} \quad \text{for } v \in [0, 1],$$

and their relation; limit process results for W_n may then be used for testing model conditions. A Kolmogorov–Smirnov type test would then be to compute $W_{n,\max} = \sqrt{n} \max_v |G_n(v) - v|$, and reject the model if this is bigger than an appropriate high quantile in the limit distribution $W_{\max} = \max_v |W(v)|$, with this limit W identified below.

Below we shall partly simplify notation, using $f(y, x)$ and $F(y, x)$ for $f(y, x, \theta_0)$ and $F(y, x, \theta_0)$, i.e. the density and c.d.f. at the true parameter, and similarly $H(y, x)$ for $\partial F(y, x, \theta_0)/\partial \theta$; in particular, $H(y_i, x_i) = \partial v_i/\partial \theta$. There will also be use for the ML representation result that $\hat{\theta} - \theta_0 = J_n^{-1}U_n = o_{\text{pr}}(n^{-1/2})$, with $U_n = n^{-1} \sum_{i=1}^n u(Y_i, x_i, \theta_0)$ the average score function at the true parameter and J_n the Fisher information matrix $-n^{-1} \sum_{i=1}^n E \partial^2 \log f(Y_i, x_i, \theta_0)/\partial \theta \partial \theta^t$; this is the regression version of what we learn in Ex. 5.11. It is also assumed below that the x_i are regular enough to ensure that averages $\bar{g}_n = n^{-1} \sum_{i=1}^n g(x_i)$ tend to well-defined limits $g_0 = \int g(x) dQ(x)$, means with respect to the the long-term covariate distribution.

(a) Intuition says G_n should be only a delta method step away from F_n , but analysis is not easy in that the $I(\hat{\varepsilon}_i \leq v)$ functions do not have simple derivatives. For this technical reason we introduce a smoothing operation. Let $k(u)$ be a kernel density function, having smooth derivatives, symmetric and supported on $[-1, 1]$, with c.d.f. K . Explain that $K_h(u) = K(u/h)$ tends to $K_0(u)$, which is 0 for $u < 0$, $\frac{1}{2}$ for $u = 0$, and 1 for $u > 0$, as h shrinks to zero. Explain that in $\sum_{i=1}^n \{K_h(v - \varepsilon_i) - K_0(v - \varepsilon_i)\}$, most of terms are zero, and that only those for which $|v - \varepsilon_i| \leq h$ make a difference. Explain further that the number of such terms is a binomial $(n, p_{n,h}(v))$, with $p_{n,h}(v) \doteq 2\phi(v)h$ for small h . Deduce from this that with

and only for
this technical
reason

$$F_{n,h}(v) = n^{-1} \sum_{i=1}^n K_h(v - \varepsilon_i) \quad \text{and} \quad G_{n,h}(v) = n^{-1} \sum_{i=1}^n K_h(v - \hat{\varepsilon}_i)$$

the kernel smoothed versions of F_n and G_n , we have

$$\sqrt{n}\{F_{n,h}(v) - F_n(v)\} \rightarrow_{\text{pr}} 0 \quad \text{and} \quad \sqrt{n}\{G_{n,h}(v) - G_n(v)\} \rightarrow_{\text{pr}} 0,$$

as long as $\sqrt{nh} \rightarrow 0$, as e.g. with $h \propto 1/n^c$ with $c \in (\frac{1}{2}, 1)$.

(b) Explain that we know from Ex. 10.26 that V_n tends to $V = W^0$, the Brownian bridge. Show that $V_{n,h}(t) = \sqrt{n}\{F_{n,h}(v) - v\}$ tends to the same limit V . Show further that

$$G_{n,h}(v) = F_{n,h}(v) - A_{n,h}(v)^t(\hat{\theta} - \theta_0) + o_{\text{pr}}(n^{-1/2}),$$

with $A_{n,h}(v) = n^{-1} \sum_{i=1}^n k_h(v_i - v)H(y_i, x_i)$. To understand the $A_{n,h}(v)$ random function, show that $k_h(v_i - v)H(y_i, x_i)$ has expected value

$$\int k_h(w - v)H(F^{-1}(w, x_i), x_i) dw = \int k(u)H(F^{-1}(v + hu, x_i), x_i) du,$$

and that this for small h leads to

$$\mathbb{E} A_{n,h}(v) = A_n(v) + O(h^2), \quad \text{with } A_n(v) = n^{-1} \sum_{i=1}^n H(F^{-1}(v, x_i), x_i).$$

Explain that the $A_n(v)$ tends in probability to a well-defined limit function $A(v)$.

(c) Establish that

$$W_{n,h}(v) = \sqrt{n}\{G_{n,h}(t) - v\} = V_{n,h}(v) - A_{n,h}(v)\sqrt{n}J_n^{-1}U_n + o_{\text{pr}}(1)$$

with $A_{n,h}(v) \rightarrow_{\text{pr}} A(v)$. Furthermore,

$$\text{cov}\{I(v_i \leq v), u(Y_i, x_i, \theta_0)\} = \int_{-\infty}^{F^{-1}(v, x_i)} f(y, x_i)u(y, x_i) dy = H(F^{-1}(v, x_i), x_i).$$

From this establish that $\text{cov}\{F_n(v), U_n\} = n^{-1}A_n(v)$. With $\sqrt{n}U_n \rightarrow_d U \sim N_p(0, J)$, conclude that

$$W_n(v) \rightarrow_d V(v) - A(v)^t J^{-1}U, \quad \text{where } \text{cov}\{V(v), U\} = A(v).$$

For this limit process, establish that

$$\text{cov}\{W(v_1), W(v_2)\} = v_1(1 - v_2) - A(v_1)^t J^{-1}A(v_2) \quad \text{for } v_1 \leq v_2.$$

(d) Show that W can be represented as the Brownian bridge V conditional on $U = 0$. The more complicated the model, the more constraints, and the smaller are the bridges in size.

Ex. 10.41 *Monitoring processes for linear regression.* Here we apply the general results developed in Ex. 10.40 to the classic linear regression model, with y given x having c.d.f. $F(y, x, \beta, \sigma) = \Phi((y - x^t \beta)/\sigma)$, with ML estimators $\hat{\beta}$ and $\hat{\sigma}$ as per Ex. 3.33–3.34. We compute the estimated residuals $\hat{r}_i = (y_i - x_i^t \hat{\beta})/\hat{\sigma}$ and from these $\hat{v}_i = \Phi(\hat{r}_i)$. For their empirical c.d.f. $G_n(v) = n^{-1} \sum_{i=1}^n I(\hat{v}_i \leq v)$ we shall identify the limit for the model monitoring process, $W_n(v) = \sqrt{n}\{G_n(v) - v\}$, under model conditions. (xx note. assume x_i has a 1, intercept, $\bar{x}^t \Sigma_n^{-1} \bar{x} = 1$. can also use $F_{n-p}(\hat{r}_i)$, with the c.d.f. for the t_{n-p} , with the same limit. xx)

(a) (xx edit and clean; will need some of this. xx) With $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t$, write the hat matrix of Ex. 3.34 as $H = X(X^t X)^{-1} X^t = n^{-1} M$, with $M = X \Sigma_n^{-1} X^t$. Explain that $\hat{\varepsilon} = Y - X \hat{\beta} = (I - H)Y \sim N_n(0, I - n^{-1} M)$; in particular, $\hat{\varepsilon}_i \sim N(0, 1 - n^{-1} m_{i,i})$. Show that $n^{-2} \sum_{i,j} m_{i,j} = 1$, the average value of the M matrix. For one more useful fact here, let as usual $\bar{x} = n^{-1} \sum_{i=1}^n x_i$, the first component of which being 1. Show that $\bar{x}^t \Sigma_n^{-1} \bar{x} = 1$. Below we shall assume that the $m_{i,j}$ are essentially stable in size, in the sense that $\max_{i,j} m_{i,j}/n \rightarrow 0$.

(b) Show that $F^{-1}(v, x_i) = x_i^t \beta + \sigma \Phi^{-1}(v)$, and that $H(y, x)$ has components $-(1/\sigma)x\phi(\varepsilon)$ and $-(1/\sigma)\varepsilon\phi(\varepsilon)$, in terms of $\varepsilon = (y - x^t \beta)/\sigma$. With notation as in the general exercise pointed to, show that

$$H(F^{-1}(v, x), x) = -(1/\sigma) \begin{pmatrix} \phi(\Phi^{-1}(v))x \\ \phi(\Phi^{-1}(v))\Phi^{-1}(v) \end{pmatrix},$$

and that this leads to

$$A_n(v) = n^{-1} \sum_{i=1}^n H(F^{-1}(v, x_i), x_i) = -(1/\sigma)\phi(\Phi^{-1}(v)) \begin{pmatrix} \bar{x} \\ \Phi^{-1}(v) \end{pmatrix}.$$

Use this to establish that $W_n \rightarrow_d W$, a zero-mean Gaussian process with covariance function

$$v_1(1 - v_2) - \phi(\Phi^{-1}(v_1))\phi(\Phi^{-1}(v_2))\{1 + \frac{1}{2}\Phi^{-1}(v_1)\Phi^{-1}(v_2)\} \quad \text{for } v_1 \leq v_2.$$

(xx should now simulate these from general recipe, yielding upper quantiles for checking regression assumptions. xx)

(c) (xx another illustration: exponential regression, to be polished. perhaps we have a story with such data. xx) Consider the setup with $Y_i \sim \text{Expo}(\theta_i)$, with $\theta_i = \exp(x_i^t \beta)$. With $F_0(u) = 1 - \exp(-u)$ and $f_0(u) = \exp(-u)$ the c.d.f. and density for the unit exponential, show that $F(y, x_i) = F_0(\theta_i y)$, $f(y, x_i) = f_0(\theta_i y)\theta_i$, $v_i = F_0(\theta_i y_i)$, $H(y, x_i) = f_0(\theta_i y)\theta_i y x_i$, $F^{-1}(v, x_i) = \theta_i^{-1} F_0^{-1}(v)$, $H(F^{-1}(v, x_i), x_i) = f_0(F_0^{-1}(v))F_0^{-1}(v)x_i$. this yields $A_n(v) = F_0^{-1}(v)(1 - v)\bar{x}_n$. With $\hat{\beta}$ the ML estimator, leading to $\hat{\theta}_i = \exp(x_i^t \hat{\beta})$, let G_n the empirical c.d.f. for $\hat{v}_i = F_0(\hat{\theta}_i Y_i)$. With $J_n = n^{-1} \sum_{i=1}^n x_i x_i^t$, show that

$$W_n(v) = \sqrt{n}\{G_n(v) - v\} = \sqrt{n}\{F_n(v) - v\} - F_0^{-1}(v)(1 - v)\bar{x}_n^t J_n^{-1} \sqrt{n}U_n + o_{\text{pr}}(1),$$

with $\text{cov}\{F_n(v), U_n\} = n^{-1}A_n(v)$.

Ex. 10.42 *Chi-squared processes and F processes.* Just as various distributions may be constructed and used based on the normal distribution, like the log-normal, the chi-squared, the t, the F, we may also use Brownian motions to construct further processes of interest.

(a) With W a Brownian motion, consider $Z(t) = W(t)^2/t$. Show that $Z(t)$ has a χ_1^2 distribution, for each t .

(b) More generally, let $W_1, \dots, W_p$ be independent Brownian motions. Show that $Z_p(t) = \sum_{j=1}^p W_j(t)^2/t$ has a χ_p^2 distribution, for each t . Simulate and display some paths, for say $p = 6$.

(c) With integers p, q , construct a process $F = \{F(t) : t > 0\}$ such that for each t , $F(t)$ has the F distribution with parameters p, q .

(d) (xx briefly on the process $V(t) = \exp\{W(t)\}$. its multiplicative nature. xx)

Monitoring bridges, likelihood bridges, running t tests

Ex. 10.43 *Bridges from partial-sum processes.* Various naturally constructed empirical processes have Gaussian bridges as their limits.

(a) Let $U_1, U_2, \dots$ be i.i.d. with mean zero and variance one. Consider the process $V_n(t) = n^{-1} \sum_{i \leq [nt]} (U_i - \bar{U}_n)$, with the sample mean $\bar{U}_n = n^{-1} \sum_{i=1}^n U_i$, and note that it starts and ends at zero. Show indeed that $V_n \rightarrow_d W^0$.

(b) Now suppose $Y_1, \dots, Y_n$ are i.i.d. with mean θ and standard deviation σ . Show that

$$H_n(t) = n^{-1/2} \sum_{i \leq [nt]} (Y_i - \bar{Y}_n) \rightarrow_d \sigma W^0(t). \quad (10.8)$$

Show hence that both $H_n(t)/\sigma$ and $H_n(t)/\hat{\sigma}$ have the same Brownian bridge limit process, where $\hat{\sigma}$ is any consistent estimator for σ . Note that the θ parameter cancels out in the H_n construction.

(c) Consider the null hypothesis H_0 of constancy, that the distribution of the Y_i has not changed over time. Explain why the test which rejects the null if $\max_{0 \leq t \leq 1} |H_n(t)|/\hat{\sigma} > 1.358$ has test level 0.05.

(d) Using $H_n/\hat{\sigma} \rightarrow_d W^0$, valid under H_0 , construct a couple of other 0.05 level test statistics.

Ex. 10.44 *Detecting non-constant behaviour.* [xx a couple of things, to be polished and edited. xx] Suppose $Y_1, \dots, Y_n$ are independent, but with means $\theta_1, \dots, \theta_n$ which do not necessarily stay constant over time. The monitoring process $H_n(t) = n^{-1/2} \sum_{i \leq [nt]} (Y_i - \bar{Y})$ studied in Ex. 10.43 can be used to test the constancy hypothesis H_0 that $\theta_1 = \dots = \theta_n$. To learn about its detection abilities outside H_0 , suppose now that $\theta_i = \theta_0 + \delta m(i/n)/\sqrt{n}$, for some underlying function $m(s)$ on $[0, 1]$.

(a) Show that the mean of $n^{-1/2} \sum_{i \leq [nt]} Y_i$ tends to $\theta_0 + \delta \int_0^t m(s) ds$, and that

$$E H_n(t) \rightarrow \delta \int_0^t \{m(s) - \bar{m}\} ds, \quad \text{writing } \bar{m} = \int_0^1 m(s) ds.$$

(b) Now consider the particular case of there being a changepoint in the mean structure, with say $\theta_i = a$ for $i \leq \tau$ and $\theta_i = b$ for $i \geq \tau + 1$, which we translate to $m(s)$ having values a and b on $[0, t_0]$ and $(t_0, 1]$, with $t_0 = \tau/n$. Find and display the limiting mean function for $H_n(t)$. Show in particular that it has a triangular shape, with the triangle pointing upwards if $a > b$ and downwards if $a < b$; also, the point where the mean function attains its maximum (for $a > b$) or minimum (for $a < b$) is precisely the changepoint position t_0 .

(c) Consider the case where $Y_i = \theta_0 + \delta m(i/n)/\sqrt{n} + U_i$, where the U_i are i.i.d. with mean zero and standard deviation σ . Show that

$$H_n(t) = n^{-1/2} \sum_{i \leq [nt]} (Y_i - \bar{Y}) \rightarrow_d \delta \int_0^t (m - \bar{m}) ds + \sigma W^0(t).$$

Ex. 10.45 *More bridges.* Here are yet further monitoring bridges, useful for checking constancy of distributions over time, and for identifying changepoints when present.

(a) Consider again the setup of Ex. 10.43, for i.i.d. observations $Y_1, \dots, Y_n$ with variance σ^2 . We saw there that $H_n(t) = n^{-1/2} \sum_{i \leq [nt]} (Y_i - \bar{Y}_n)$ tends to $\sigma W^0(t)$. Writing $\bar{Y}_j$ for the mean of the first j data points, show that the value of H_n at position $t = j/n$ can be expressed as

$$H_n(t) = n^{-1/2} j(\bar{Y}_j - \bar{Y}_n) = n^{-1/2} j(1 - j/n)(\bar{Y}_j - \tilde{Y}_j) = \sqrt{nt}(1 - t)(\bar{Y}_j - \tilde{Y}_j),$$

in which $\tilde{Y}_j$ is the mean of the $n - j$ observations falling to the right of j . Show that the variance of $H_n(t)$ is $\sigma^2 t(1 - t)$.

(b) These findings, where several natural constructions for checking constancy of means lead to Brownian bridges, can now be lifted to other settings, and specifically to classes

of estimators. Suppose $\hat{\theta}_j$ is an estimator of some θ , after j data points. Assume these have the property that $\hat{\theta}_j - \theta_0 = (1/j) \sum_{i \leq j} U_i + \varepsilon_j$, for i.i.d. U_i with zero mean and some variance τ^2 , with $j^{1/2} \varepsilon_j \rightarrow_{\text{pr}} 0$ as j increases. Construct the process H_n with value $H_{n,j} = n^{-1/2} j(\hat{\theta}_j - \hat{\theta}_n)$ at $t = j/n$, and then constant over cells $[j/n, (j+1)/n]$. Show that $H_n \rightarrow_d \tau W^0$, over each $D[\delta, 1]$, for small δ .

(c) Explain that for any estimator $\hat{\theta}_n$ having an influence function with finite variance τ^2 , see Ex. 6.17, the monitoring process M_n with values $(1/\sqrt{n})j(\hat{\theta}_j - \hat{\theta}_n)/\hat{\tau}$ at $t = j/n$, will converge to the Brownian bridge. Here $\hat{\tau}$ is some consistent estimator of τ . Departure from Brownian bridge behaviour of this bridge plot thus signals that the model has not remained constant.

(d) Construct a dataset $X_1, \dots, X_n$ where the standard deviation σ changes midway. Check if the changepoint is visible in the monitoring plot $n^{-1/2}j(\hat{\sigma}_j - \hat{\sigma}_n)/\hat{\tau}$.

Ex. 10.46 *A running t test process to check left versus right.* Consider an i.i.d. sequence $Y_1, \dots, Y_n$ with mean μ and standard deviation σ . For any $\tau \in \{1, \dots, n\}$, divide data into a left part, $Y_1, \dots, Y_\tau$, and a right part, $Y_{\tau+1}, \dots, Y_n$, with estimates $\hat{\mu}_L = (1/\tau) \sum_{i \leq \tau} Y_i$ and $\hat{\mu}_R = (1/(n - \tau)) \sum_{i > \tau} Y_i$. The usual way of checking that the left and right parts actually have the same mean is by comparing $\hat{\mu}_R - \hat{\mu}_L$ to an estimate of its standard deviation, as with the classical t test, say

$$t_n(\tau) = \frac{\hat{\mu}_L - \hat{\mu}_R}{\{\hat{\sigma}_L^2/\tau + \hat{\sigma}_R^2/(n - \tau)\}^{1/2}}.$$

One may use separate standard deviation estimates for the left and right parts, or a common one. The idea now is to consider the full running t test process. Below we work under the null hypothesis assumption that the full sequence is i.i.d., but the point will later be to find changepoints, where the mean changes from one value to another one.

(a) With one test at the time, so to speak, show that $t_n(\tau) \rightarrow_d N(0, 1)$, as long as τ and $n - \tau$ both increase.

(b) Writing $s = \tau/n$, form the process

$$T_n(s) = \frac{\hat{\mu}_L - \hat{\mu}_R}{\{\hat{\sigma}_L^2/\tau + \hat{\sigma}_R^2/(n - \tau)\}^{1/2}} = \frac{\bar{Y}_{1,[ns]} - \bar{Y}_{[ns]+1,n}}{\{\hat{\sigma}_{1,[ns]}^2/[ns] + \hat{\sigma}_{[ns]+1,n}^2/(n - [ns])\}^{1/2}},$$

where subscripts L and R as above refer to sample parts $Y_1, \dots, Y_{[ns]}$ and $Y_{[ns]+1}, \dots, Y_n$. Show that $T_n(s) \rightarrow_d N(0, 1)$ for each $s \in (0, 1)$.

(c) From the Donsker theorem, we know that the process $A_n(s) = n^{-1/2} \sum_{i \leq [ns]} (Y_i - \mu)$ tends to $\sigma W(s)$. With $s = \tau/n$, show that

$$\sqrt{n}(\hat{\mu}_L - \mu) = A_n(s)/s \quad \text{and} \quad \sqrt{n}(\hat{\mu}_R - \mu) = \{A_n(1) - A_n(s)\}/(1 - s).$$

(d) Show from this that

$$\sqrt{n}(\hat{\mu}_L - \hat{\mu}_R) = \frac{A_n(s)}{s} - \frac{A_n(1) - A_n(s)}{1 - s} = \frac{A_n(s) - sA_n(1)}{s(1 - s)}.$$

(e) Let $v_n(s)$ be a suitable weight function, deterministic or depending on the data, in such a way that there is a limit in probability function $v_n(s) \rightarrow_{\text{pr}} v(s)$. Show that the following weighted difference test process converges in distribution:

$$Z_n(s) = \sqrt{n}v_n(s)(\hat{\mu}_L - \hat{\mu}_R) \rightarrow_d Z(s) = v(s)\sigma \frac{W(s) - sW(1)}{s(1 - s)} = \sigma \frac{v(s)W^0(s)}{s(1 - s)},$$

with the Brownian bridge $W^0(s) = W(s) - sW(1)$. Show also there is full process convergence $Z_n \rightarrow_d Z$ in each $D[\varepsilon, 1 - \varepsilon]$, for $\varepsilon > 0$, provided $\max_{\varepsilon \leq s \leq 1 - \varepsilon} |v_n(s) - v(s)| \rightarrow_{\text{pr}} 0$, i.e. we need uniform convergence in probability of the weight function.

(f) For the special case of the running t test process $T_n(s)$, write out the weight function $v_n(s)$ indirectly used there, and show that $v_n(s) \rightarrow_p \{s(1 - s)\}^{1/2}/\sigma$. Hence

$$T_n(s) = \frac{\hat{\mu}_L - \hat{\mu}_R}{\{\hat{\sigma}_L^2/\tau + \hat{\sigma}_R^2/(n - \tau)\}^{1/2}} \rightarrow_d T(s) = \frac{W^0(s)}{\sqrt{s(1 - s)}},$$

the naturally normalised Brownian bridge, to make the variance constant and equal to one. This is a process we met in Ex. 10.18.

(g) Suppose $Y_i = \mu_i + \varepsilon_i$, with the ε_i being i.i.d., and that we are to test the null hypothesis that all the means μ_i are equal, against the alternative that there is changepoint τ with one value μ_L for $i \leq \tau$ and another value μ_R for $i \geq \tau + 1$. A natural test statistic is

$$t_{\max} = \max_{\varepsilon \leq s \leq 1 - \varepsilon} |T_n(s)| = \max_{c \leq \tau \leq d} |T_n(\tau)|,$$

with $c = [n\varepsilon]$ and $d = [n(1 - \varepsilon)]$, for a suitable and perhaps small ε . Under the null hypothesis, show that $t_{\max} \rightarrow_d \max_{\varepsilon \leq s \leq 1 - \varepsilon} |T(s)|$, in the notation above. Exemplify this by drawing $y_i \sim N(a, 1)$ for $i = 1, \dots, 75$ and $y_i \sim N(b, 1)$ for $i = 76, \dots, 100$, for a, b of your own choice. Draw the $t_n(\tau)$ curve, and test the null, using $[\varepsilon, 1 - \varepsilon] = [0.05, 0.95]$. This calls for separate simulation of $\max_{0.05 \leq s \leq 0.95} |T(s)|$; see Ex. 10.18.

(h) For some applications we encounter processes of the above type, with running tests based on $\hat{\mu}_R - \hat{\mu}_L$, but based on observations with different variances. This happens e.g. in Story #40, where we compare word lengths for a sequence of chapters of different sizes. Consider therefore Y_i with mean μ_i and variance σ^2/m_i , with left-and-right averages $\hat{\mu}_L = \sum_{i \leq \tau} m_i Y_i / \sum_{i \leq \tau} m_i$ and $\hat{\mu}_R = \sum_{i \geq \tau+1} m_i Y_i / \sum_{i \geq \tau+1} m_i$. As above, write $s = \tau/n$, with τ growing with n . The natural test ratio to monitor is then

$$T_n(s) = \frac{\hat{\mu}_L - \hat{\mu}_R}{(\hat{\sigma}_L^2 / \sum_{i \leq \tau} m_i + \hat{\sigma}_R^2 / \sum_{i \geq \tau+1} m_i)^{1/2}},$$

where $\hat{\sigma}_L^2 = \tau^{-1} \sum_{i \leq \tau} m_i (Y_i - \hat{\mu}_L)^2$ and $\hat{\sigma}_R^2 = (n - \tau)^{-1} \sum_{i \geq \tau+1} m_i (Y_i - \hat{\mu}_R)^2$. The relevant empirical process to examine is

$$A_n(s) = \left(\sum_{i=1}^n m_i \right)^{-1/2} \sum_{i \leq ns} m_i (Y_i - \mu) \quad \text{for } s \in [0, 1].$$

For technical reasons we assume the background sample sizes $m_1, \dots, m_n$ are well-behaved, with $\bar{m}_n = n^{-1} \sum_{i=1}^n m_i$ tending to some m_0 . Show via Ex. 10.48 that $A_n \rightarrow_d \sigma W$, and that arguments as above extend to the result $T_n(s) \rightarrow_d \sigma W^0(s) / \{s(1 - s)\}^{1/2}$. Explain how this leads to overall running t tests type statistics.

Ex. 10.47 *Generalisations and uses of the running t test process.* The apparatus of Ex. 10.46 can be used to test the homogeneity hypothesis against the existence of one of more changepoints for the mean. The machinery extends nicely and fruitfully to other parameters, as developed here. Consider in general terms $Y_1, \dots, Y_n$ being i.i.d. from some distribution F , with $\theta = \theta(F)$ a parameter of interest, like the mean, the standard deviation, a quantile, or a threshold probability. We also assume that there is a well-defined estimator $\hat{\theta}_{a,b}$ for θ based on the data stretch $Y_a, \dots, Y_b$. We work under the null

hypothesis here, but the aim is to build a running t test process for monitoring the θ_i associated with Y_i . Such a process should lead to tests to pick up that the θ_i might have changed.

(a) We assume that the estimand θ and its estimator admit an influence function $\text{IF}(F, y)$, with finite variance, as per Ex. 6.17. Show that

$$\hat{\theta}_{a,b} = \theta + \frac{1}{b-a+1} \sum_{i=a}^b \text{IF}(F, Y_i) + \varepsilon_{a,b},$$

with the remainder term small enough to have $(b-a+1)^{1/2} \varepsilon_{a,b} \rightarrow_{\text{pr}} 0$ when the data stretch length $b-a+1$ increases. Show that $(b-a+1)^{1/2}(\hat{\theta}_{a,b} - \theta) \rightarrow_d N(0, \kappa^2)$, where $\kappa^2 = \text{Var}_F \text{IF}(F, Y)$.

(b) Now construct, for each τ , the t type test statistic

$$t_n(\tau) = \frac{\hat{\theta}_L - \hat{\theta}_R}{\{\hat{\kappa}_L^2/\tau + \hat{\kappa}_R^2/(n-\tau)\}^{1/2}},$$

with $\hat{\theta}_L$ and $\hat{\kappa}_L$ defined for the left data $Y_1, \dots, Y_\tau$ and similarly $\hat{\theta}_R$ and $\hat{\kappa}_R$ for the right data $Y_{\tau+1}, \dots, Y_n$. [xx then the point is to find the same limit as for Ex. 10.46. xx]

(c) (xx illustration: normal data, with shift in μ and/or σ . xx)

(d) (xx illustration: following the median over time. xx)

Ex. 10.48 A Lindeberg type extension of the Donsker theorem. Often enough one encounters partial-sum processes where the summands are not i.i.d. Assume that $U_1, U_2, \dots$ are independent with zero means, but perhaps with different distributions. With $\sigma_1, \sigma_2, \dots$ their standard deviations, we suppose these are sufficiently regular to have $n^{-1} \sum_{i=1}^n \sigma_i^2 \rightarrow \kappa^2$, say. Now build the process $X_n(t) = n^{-1/2} \sum_{i \leq [nt]} U_i$ for $t \geq 0$.

(a) Show that for each t , $X_n(t) \rightarrow_d N(0, \kappa^2 t)$. Under the Lindeberg type condition that $n^{-1} \sum_{i=1}^n E U_i^2 I(|U_i| \geq \sqrt{n}\varepsilon) \rightarrow 0$ for each positive ε , show that $X_n \rightarrow_d \kappa W$.

(b) (xx give an example. point to monitoring bridges for regression models. and perhaps generalise to weaker conditions. perhaps the most natural bridge uses $m_n(t) = \sum_{i \leq [nt]} \kappa_i^2 / \sum_{i=1}^n \kappa_i^2$, which tends to t , and then the bridge constriction $Z_n(t) = X_n(t) - m_n(t)X_n(1)$, tending to $Z(t) = X(t) - tX(1) = \kappa W^0(t)$. xx)

Ex. 10.49 A Donsker type theorem for regression coefficients. Consider a regression type sequence (x_i, y_i) , where the x_i are increasing, as in a time series context, and where it is assumed that $y_i = bx_i + \varepsilon_i$, with the ε_i being i.i.d. with zero mean and variance σ^2 . We may then estimate the regression coefficient based on the first j datapoints, as $\hat{b}_j = \sum_{i \leq j} x_i y_i / M_j$, with $M_j = \sum_{i \leq j} x_i^2$, and these may be monitored over time. We assume that the x_i are well-behaved, with (i) $n^{-1} M_n$ tending to some positive m_0 for growing n and (ii) $\max_{i \leq n} x_i^2 / M_n \rightarrow 0$.

(a) Show first that $n^{-1} M_{[nt]} \rightarrow m_0 t$. Consider then the process $A_n(t) = n^{-1/2} \sum_{i \leq [nt]} x_i \varepsilon_i$. Explain that it also be written

$$n^{-1/2} \sum_{i \leq [nt]} x_i (y_i - bx_i) = n^{-1} M_{[nt]} \sqrt{n} (\hat{b}_{[nt]} - b).$$

Show that variance of $A_n(t)$ tends to $m_0 \sigma^2 t$, and that in fact $A_n \rightarrow_d m_0^{1/2} \sigma W$, a scaled Brownian motion. You may use results from Ex. 2.41.

(b) We then generalise to the more natural setup with a full two-parameter regression line, with $y_i = a + bx_i + \varepsilon_i$, and again with the ε_i being i.i.d. with zero mean and variance σ^2 . Writing $\bar{x}_j = j^{-1} \sum_{i \leq j} x_i$, we assume that $M_n = \sum_{i=1}^n (x_i - \bar{x}_n)^2$ tends to some m_0 , and that $\max_{i \leq n} (x_i - \bar{x}_n)^2 / M_n \rightarrow 0$. Show that

$$B_n(t) = n^{-1/2} \sum_{i \leq [nt]} (x_i - \bar{x}_{[nt]}) \varepsilon_i \rightarrow_d m_0^{1/2} \sigma W(t),$$

again a scaled Brownian motion. With $\hat{b}_j = \sum_{i \leq j} (x_i - \bar{x}_j) y_i$ estimating b , show also that $B_n(t) = n^{-1} M_{[nt]} \sqrt{n} (\hat{b}_{[nt]} - b)$.

(c) Explain that the methods and results of Ex. 10.46–10.47 may be extended to the present case, monitoring the regression estimators $\hat{b}_j$ over time. For given τ , divide data into left, with $i \leq \tau$, and right, with $i \geq \tau + 1$, and ensuing $\hat{b}_L$ and $\hat{b}_R$, with variances $\sigma^2 \kappa_L^2 / \tau$ and $\sigma^2 \kappa_R^2 / (n - \tau)$, writing $\kappa_L^2 = \sum_{i \leq \tau} (x_i - \bar{x}_L)^2$ and $\kappa_R^2 = \sum_{i \geq \tau+1} (x_i - \bar{x}_R)^2$, in which $\bar{x}_L$ and $\bar{x}_R$ are the averages to the left and right. Then define the monitoring process

$$T_n(\tau) = \frac{\hat{b}_R - \hat{b}_L}{\hat{\sigma} \{ \kappa_L^2 / \tau + \kappa_R^2 / (n - \tau) \}^{1/2}},$$

using $\hat{\sigma}^2 = (\tau/n) \hat{\sigma}_L^2 + (1 - \tau/n) \hat{\sigma}_R^2$. Explain that with $\tau/n \rightarrow s$, we have $T_n(\tau) \rightarrow_d T(s) = W^0(s) / \{s(1-s)\}^{1/2}$, and that $T_{n,\max} = \max_{c \leq \tau/n \leq d} |T_n(\tau)|$ tends in distribution to $T_{\max} = \max_{c \leq s \leq d} |T(s)|$. Quantiles for this distribution, for some intervals $[c, d]$, were found in Ex. 10.18. (xx point to Oslo temperatures story. xx)

Ex. 10.50 *A log-likelihood maxima bridge.* Suppose i.i.d. data $Y_1, \dots, Y_n$ arise in orderly sequential fashion, and consider a parametric distribution $f(y, \theta)$ for these, of dimension say p . Here we build a log-likelihood-maxima bridge, which may be used to check whether something unusual has taken place in the course of time points $1, \dots, n$.

(a) Start out considering $\ell_j = \sum_{i \leq j} \log f(Y_i, \theta_0)$, with θ_0 the true parameter, along with the attained maximum $\hat{\ell}_j = \sum_{i \leq j} \log f(Y_i, \hat{\theta}_j)$, where $\hat{\theta}_j$ is the ML estimator after j datapoints. Show that ℓ_j/j as well as $\hat{\ell}_j/j$ tends to $\xi = E \log f(Y_i, \theta_0)$. Plotting $\hat{\ell}_j$ as a function of j should accordingly be approximately a straight line, with derivative ξ .

(b) Use Ex. 10.43 to show that the process $A_n(t) = n^{-1/2} \{ \ell_{[nt]} - ([nt]/n) \ell_n \}$, with values $A_{n,j} = n^{-1/2} \{ \ell_j - (j/n) \ell_n \}$ at j/n , tends to $\kappa W^0(t)$, with $\kappa^2 = \text{Var} \log f(Y_i, \theta_0)$.

(c) From the log-likelihood maxima construct the process B_n with values $B_{n,j} = n^{-1/2} \{ \hat{\ell}_j - (j/n) \hat{\ell}_n \}$ at j/n . The $B_n(t)$ process needs at least p data points to get started, so technically its time domain is $[p/n, 1]$, and if pedantry demands can be put to zero in the small interval before p/n . Show that the $\hat{\ell}_j - \ell_j$ differences are bounded in probability (xx point to Wilks Ch5 and chi-squared things here xx). Use this to demonstrate that B_n and A_n are close enough for them to have the same process limit, i.e. κW^0 .

(d) Show how this leads to model monitoring bridge plots

$$M_{n,j} = B_{n,j} / \hat{\kappa} = (1/\sqrt{n}) \{ \hat{\ell}_j - (j/n) \hat{\ell}_n \} / \hat{\kappa} \quad \text{for } j = p, \dots, n,$$

which under nothing-has-changed model conditions will behave as a Brownian bridge. In particular, $\max_{p \leq j \leq n} M_{n,j} \rightarrow_d M = \max_{0 \leq s \leq 1} |W^0(s)|$, with 1.358 as upper 0.05 point, etc., see Ex. 10.26. Here $\hat{\kappa}$ is an estimator for κ , computed either under model conditions, i.e. as the appropriate $\kappa(\hat{\theta})$, or as the empirical standard deviation of the observed $\log f(y_i, \hat{\theta})$.

(e) Argue via Wilks theorems of Ch. 5 that $\widehat{\ell}_j - \ell_j = \frac{1}{2}W_j + \varepsilon_j$, say, with W_j tending to a χ_p^2 and the ε_j small. This invites the bias correction $\widehat{\ell}_j - \frac{1}{2}p$, to come closer to ℓ_j when constructing the bridge. Show that this leads to the modified bridge plot

$$M'_{n,j} = B'_{n,j}/\widehat{\kappa} = (1/\sqrt{n})\{\widehat{\ell}_j - (j/n)\widehat{\ell}_n - \frac{1}{2}p(1 - j/n)\}.$$

Verify that it has the same Brownian bridge limit; the intention is to come closer to this limit.

(f) The ℓ_j and $\widehat{\ell}_j$ grow linearly over time, so it may be convenient to compute and present the plots in terms of $H_j = \widehat{\ell}_j/j$. Show that $M_{n,j} = n^{-1/2}j(H_j - H_n)/\widehat{\kappa}$.

Ex. 10.51 *Likelihood monitoring bridges for some models.* Here we use the general setup of Ex. 10.50 to set up clear likelihood monitoring bridges for a few parametric models.

(a) For the case of normal $N(\xi, \sigma^2)$ data, show that $M_{n,j} = n^{-1/2}j(\log \widehat{\sigma}_n - \log \widehat{\sigma}_j)/\widehat{\kappa}$, where $\widehat{\sigma}_j^2 = (1/j) \sum_{i \leq j} (y_i - \bar{y}_j)^2$ is the empirical variance after j data points. Also, show that $\kappa = 1/\sqrt{2}$, under model conditions, but we may also take $\widehat{\kappa}$ to be the empirical standard deviation for the observed $\frac{1}{2}(y_i - \widehat{\xi})^2/\widehat{\sigma}^2$.

(b) For the $\text{Pois}(\theta)$ model, show that $M_{n,j} = n^{-1/2}j\{H(\bar{y}_j) - H(\bar{y}_n)\}/\widehat{\kappa}$, with $\bar{y}_j$ the mean of the first j data points and $H(u) = u \log u - u$. Also, $\widehat{\kappa} = (\bar{y}_n)^{1/2} \log \bar{y}_n$, under model conditions, or one may use the model robust $\widehat{\kappa} = \widehat{\sigma}_n \log \bar{y}_n$, with $\widehat{\sigma}_n$ the standard deviation of the data.

(c) You observe Bernoulli 0-1 outcomes $Y_1, \dots, Y_n$ and wonder whether the underlying probability p has remained unchanged during the process. With $K(p) = p \log p + (1 - p) \log(1 - p)$, show that the monitoring bridge takes the form $M_{n,j} = n^{-1/2}j\{K(\widehat{p}_j) - K(\widehat{p}_n)\}/\widehat{\kappa}$, with $\widehat{p}_j$ the estimate after j data points, and $\widehat{\kappa}^2$ the estimate of $\kappa^2 = p(1 - p)[\log\{p/(1 - p)\}]^2$. Set up a simple simulation experiment, to check out how this works: take $n = 200$, with p equal to 0.33 for the first 100 and 0.44 for the next hundred. See if the $M_{n,j}$ plot picks up that p has not remained constant.

(d) Consider a trinomial process, with $(Y_{i,1}, Y_{i,2}, Y_{i,3})$ being $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ in experiment i , for cases A, B, C, say, with probabilities p_1, p_2, p_3 summing to one. Show that the likelihood monitoring plot takes the form

$$M_{n,j} = n^{-1/2}j\{K(\widehat{p}_{1,j}, \widehat{p}_{2,j}, \widehat{p}_{3,j}) - K(\widehat{p}_{1,n}, \widehat{p}_{2,n}, \widehat{p}_{3,n})\}/\widehat{\kappa},$$

with $(\widehat{p}_{1,j}, \widehat{p}_{2,j}, \widehat{p}_{3,j})$ the estimated probabilities after j experiments, and $K(p_1, p_2, p_3) = p_1 \log p_1 + p_2 \log p_2 + p_3 \log p_3$. Also, show that the required $\widehat{\kappa}$ should be the estimate of $\kappa^2 = \sum_j p_j (\log p_j)^2 - (\sum_j p_j \log p_j)^2$. Set up a simple simulation experiment to see how the plots look and to see if they can pick up a modest change of (p_1, p_2, p_3) .

(e) (xx then one more. find a dataset to work through, where the pedagogical point would be to demonstrate that these monitoring plots are fine to set up and use just using the numerical tools from ML estimation, without clear formulae. something with history, or emil's basketball data over time. point to Stories #38 and #40. xx)

Ex. 10.52 *Likelihood monitoring bridges for regression models.* (xx point here is to generalise the above monitoring plots to regression setups in the right way. at the moment nils ranting away. see nils note; sum of two Brownian bridges in the limit. xx) consider a model $f(y_i | x_i, \theta_0)$, with mean ξ_i and variance κ_i^2 for $\log f(Y_i | x_i, \theta_0)$. then build the partial-sum process $A_n(t)$, taking values

$$n^{-1/2}(\ell_j - a_j) = n^{-1/2} \sum_{i \leq j} \{\log f(y_i | x_i, \theta_0) - \xi_i\}$$

at positions $t = j/n$, with $a_j = \sum_{j \leq i} \xi_j$. the variance is $n^{-1} \sum_{i \leq [nt]} \kappa_i^2$, tending to say $\kappa^2 t$, under ergodic assumptions. by the Lindeberg extension of the Donsker theorem, see Ex. 10.48. $A_n(t) \rightarrow_d W(\kappa^2 t)$. from this the bridge construction, but careful. $B_n(t) = A_n(t) - m_n(t)A_n(1)$.

(a) (xx check nils note; sum of two Brownian bridges. xx)

(b) (xx but this is going to be super useful for time series models, where the basic construction still works, though needing more arguments for the basic Brownian bridge lemma. xx)

Ex. 10.53 *Monitoring processes for constancy or parameters over time.* [xx the basics from Hjort and Koning (2002). processes tending to Brownian bridges. application to goodness of fit, and to testing the hypothesis of θ_i having remained unchanged over time. xx]

Ex. 10.54 *Convergence in distribution, for random indexes.* Suppose $Z_n = \sqrt{n}(A_n - a) \rightarrow_d Z_0$, say, and that N_n is a sequence of random indexes, growing to infinity. Will it be the case that $Z_{N_n} = N_n^{1/2}(A_{N_n} - a)$ tends to the same limit?

(a) To realise that the statement above cannot be taken for granted, without further conditions, consider $Y_1, Y_2, \dots$ being independent standard normal, for which of course $Z_n = \sqrt{n}\bar{Y}_n \rightarrow_d N(0, 1)$. Show that there will be infinitely many n at which Z_n is positive, and let $N_1 < N_2 < \dots$ be these time points. Show that the N_n sequence increases without bounds. Clearly Z_{N_n} does not have the normal limit; the CLT does not work for this increasing subsequence.

(b) Suppose $Z_n = \sqrt{n}(A_n - a) \rightarrow_d Z_0$, and that a sequence of random indexes N_n satisfies $N_n/n \rightarrow_{\text{pr}} c$, with $c \in (0, 1)$. Define the random functions $Z_n(t) = \sqrt{n}(A_{[nt]} - a)$ and $V_n(t) = N_n t/n$, in the function space $D[\varepsilon, 1]$, with some fixed $\varepsilon \in (0, c)$. Show that $V_n \rightarrow_d V$, where V is the deterministic function $V(t) = ct$, and that $Z_n \rightarrow_d Z$, where $Z(t) = Z_0/t^{1/2}$.

(c) Show that $(Z_n, V_n) \rightarrow_d (Z, V)$, in the product function space $D[\varepsilon, 1]^2$, and that this limit is in the subspace $C[\varepsilon, 1]^2$. (xx i hope this is correct. xx) Deduce from this that $Z_n \circ V_n \rightarrow_d Z \circ V$, i.e. that there is process convergence $Z_n(V_n(t)) \rightarrow_d Z(V(t))$. In particular, show from this that

$$Z_n(V_n(1)) = \sqrt{n}(A_{N_n} - a) \rightarrow_d Z(V(1)) = Z_0/c^{1/2},$$

and that $N_n^{1/2}(A_{N_n} - a) \rightarrow_d Z_0$ follows from this. Show in particular that $A_{N_n} \rightarrow_{\text{pr}} a$.

(d) (xx give one or two easy applications. xx)

Notes and pointers

A few remarks. Read more about risk functions in DeGroot (1970).

Point to Beta processes, Gamma processes, more general than the standard case tended to in Ex. 10.21, Dirichlet processes in Ch. 14; these live in $D[0, \tau]$ spaces.

(xx pointers, more. to Cunen et al. (2020a) and to Story #40. xx)

(xx for Ex. 10.35, bootstrapping: point to Efron and EfronTibs and more. there are modifications of the direct method, but this exercise has the basics. also parallel to Bayesian bootstrap, via BvM theorem for Dirichlet process priors, in Ch14. also: the basic thing holds for all Hadamard differentiable $T(F)$. xx)

(xx for Ex. 10.32: Further technical issues, also for tightness of the $A_n(u)$ process, to complete these arguments; see Doss and Gill (1992). xx)

(xx point to Parthasarathy (1967) and compare with Billingsley (1968), equally good at the same time. xx)

(xx this is too remarky; perhaps push to Notes. xx) Note in particular that since D_n has a limit distribution, we must have $D_n/\sqrt{n} \rightarrow_{\text{pr}} 0$; the maximal difference, between the e.c.d.f. F_n and the true F , goes to zero. The stronger version of this statement, that this happens with probability 1, is the Glivenko–Cantelli theorem of Ex. 2.60. It tells us that with lots of data, we can uncover any tiny little aspect of their underlying distribution, we can be as sophisticated as we might wish to. The result above is however more informative, since we learn how quickly it goes to zero, namely with speed $1/\sqrt{n}$. This is incidentally impossible for density estimators, where the best one can hope for is that the maximal distance $\max_t |f_n(t) - f(t)|$ goes to zero with speed $1/n^{2/5}$.

Brown, Einstein, Wiener, Donsker, Billingsley. Empirical processes with estimated parameters: Durbin 1973, Hjort 1990, Loynes 1980.

Robert Brown (1773–1858) is the British botanist, who once had a cup of tea; Norbert Wiener (1894–1964) the father of cybernetics etc.; also Albert Einstein (1879–1955) belongs on the list of famous scientists who early on worked with this process.

point to Story #100: An extension of the Donsker theorem with positive kurtosis. [xx to come later. nils construction, perhaps touched in chapter 2, with

$$X_n(t) = \sum_{i \leq [nt]} J_i(i/n)^\alpha U_i, \quad \text{for } t \geq 0,$$

with the U_i being i.i.d. with mean zero and variance one, and the J_i being an independent sequence of independent Bernoulli variables, with $\Pr(J_i = 1) = 1/i$. There is $X_n \rightarrow_d X$ process convergence, with extra kurtosis. xx]

The tightness criterion (10.3) of Ex. 10.19 is a useful extension of Billingsley (1968, Th. 15.6).

[xx make a point of Glivenko–Cantelli, $D_n/\sqrt{n} \rightarrow_{\text{a.s.}} 0$, which we perhaps have as an exercise in Ch. 2, and that Kolmogorov’s $D_n \rightarrow_d D$ is more informative. i think both results appeared in 1933. xx]

[xx cool log-likelihood monitoring bridge: from Cunen et al. (2018). pointer to Story #38. xx]

[xx should have gamma process and poisson process, to illustrate the use of such limit processes inside $D[0, 1]$. need care with what to use where etc. cross-check with other chapters, also Ch. 14. xx]

[xx may consider pushing math details for C and D spaces to an envisaged Technical Appendix, with various nitty-gritty details. here description of compacta; the $w''(x, \delta)$ modulus for D ; the Skorohod topology; the Prohorov theorem about tightness. xx]

[xx at some point we need to point to compactness characterisations, in C and in D , without making too much of it. mention Hanche-Olsen and Holden (2010), which binds together the Arzelà–Ascoli theorem and the Kolmogorov compactness theorem. xx]

ToDo: nils does more cleaning. nils makes some exercises with Poisson process, with Gamma process, pointing also to chapter on Bayesian nonparametrics.

Part II

100 Statistical Stories

Part III

Appendix

References

- Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726.
- Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. In Klonecki, W., Kozek, A., and Rosiński, J., editors, *Mathematical Statistics and Probability Theory. Lecture Notes in Statistics*, pages 1–25. Springer.
- Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925.
- Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2:951–972.
- Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis. A process point of view*. Springer, New York.
- Aalen, O. O. and Gjessing, H. K. (2004). Survival models based on the Ornstein–Uhlenbeck process. *Lifetime Data Analysis*, 10:407–423.
- Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover, New York.
- Aït-Sahalia, Y. and Jacod, J. (2014). *High-Frequency Financial Econometrics*. Princeton University Press, Princeton.
- Andersen, P. K. and Borgan, Ø. (1985). Counting process models for life history data: A review [with discussion and reply]. *Scandinavian Journal of Statistics*, 12:97–158.
- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.
- Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24:3–30.
- Ashworth, S., Berry, C. R., and de Mesquita, E. B. (2021). *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press, Princeton.
- Aursnes, I., Tvette, I. F., Gåsemyr, J., and Natvig, B. (2005). Suicide attempts in clinical trials with paroxetine randomised against placebo. *BMC Medicine*, xx:1–5.
- Aursnes, I., Tvette, I. F., Gåsemyr, J., and Natvig, B. (2006). Even more suicide attempts in clinical trials with paroxetine randomised against placebo. *BMC Psychiatry*, xx:1–3.
- Ball, P. (1999). *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*. American Academy for the Advancement of Science, Washington.
- Barfort, S., Klemmensen, R., and Larsen, E. G. (2020). Longevity returns to political office. *Political Science Research and Methods*, 9:658–664.
- Barry, D. and Hartigan, J. A. (1987a). Asynchronous distance between homologous dna sequences. *Biometrics*, 43:261–276.

- Barry, D. and Hartigan, J. A. (1987b). Statistical analysis of hominoid molecular evolution. *Statistical Science*, 2:191–210.
- Bartolucci, F. and Lupparelli, M. (2008). Focused Information Criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistics*, 35:629–649.
- Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Press, London.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, 15:377–180.
- Bernstein, F. (1924). Ergebnisse einer biostatistischen zusammenfassenden Betrachtung über die erblichen Blutstrukturen des Menschen. *Klinische Wochenschriften*, 3:1467–1495.
- Bickel, P. J. and Doksum, K. A. (2015). *Mathematical Statistics [2nd ed.]*. CRC Press, New York.
- Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Chicago University Press, Chicago.
- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- Billingsley, P. (1995). *Probability and Measure. Third Edition*. Wiley, New York.
- Blower, J. G., Cook, L. M., and Bishop, J. A. (1981). *Estimating the Size of Animal Populations*. Allen & Unwin, London.
- Boitsov, V. D., Karsakov, A. L., and Trofimov, A. G. (2012). Atlantic water temperature and climate in the barents sea, 2000–2009. *ICES Journal of Marine Science*, 69:833–840.
- Bolt, U. (2013). *Faster Than Lightning: My Autobiography*. HarperSport, London.
- Borgan, Ø., Fiaccone, R. L., Henderson, R., and Barreto, M. L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in brazil. *Scandinavian Journal of Statistics*, 34:53–69.
- Brazzale, A. R. and Davison, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science*, 23:465–484.
- Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied asymptotics: Case studies in small-sample statistics*. Cambridge University Press, Cambridge.
- Breiman, L. (2001). Statistical modeling: The two cultures [with comments and a rejoinder by the author]. *Statistical Science*, 16:199–231.
- Brown, L. D. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, 9:i–279.
- Brunborg, H., Lyngstad, T. H., and Urdal, H. (2003). Accounting for genocide: How many were killed in Srebrenica? *European Journal of Population*, 19:229–248.
- Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84:772–793.
- Casella, G. and Berger, R. L. (2013). *Statistical Inference [2nd ed.]*. Duxbury, Canberra.
- Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *American Statistician*, 46:167–174.
- Claeskens, G., Cunen, C., and Hjort, N. L. (2019). Model selection via Focused Information Criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution*, 7:415–428.
- Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:900–916.
- Claeskens, G. and Hjort, N. L. (2008a). Minimizing average risk in regression. *Econometric Theory*, 24:493–527.

- Claeskens, G. and Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.
- Claeskens, G. and Hjort, N. L. (2027). The focused information criterion. *Annual Review of Statistics and Its Application*, xx:xx–xx.
- Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances*, 4:1–9.
- Clauset, A. (2020). On the frequency and severity of interstate wars. In Gleditsch, N. P., editor, *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences*, pages 113–128. Springer, Berlin.
- Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous estimation of the means of independent Poisson laws. *Journal of the American Statistical Association*, 70:698–705.
- Cox, D. R. (1958). Some problems with statistical inference. *The Annals of Mathematical Statistics*, 29:357–372.
- Cox, D. R. (1972). Regression models and life-tables [with discussion]. *Journal of the Royal Statistical Society: Series B*, 34:187–202.
- Cox, D. R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. *Journal of the Royal Statistical Society Series B*, 21:195–200.
- Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall, London.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.
- Cramér, H. (1976). Half a century with probability theory: some personal reflections. *Annals of Probability Theory*, 4:509–546.
- Cunen, C. (2015). Mortality and Nobility in the Wars of the Roses and Game of Thrones. *FocuStat Blog, University of Oslo*, iv.
- Cunen, C., Hermansen, G. H., and Hjort, N. L. (2018). Confidence distributions for change points and regime shifts. *Journal of Statistical Planning and Inference*, 195:14–34.
- Cunen, C. and Hjort, N. L. (2015). Optimal inference via confidence distributions for two-by-two tables modelled as Poisson pairs: fixed and random effects. In Nair, V., editor, *Proceedings of the 60th World Statistics Congress, ISI Rio*, pages xx–xx. Springer, Rio.
- Cunen, C. and Hjort, N. L. (2022). Combining information from diverse sources: the II-CC-FF paradigm. *Scandinavian Journal of Statistics*, 49:625–656.
- Cunen, C. and Hjort, N. L. (2025). Survival and event history models and methods via Gamma processes. Technical report, University of Oslo. Technical report.
- Cunen, C., Hjort, N. L., and Nygård, H. M. (2020a). Statistical sightings of better angels. *Journal of Peace Research*, 57:221–234.
- Cunen, C., Hjort, N. L., and Schweder, T. (2020b). Confidence in confidence distributions! *Proceedings of the Royal Society, A*, 476:1–5.
- Cunen, C., Walløe, L., and Hjort, N. L. (2020c). Focused model selection for linear mixed models, with an application to whale ecology. *Annals of Applied Statistics*, 14:872–904.
- Dæhlen, I. and Hjort, N. L. (2025). Model robust hybrid likelihood. *Journal of Statistical Planning and Inference*, 241:xx–xx.
- Dæhlen, I., Hjort, N. L., and Hobæk Haff, I. (2024). Accurate bias estimation with applications to focused model selection [to appear]. *Scandinavian Journal of Statistics*, 51:724–759.
- Dagsvik, J. K., Fortuna, M., and Moen, S. H. (2020). How does temperature vary over time?: Evidence on the stationary and fractal nature of temperature fluctuations. *Journal of the Royal Statistical Society A*, pages 883–908.

- Daniell, P. J. (1920). Observations weighted according to order. *American Journal of Mathematics*, pages 222–236.
- Davison, A. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
- De Blasi, P. and Hjort, N. L. (2007). Bayesian survival analysis in proportional hazard models with logistic relative risk. *Scandinavian Journal of Statistics*, 34:229–257.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. John Wiley & Sons, Hoboken, N.J.
- Dickter, R. (2014). *Number Time Archetype: The Significance of Magic Squares in China and the West*. Grass Valley, Grass Valley.
- Dickter, R. (2026). *The Secret of the Magic Square: Numerology in Art and Architecture*. Grass Valley, Grass Valley.
- Doksum, K. A. (1974a). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2:267–277.
- Doksum, K. A. (1974b). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2:183–201.
- Doksum, K. A. and Stevers, G. L. (2006). Plotting with confidence: graphical comparisons of two populations. *Biometrika*, 63:421–434.
- Doss, H. and Gill, R. D. (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data. *Journal of the American Statistical Association*, 87:869–877.
- Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge.
- Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.
- Edwards, A. W. F. (1974). The history of likelihood. *International Statistical Review*, 42:9–15.
- Edwards, A. W. F. (2025). *Latin Square: Essays in Defense of R.A. Fisher*. River Cam Books, Cambridge.
- Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, Cambridge.
- Efron, B. (2023). *Exponential Families in Theory and Practice*. Cambridge University Press, Cambridge.
- Efron, B. and Morris, C. (1977). Stein’s paradox in statistics. *Scientific American*, 236:119–127.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Einmahl, J. H. J. and Smeets, S. G. W. R. (2011). Ultimate 100 m world records through extreme-value theory. *Statistica Neerlandica*, 65:32–42.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, London.
- Fagerland, M., Lydersen, S., and Laake, P. (2017). *Statistical Analysis of Contingency Tables*. Chapman and Hall/CRC, New York.
- Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.
- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, 2:615–629.
- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

- Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770.
- Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.
- Franklin, B. (1793). *The Autobiography of Benjamin Franklin*. Dover, New York. Reprinted from Dover, New York, 1996.
- Friesinger, A. (2004). *Mein Leben, mein Sport, meine besten Fitness-Tipps*. Goldmann, Berlin.
- Frigessi, A. and Hjort, N. L. (2002). Statistical methods for discontinuous phenomena. *Journal of Nonparametric Statistics*, 14:1–5.
- Gao, L., Shen, H., and Liu, C. (2021). Statistical analysis of songs by composers of the Romantic era: A comparison of F. Schubert and his contemporaries. *Música Hodie*, 21:xx–xx.
- Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Friedrich Perthes und I.H. Besser, Hamburg.
- Geißler, A. (1889). Beiträge zur Frage des Geschlechts verhältnisses der Geborenen. *Zeitschrift des königlichen sächsischen statistischen Bureaus*, 35:1–24.
- Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and Other Stories*. Cambridge University Press, Cambridge.
- Gelman, A., Hill, J., and Vehtari, A. (2022). *Regression and Other Stories*. Cambridge University Press, Cambridge.
- Gelman, A. and Nolan, D. (2002). A probability model for golf putting. *Teaching Statistics*, 24:93–95.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.
- Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.
- Ghosh, J. K., Hjort, N. L., Messan, C., and Ramamoorthi, R. V. (2006). Bayesian bivariate survival estimation. *Journal of Statistical Planning and Inference*, 136:2297–2308.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Verlag, Berlin.
- Ghosh, M. (2002). Basu’s theorem with applications: a personalistic review. *Sankhya*, 35:721–741. Special issue in memory of D. Basu.
- Gilby, W. H. and Pearson, K. (1911). On the significance of the teacher’s appreciation of general intelligence. *Biometrika*, 8:94–108.
- Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Annals of Statistics*, 18:1501–1555.
- Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314.
- Gjessing, H. K., Aalen, O. O., and Hjort, N. L. (2003). Frailty models based on Lévy processes. *Advances in Applied Probability*, 35:532–550.
- Gleditsch, N. P. (2020). *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences (edited book)*. Springer, Berlin.
- Goudie, I. B. J. and Goudie, M. (2007). Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society, Series A*, 170:825–839.
- Green, P. J., Hjort, N. L., and Richardson, S., editors (2003). *Highly Structured Stochastic Systems*. Oxford University Press.

- Grønneberg, S. and Hjort, N. L. (2012). On the errors committed by sequences of estimator functionals. *Mathematical Methods of Statistics*, 20:327–346.
- Groß, J. and Möller, A. (2024). The zero degree of freedom non-central chi squared distribution for ensemble postprocessing. *EMOS Meteorologische Zeitschrift*, xx:1–20.
- Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrics*, 11:1–12.
- Hall, P. (1927). The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, 19:240–245.
- Hanche-Olsen, H. and Holden, H. (2010). The Kolmogorov-Riesz compactness theorem. *Expositiones Mathematicae*, 28:385–394.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 9:97–109.
- Heger, A. (2011). Jeg og jordkloden. *Dagsavisen*, Dec. 16.
- Hermansen, G. H. and Hjort, N. L. (2015). Bernstein–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach. *Journal of Statistical Planning and Inference*, 166:138–157.
- Hermansen, G. H., Hjort, N. L., and Kjesbu, O. S. (2016). Modern statistical methods applied on extensive historic data: Hjort liver quality time series 1859–2012 and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences*, 73:273–295.
- Hindemith, P. (1950). *Johann Sebastian Bach: Ein verpflichtendes Erbe: Rede zum Bachfest in Hamburg am 12. September 1950*. Insel-Verlag, Wiesbaden.
- Hjort, J. (1914). *Fluctuations in the Great Fisheries of Northern Europe, Viewed in the Light of Biological Research*. Conseil Permanent International Pour l’Exploration de la Mer, Copenhagen.
- Hjort, J. and Petersen, C. G. J. (1905). *Short review of the results of the international fisheries investigation*, volume 3. Conseil Permanent International Pour l’Exploration de la Mer, Rapport et Proces-Verbaux des Reunions, København.
- Hjort, N. L. (1976). The Dirichlet Process Applied to Nonparametric Estimation Problems. Technical report, University of Tromsø, University of Tromsø. Cand. real. thesis, Universities of Tromsø and Oslo.
- Hjort, N. L. (1985a). Bayesian nonparametric bootstrap confidence intervals. Technical report, Department of Statistics, Stanford University.
- Hjort, N. L. (1985b). Discussion contribution to P.K. Andersen and Ø. Borgan’s paper ‘Counting process models for life history data: A review’. *Scandinavian Journal of Statistics*, 12:141–151.
- Hjort, N. L. (1986a). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics*, 13:63–85.
- Hjort, N. L. (1986b). *Notes on the Theory of Statistical Symbol Recognition*. Norwegian Computing Centre, Oslo.
- Hjort, N. L. (1988a). The eccentric part of the non-central chi square. *The American Statistician*, 42:130–132.
- Hjort, N. L. (1988b). On large-sample multiple comparison methods. *Scandinavian Journal of Statistics*, 15:259–271.
- Hjort, N. L. (1990a). Goodness of fit tests for life history data based on cumulative hazard rates. *Annals of Statistics*, 18:1221–1258.
- Hjort, N. L. (1990b). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, 18:1259–1294.

- Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review*, xx:355–387.
- Hjort, N. L. (1994a). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association*, 89:665–675.
- Hjort, N. L. (1994b). Should the Olympic speedskaters run the 500 meter twice? Technical report, Department of Mathematics, University of Oslo. Also on arXiv, 2603.22298.
- Hjort, N. L. (2003). Topics in nonparametric Bayesian statistics. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*. Oxford University Press.
- Hjort, N. L. (2007). And Quiet Does Not Flow the Don: Statistical analysis of a quarrel between Nobel laureates. In Østreng, W., editor, *Conciliation*, pages 134–140. Centre for Advanced Research, Oslo.
- Hjort, N. L. (2008). Discussion of P.L. Davies’ article ‘Approximating data’. *Journal of the Korean Statistical Society*, 37:221–225.
- Hjort, N. L. (2014). Discussion of Efron’s article ‘Estimation and accuracy after model selection’. *Journal of the American Statistical Association*, 110:1017–1020.
- Hjort, N. L. (2017a). Cooling of Newborns and the Difference Between 0.244 and 0.278. *FocuStat Blog, University of Oslo*, xv.
- Hjort, N. L. (2017b). The Semifinals Factor for Skiing Fast in the Finals. *FocuStat Blog, University of Oslo*, xv.
- Hjort, N. L. (2018a). Overdispersed Children. *FocuStat Blog, University of Oslo*, xxi.
- Hjort, N. L. (2018b). Towards a More Peaceful World [insert ‘!’ or ‘?’ here]. *FocuStat Blog, University of Oslo*, xvii.
- Hjort, N. L. (2019a). The Magic Square of 33. *FocuStat Blog, University of Oslo*, xxi.
- Hjort, N. L. (2019b). Sudoku Solving by Probability Models and Markov Chains. *FocuStat Blog, University of Oslo*, xxi.
- Hjort, N. L. (2026). Counting the uncounted: How many were killed in Guatemala, 1978-1995? In de Bin, R., editor, *Proceedings of the 40th International Workshop on Statistical Modelling, Oslo, 2026*, pages 1–10. University of Oslo, Department of Mathematics, Oslo.
- Hjort, N. L. and Claeskens, G. (2003a). Frequentist model averaging [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:879–899.
- Hjort, N. L. and Claeskens, G. (2003b). Rejoinder to the discussion of the Hjort and Claeskens and Claeskens and Hjort papers. *Journal of the American Statistical Association*, 98:917–925.
- Hjort, N. L. and Fenstad, G. (1992). On the last time and the number of times an estimator is more than ε from its target value. *The Annals of Statistics*, 20:469–489.
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.
- Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24:1619–1647.
- Hjort, N. L. and Koning, A. J. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, 14:113–132.
- Hjort, N. L. and Lumley, T. (1993). Normalised local hazard plots. Technical report, Department of Statistics, University of Oxford, Oxford.
- Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics*, 37:1079–1111.

- Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica*, 27:2389–2407.
- Hjort, N. L. and Ongaro, A. (2005). Exact inference for random Dirichlet means. *Statistical Inference for Stochastic Processes*, 8:227–254.
- Hjort, N. L. and Ongaro, A. (2006). On the distribution of random Dirichlet jumps. *Metron*, LXIV:61–92.
- Hjort, N. L. and Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In Nair, V., editor, *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell Doksum*, pages 463–492. World Scientific, New Jersey.
- Hjort, N. L. and Pollard, D. B. (1993). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.
- Hjort, N. L. and Schweder, T. (2018). Confidence distributions and related themes: introduction to the special issue. *Journal of Statistical Planning and Inference*, 195:1–13.
- Hjort, N. L. and Stoltenberg, E. A. (2021). The partly parametric and partly nonparametric additive risk model. *Lifetime Data Analysis*, 27:1–31.
- Hjort, N. L. and Stoltenberg, E. A. (2026). Probability proofs for the Stirling (and more): the ubiquitous rolde of $\sqrt{2\pi}$. *American Statistician*, xx:xx–xx.
- Hjort, N. L. and Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics*, 35:64–82.
- Hjort, N. L. and Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics*, 37:105–131.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.
- Holum, D. (1984). *The Complete Handbook of Speed Skating*. High Peaks Cyclery, Lake Pacid.
- Hosmer, D. W. and Lemeshow, S. (1999). *Applied Logistic Regression*. Wiley, New York.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals Mathematical Statistics*, 53:73–101.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York.
- Hveberg, K. (2019). *Lene din ensomhet langsomt mot min*. Aschehoug, Oslo.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge.
- Inlow, M. (2010). A moment generating function proof of the Lindeberg–Lévy central limit theorem. *American Statistician*, 64:228–230.
- Irwin, J. O. (1927). On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson’s type II. *Biometrika*, 19:225–239.
- Jacod, J. and Protter, P. (2004). *Probability Essentials. Second Edition*. Springer, Berlin.
- Jacod, J. and Shiryaev, A. (2013). *Limit Theorems for Stochastic Processes. Second Edition*. Springer, Berlin.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379.
- Jamtveit, B., Jacobsen, A. U., and Wyller, T. B. (2018). Utvikling i andel administrativt personale i norske helseforetak. *Samfunnsøkonomen*, 6:17–21.

- Jamtveit, B., Jettestuen, E., and Mathiesen, J. (2009). Scaling properties of European research units. *Proceedings of the National Academy of Sciences*, 106:13160–13163.
- Jansen, D. (1994). *Full Circle*. Villard Books, New York.
- Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, 27:951–981.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Kahneman, D., Sibony, O., and Sunstein, C. R. (2020). *Noise: A Flaw in Human Judgment*. William Collins, London.
- Kallenberg, O. (2002). *Foundations of Modern Probability. Second Edition*. Springer, Berlin.
- Keener, R. W. (2011). *Theoretical Statistics: Topics for a Core Course*. Springer, New York.
- Kjesbu, O. S., Opdal, A. F., Korsbrekke, K., Devine, J. A., and Skjæraasen, J. E. (2014). Making use of Johan Hjort’s ‘unknown’ legacy: reconstruction of a 150-year coastal time-series on northeast Arctic cod (*Gadus morhua*) liver data reveals long-term trends in energy allocation patterns. *ICES Journal of Marine Science*, 71:2053–2063.
- Kjetsaa, G., Gustavson, S., Beckman, B., and Gil, S. (1984). *The Authorship of The Quiet Don [also published in Russian]*. Solum/Humanities Press, Oslo.
- Klein, R., Knudtson, M. D., Lee, K. E., Gangnon, R., and Klein, B. E. (2008). The Wisconsin epidemiologic study of diabetic retinopathy: XXII the twenty-five-year progression of retinopathy in persons with type 1 diabetes. *Ophthalmology*, 115:1859–1868.
- Klotz, J. (1972). Markov chain clustering of births by year. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability Theory*, 4:173–185.
- Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *Annals of Statistics*, 1:373–379.
- Koehler, J. J. and Conley, C. A. (2003). The “hot hand” myth in professional basketball. *Journal of Sport and Exercise Psychology*, 25:253–259.
- Kolmogorov, A. N. (1933a). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer, Berlin.
- Kolmogorov, A. N. (1933b). Sulla determinazione empirica di una legge di distribuzione. *Giorn Ist Ital Attuar*, 4:83–91.
- Kusolitsch, N. (2010). Why the theorem of Scheffé should be rather called a theorem of Riesz. *Periodica Mathematica Hungarica*, 61:225–229.
- Laake, P., Laake, K., and Aaberge, R. (1985). On the problem of measuring the distance between distribution functions: Analysis of hospitalization versus mortality. *Biometrics*, 41:515–523.
- Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les évènements. Technical report, Mémoires de l’Académie Royale des Sciences, Présenté par Divers Savan, vol. 6, Paris.
- Laptook, A. e. a. (2017). Effect of therapeutic hypothermia initiated after 6 hours of age on death and disability among newborns with hypoxic-ischemic encephalopathy: A randomized clinical trial. *Journal of the American Medical Association*, 318:1550–1560.
- Larkey, P. D., Smith, R. A., and Kadane, J. B. (1989). It’s okay to believe in the “hot hand”. *Chance*, 2:22–30.
- Larsen, R. L. and Marx, M. L. (1981). *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall, New York.
- Le May Doan, C. (2002). *Going For Gold*. McClelland & Stewart Publisher, Toronto.
- LeCam, L. (1986). The Central Limit Theorem around 1935. *Statistical Science*, 1:78–91.

- Lehmann, E. L. (1950). *Notes on the Theory of Estimation*. Berkeley University Press, Berkeley. Notes recorded by Colin Blyth.
- Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.
- Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.
- Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer-Verlag, Berlin.
- Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation [2nd ed.]*. Springer, Berlin.
- Leike, A. (2001). Demonstration of the exponential decay law using beer froth. *European Journal of Physics*, 23:1–21.
- Lessing, D. (1997). *Walking in the Shade: Volume Two of My Autobiography, 1949 to 1962*. xx, xx.
- Li, G., Reid, N., and Varin, C. (2027). *Likelihood and Its Extensions*. CRC Press, New York.
- Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15:211–225.
- Lindley, D. V. (1972). *Bayesian Statistics: A Review*. Society for Industrial and Applied Mathematics, Philadelphia.
- Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model [with discussion]. *Journal of the Royal Statistical Society Series B*, 34:1–41.
- Lindqvist, B. H. (1978). A note on Bernoulli trials with dependence. *Scandinavian Journal of Statistics*, 5:205–208.
- Linero, A. and Rosalsky, A. (2013). On the Toeplitz lemma, convergence in probability, and mean convergence. *Stochastic Analysis and Applications*, 31:684–694.
- Lum, K., Price, M. E., and Banks, D. (2013). Applications of multiple systems estimation in human rights research. *American Statistician*, 24:191–200.
- Mardia, K. V., Kent, J. T., and Taylor, C. C. (1979). *Multivariate Analysis*. Academic Press, London.
- Markov, A. A. (1906). Распространение закона больших чисел на величины, зависящие друг от друга [Extending the law of large numbers for variables that are dependent of each other]. *Известия Физико-математического общества при Казанском университете (2-я серия)*, 15:124–156.
- Markov, A. A. (1913). Пример статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь [Example of a statistical investigation illustrating the transitions in the chain for the ‘Evgenii Onegin’ text]. *Известия Академии Наук, Санкт-Петербург (6-я серия)*, 7:153–162.
- Marshall, E. C. and Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, 2:409–444.
- McCullagh, P. (2002). What is a statistical model? [with discussion]. *Annals of Statistics*, 30:1225–1310.
- McCullagh, P. (2022). *Ten Projects in Applied Statistics*. Springer, Berlin.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models [2nd. ed.]*. Chapman and Hall, London.
- Miller, J. B. and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86:2019–2047.
- Miller, J. B. and Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the NBA three-point contest? *European Economic Review*, 138:103771.

- Mykland, P. A. and Zhang, L. (2012). The econometrics of high frequency data. In Kessler, M., Lindner, A., and Sørensen, M., editors, *Statistical Methods for Stochastic Differential Equations*, pages 109–190. CRC Press.
- Mykland, P. A., Zhang, L., and Chen, D. (2019). The algebra of two scales estimation, and the S-TSRV: High frequency estimation that is robust to sampling times. *Journal of Econometrics*, 208:101–119.
- Newton, M. J. (2002). A conversation with Emanuel Parzen. *Statistical Science*, 17:357–378. Correction, op. cit., 467.
- Neyman, J. and Pearson, E. (1933). On the problem of the most efficient statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 68:289–337.
- Normand, S.-L. T. (1999). Tutorial in biostatistics: Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18:321–359.
- Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.
- Parzen, E. (1979). Nonparametric statistical data modeling [with discussion]. *Journal of the American Statistical Association*, 74:101–131.
- Parzen, E. (1982). Data modeling using quantile and density-quantile functions. In *Some Recent Advances in Statistics: Symposium Lisbon 1980*, pages 23–52. xx.
- Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science*, 19:652–662.
- Pearson, K. (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 Series*, 5(302):157–175.
- Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, 1:261–264.
- Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6:5–84.
- Peterson, A. V. (1975). Nonparametric estimation in the competing risks problem. Technical report, Department of Statistics, Stanford University.
- Phadia, E. G. (1973). Minimax estimation of a cumulative distribution function. *Annals of Statistics*, 6:1149–1157.
- Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books, Toronto.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
- Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 11:453–466.
- Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletins of the Calcutta Mathematical Society*, pages 81–91.
- Reeves, R. V. (2022a). *Of Boys and Men: Why the Modern Male is Struggling, Why it Matters, and What to Do About It*. Brookings Institution Press, Washington, D.C.
- Reeves, R. V. (2022b). Redshirt the boys. *The Atlantic*, October.
- Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, Berlin.
- Robert, C. and Casella, G. (2011). A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26:102–115.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

- Royden, H. L. and Fitzpatrick, P. M. (2010). *Real Analysis [4th ed.]*. Pearson Education Asia, Beijing.
- Rubin, H. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9:130–134.
- Rydén, J. (2020). On features of fugue subjects: A comparison of J.S. Bach and later composers. *Journal of Mathematics and Music*, pages 1–20.
- Sala-i Martin, X. (1997). I just ran two million regressions. *The American Economic Review*, 87:179–183.
- Saleh, J. H. (2019). Statistical reliability analysis for a most dangerous occupation: Roman emperor. *Palgrave Communication*, 5:1–7.
- Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43:142–1542.
- Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18:434–438.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer, New York.
- Schömig, A., Mehili, J., de Waha, A., Seyfarth, M., Pahce, J., and Kastrati, A. (2008). A meta-analysis of 17 randomized trials of a percutaneous coronary intervention-based strategy in patients with stable coronary artery disease. *Journal of the American College of Cardiology*, 52:894–904.
- Schweder, T. (1980). Scandinavian statistics, some early lines of development. *Scandinavian Journal of Statistics*, 7:113–129.
- Schweder, T. (1999). Early statistics in the Nordic countries – when did the Scandinavians slip behind the British? *Bulletin of the International Statistical Institute*, 58:1–4.
- Schweder, T. and Hjort, N. L. (2002). Likelihood and confidence. *Scandinavian Journal of Statistics*, 29:309–322.
- Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.
- Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.
- Sethuraman, J. and Tiwari, R. C. (1982). Convergence of Dirichlet measures and interpretation of their parameter. *Statistical Decision Theory and Related Topics, III*, 2:305–315.
- Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, 1:185–202.
- Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimation. *Journal of the American Statistical Association*, 80.
- Shiryayev, A. N. (1996). *Probability. Second edition*. Springer, Berlin.
- Shumway, R. H. and Stoffer, D. S. (2016). *Time Series Analysis and Its Applications [4th ed.]*. Springer, Heidelberg.
- Silver, N. (2012). *The Signal and the Noise: Why so Many Predictions Fail, but Some Don't*. Penguin.
- Simpson, R. J. S. and Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3:1243–1246.
- Sims, C. A. (2012a). Appendix: inference for the Haavelmo model. Technical report, Public Policy & Finance, Princeton University, Princeton, NJ.
- Sims, C. A. (2012b). Statistical modeling of monetary policy and its effects [Sveriges Riksbank Prize in Memory of Alfred Nobel lecture]. *American Economic Review*, xx:1–22.

- Singh, K., Xie, M., and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Annals of Statistics*, 33:159–183.
- Slud, E. (1989). Clipped Gaussian processes are never M-step Markov. *Journal of Multivariate Analysis*, 29:1–14.
- Smith, T. D. (1994). *Scaling Fisheries: The Science of Measuring the Effects of Fishing 1855–1955*. Cambridge University Press, Cambridge.
- Spiegelberg, W. (1901). *Aegyptische und Griechische Eigennamen aus Mumientiketten der Römischen Kaiserzeit*. Greek Inscriptions, Cairo.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206.
- Stigler, S. M. (1973). Studies in the history of probability and statistics. xxxii: Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika*, 60:439–445.
- Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, 27:1055–1098.
- Stigler, S. M. (1983). Who discovered Bayes’s Theorem? *American Statistician*, 37:290–296.
- Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.
- Stigler, S. M. (2006). How Ronald Fisher became a mathematical statistician. *Mathematics and Social Sciences*, 44:23–30.
- Stoltenberg, E. A. (2019). An MGF proof of the Lindeberg theorem. Technical report, Department of Mathematics, University of Oslo.
- Stoltenberg, E. A. and Hjort, N. L. (2020). Multivariate estimation of Poisson parameters. *Journal of Multivariate Analysis*, 175:1–19.
- Stoltenberg, E. A. and Hjort, N. L. (2021). Models and inference for on-off data via clipped Ornstein–Uhlenbeck processes. *Scandinavian Journal of Statistics*, 48:908–929.
- Stout, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.
- Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.
- Tronstad, E. (2025). Window of change: Extensions of change-point analysis. Technical report, University of Oslo. Master Thesis.
- Tversky, A. and Gilovich, T. (1989). The cold facts about the “hot hand” in basketball. *Chance*, 2:16–21.
- Van Bavel, K., Moreels, S., Van de Putte, B., and Matthijs, K. (2011). Family size and intergenerational social mobility during the fertility transition: Evidence of resource dilution from the city of Antwerp in nineteenth century Belgium. In *Social Mobility and Demographic Behaviour: A Long-Term Perspective*, volume 24, pages xx–xx. Holland Holland.
- Varian, H. R. (1975). Distributive justice, welfare economics, and the theory of fairness. *Philosophy and Public Affairs*, 4:223–247.
- Voldner, B., Frøslie, K. F., Haakstad, L., Hoff, C., and Godang, K. (2008). Modifiable determinants of fetal macrosomia: role of lifestyle-related factors. *Acta Obstetrica et Gynecologica Scandinavica*, 87:423–429.
- von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Annals of Mathematical Statistics*, xx:808–818.
- von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*. B.G. Teubner, Berlin.
- Walløe, L., Hjort, N. L., and Thoresen, M. (2019a). Major concerns about late hypothermia study. *Acta Paediatrica*, 108:588–589.

- Walløe, L., Hjort, N. L., and Thoresen, M. (2019b). Why results from Bayesian statistical analyses of clinical trials with a strong prior and small sample sizes may be misleading: The case of the NICHD Neonatal Research Network Late Hypothermia Trial. *Acta Paediatrica*, 108:1190–1191.
- Wardrop, R. L. (1995). Simpson’s paradox and the hot hand in basketball. *The American Statistician*, 49:24–28.
- Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge.
- Wissner-Gross, Z. (2020). Can you feed the hot hand? <https://fivethirtyeight.com/features/can-you-feed-the-hot-hand/>. Accessed: December 12, 2020.
- Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review [with discussion and a rejoinder]. *International Statistical Review*, 81:3–39.
- Zabriskie, B. N., Corcoran, C., and Senchaudhuri, P. (2021). A comparison of confidence distribution approaches for rare event meta-analysis. *Statistics in Medicine*, 40:5276–5297.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81:446–451.
- Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411.

Name index

DeGroot, Morris H, 353

Sverdrup, E., 271

Subject index

action space, 281
admissibility, 282
Arzelà–Ascoli theorem, 323

Bahadur’s theorem, 288
Bayes risk, 216, 282
Bayes solution, 216, 283
Blyth’s method, 293

complete metric space, 314
completeness, 287
conditionally identically distributed, 233

Daniell extension theorem, 321

empirical Bayes, 240

finite-dimensional distributions, 319

improper prior, 290, 292

Lehmann–Scheffé theorem, 287
loss function, 281

Minimax, 282
minimum Bayes risk, 216
modulus of continuity, 324

probability kernel, 321
projective property, 321
Prokhorov’s theorem, 321

Rao–Blackwell theorem, 286
regular conditional probability, 321
relatively compact, 320, 323
risk, 353
risk function, 282

sequentially compact, 323
squared error loss, 282

tightness, 321
Type I and Type II errors, 282

unbiased estimator, 282
uniformly equicontinuous, 323
uniformly minimum variance unbiased estimator, 285