

Statistical Inference: 600 Exercises, 100 Stories, and Solutions

Nils Lid Hjort and Emil Aas Stoltenberg

Department of Mathematics, University of Oslo

Here: Preface, Chapters 1, 2, 3, 4, 5, 6

– 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

I.1

Statistical models

In this chapter we study families of distributions and densities that we are to meet time and again in this book. A partial list includes the uniform, normal and multinormal, chi-squared, the t and the F, Gamma, exponential, Weibull, Beta, Dirichlet, Poisson, compound Poisson, binomial, multinomial, geometric, Pareto, Gumbel, logistic. These families have parameters, with values to be set for certain studies or illustrations, or for purposes of confidence setting and tests; more generally these parameters are estimated from data, as we return to in several later chapters. We also learn fruitful ways of extending and mixing given families of distributions; in such fashions the classical models can be building blocks for forming new ones. Mathematical techniques for deriving crucial properties include those of moment-generating functions, convolutions, and double expectation.

Key words: distributions, models, moments, moment-generating functions, parameters, quantiles, sums, transformations

The aim of this chapter is to go through a generous list of parametric statistical models, from the well-known distributions connected with the normal model, to the Beta and the Gamma, to the binomial, Poisson, and negative binomial for discrete data, etc., along with deriving their basic properties. These models turn up repeatedly in later chapters and in our Statistical Stories, with variations, as direct models for data, or as building blocks for more complicated constructions. The normal and multinormal distributions play important roles, also because these become fruitful simple-to-use approximations to sometimes much more complicated exact distributions.

These models, for probability theory and statistics, rely on deeper mathematical constructions and considerations, with random variables being measurable functions on probability spaces, measure and integration theory, etc. For this book it has been practical to organise that body of mathematical theory in Appendix B. For the present chapter on models we take certain notions and basic definitions for granted, with background and more detail in that appendix. Thus we deal here with classes of distributions, parameters, probability densities, cumulative distribution functions, conditional and marginal distributions, means and variances, quantiles, correlations, and so on. In particular, any nonnegative function $f(y)$ integrating to 1 over some interval is a probability density over that interval; if Y is drawn from this f it has a cumulative distribution function (c.d.f.) $F(y) = \Pr(Y \leq y)$, which in the continuous case is the same as $\int_{-\infty}^y f(y') dy'$; the mean of a random variable Y drawn from this distribution is $EY = \int yf(y) dy$; its median is $F^{-1}(\frac{1}{2})$; its variance $\text{Var } Y = E(Y - EY)^2$; the covariance between two random

c.d.f.

variables X and Y with means a and b is $\text{cov}(X, Y) = E(X - a)(Y - b) = E(XY) - ab$; etc. We also deal with sums of random variables, drawn from the same or different distributions. In one of its classical forms, the density $h(z)$ of $Z = X + Y$, where X and Y are independent with densities $f(x)$ and $g(y)$, is the convolution formula

$$h(z) = \int f(z - y)g(y) dy = \int f(x)g(z - x) dx, \quad (1.1)$$

see Ex. C.7 for more. There are also multiple occasions in this chapter where the rules for double expectation and variance, i.e.

$$E X = E\{E(X | Y)\} \quad \text{and} \quad \text{Var } X = E\{\text{Var}(X | Y)\} + \text{Var}\{E(X | Y)\}, \quad (1.2)$$

see Ex. C.10, come in handy.

In addition to defining and presenting a list of useful models, and diving into their properties and inter-connections, we develop certain tools, useful also in later chapters. These include transformations (see Ex. 1.12), moment-generating functions (see Ex. 1.32, with more in Ex. C.21), characteristic functions (see Ex. 1.36), conditional distributions, mixtures, and simulation. Also included is material on the general exponential family class, which has several of the classic models as special cases (see Ex. 1.57, with follow-up material in Ch. 4).

Importantly, several of the central models worked with in this introduction chapter find uses inside wider contexts, e.g. for regression situations, as we shall return to in later chapters. To indicate that direction of model building, below we learn about a random variable Y having a gamma distribution with parameters (a, b) , which we write as $Y \sim \text{Gam}(a, b)$; see Ex. 1.9. Now suppose there is a dataset consisting of measurements $(x_{i,1}, x_{i,2}, y_i)$ for individuals $i = 1, \dots, n$, where the main outcome Y_i is influenced by the covariates $x_{i,1}, x_{i,2}$. Then a gamma regression model could take the form $Y_i | (x_{i,1}, x_{i,2}) \sim \text{Gam}(a_i, b)$, with $a_i = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2})$. The traditional multiple linear regression model is of a similar type, with Y_i given the covariates having normal distributions, with mean function linear in the covariates. Tools developed in later chapters may then be applied to estimate parameters, with confidence intervals, testing, comparisons, prediction, etc.

Normal, bi- and multinomial, exponential, gamma, mixing

Ex. 1.1 *The normal distribution.* The perhaps most famous and broadly useful distribution in probability theory and statistics is the normal distribution, also called the Gaußian distribution. It is also a building block for various inferred and related models and distributions, as we learn later in the chapter. In its standard form, before we add on two more parameters, the *normal density* is

Gauß

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2) \quad \text{on the real line.}$$

We call this the standard normal distribution, and write $X \sim N(0, 1)$ to indicate this. It is standard in statistics and probability theory to use $\phi(x)$ for its density and $\Phi(x)$ for its cumulative distribution function (c.d.f.).

(a) There are myriad ways of demonstrating that $1/(2\pi)^{1/2}$ is the correct constant here, i.e. that $I = \int \exp(-\tfrac{1}{2}x^2) dx = (2\pi)^{1/2}$. You are allowed to take this for granted, but

attempt to show it via expressing I^2 as a double integral, featuring $\exp\{-\frac{1}{2}(x^2 + y^2)\}$, and then substituting $x = r \cos \theta$ and $y = r \sin \theta$, followed by the use of double integration tools from calculus.

(b) Show that for X a standard normal, its mean is zero and its variance is one.

(c) With X a standard normal, consider $Y = \xi + \sigma X$, with ξ any number and σ positive. Show that its mean and standard deviation are μ and σ , and that its density can be written

$$f(y) = \phi\left(\frac{y - \xi}{\sigma}\right) \frac{1}{\sigma} = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y - \xi}{\sigma}\right)^2\right\}.$$

We write $Y \sim N(\xi, \sigma^2)$ to indicate this distribution. Show that $\Pr(\xi - 1.96\sigma \leq Y \leq \xi + 1.96\sigma) = 0.95$. Find the c such that $\Pr(|Y - \xi| \leq c\sigma) = 0.50$.

(d) With X a standard normal, consider $Z = X^2$. Find its c.d.f., and show that its density becomes $g(z) = (2\pi)^{-1/2} \exp(-\frac{1}{2}z)/\sqrt{z}$. We learn about the chi-squared distribution in Ex. 1.48; this X^2 has such a chi-squared distribution, with degrees of freedom equal to 1, which we write as $X^2 \sim \chi_1^2$.

(e) Consider X_1, X_2, X_3 being independent and standard normal. Work out the means and variances of X_1^2 , $X_1^2 + X_2^2$, $X_1^2 + X_2^2 + X_3^2$. Simulate say 10^4 realisations of these distributions, check their histograms, and describe their different behaviour close to zero.

(f) Consider the enigmatic density $f(x) = e^{-\pi x^2}$ on the real line, featuring and combining the eternal mathematical constants e and π , integrating to 1. What is its standard deviation, and what is the probability that an X with this distribution is inside $[-1, 1]$?

(g) A statistician is helping clients to predict the future outcome Y of some variable. The predictions $\hat{Y}$ are such that $\hat{Y} - Y$ is approximately a zero-mean normal with standard deviation 5.0. What is the probability that the prediction is inside $(-1, 1)$ of the future outcome?

Ex. 1.2 Normal sums. Sums of independent normals have themselves normal distributions. This is clearly easiest to demonstrate via moment-generating functions (m.g.f.s), see Ex. 1.33 below, but it is worth doing this via convolution formulae too.

(a) Let X and Y be independent standard normals. Show that $X + Y \sim N(0, 2)$, via the convolution formula (1.1). With a bit more algebraic work, show that if $X_1 \sim N(\xi_1, \sigma_1^2)$ and $X_2 \sim N(\xi_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim N(\xi_1 + \xi_2, \sigma_1^2 + \sigma_2^2)$.

(b) Generalise this: show that if $X_i \sim N(\xi_i, \sigma_i^2)$ for $i = 1, \dots, m$, and these are independent, then $Z = \sum_{i=1}^m a_i X_i$ is also normal, with mean $\sum_{i=1}^m a_i \xi_i$ and variance $\sum_{i=1}^m a_i^2 \sigma_i^2$.

(c) Sometimes $2 + 2$ might be 5, for extremely high values of 2. If independent X and Y are 2 and 2, but observed with with some Gaussian noise on top of standard deviation level σ , find a formula $p(\sigma)$ for the probability that $X + Y$ is outside $[3.5, 4.5]$, i.e. 3 or 5 or even farther away from 4 when rounded off to the nearest integer. Plot this function.

Ex. 1.3 Binomial distribution. One of the more old, classic, and deservedly famous distributions in probability and statistics is the binomial. If there is a fixed probability $p = P(A)$ of a certain event A taking place, in a certain type of experiment, then the number Y of times A is seen, in n independent experiments, is the *binomial*, which we write as $Y \sim \text{binom}(n, p)$.

(a) Show that

$$\Pr(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \quad \text{for } y = 0, 1, \dots, n.$$

Gauß

the binomial
distribution

This involves the essential combinatorial fact that the number of ways one may place precisely y 1s in a total of n possible position is $\binom{n}{y} = n!/\{y!(n-y)!\}$. Explain that Y can be expressed as $X_1 + \dots + X_n$, where X_i is a simple 0-1 variable, with $\Pr(X_i = 1) = p$, and where these are independent. Such X_i are called *Bernoulli variables*. Use this to prove the classic formulae np and $np(1-p)$ for mean and variance. Also, deduce the $\Pr(Y = y)$ formula from the $Y = \sum_{i=1}^n X_i$ description.

Bernoulli
variables

(b) If the first question to ask concerning a distribution is about its centre (its mean, or perhaps its median), and the second is about its spread (its standard deviation, or perhaps a different measure, like its interquartile range), then the third question would be about its skewness, the degree of asymmetry. The classical skewness definition of a distribution, or equivalently of a random variable Y having that distribution, is $\text{skew} = E W^3$, where $W = (Y - EY)/(\text{Var } Y)^{1/2}$ is the normalised version of Y , i.e. linearly transformed to have mean zero and standard deviation one. Show for the binomial (n, p) case that its skewness is

the skewness

$$\text{skew} = E \left[\frac{Y - np}{\{np(1-p)\}^{1/2}} \right]^3 = \frac{1-2p}{\{np(1-p)\}^{1/2}},$$

which for fixed p goes to zero with rate $1/\sqrt{n}$ (i.e. $\sqrt{n}$ skew tends to a positive constant). Briefly discuss what this entails regarding the degree of asymmetry for the binomial distribution.

(c) After the skewness comes the so-called kurtosis, defined as $\text{kurt} = E W^4 - 3$, with W as in the previous point. The minus 3 is there in order for the kurtosis to be zero for the normal distribution; show that this is the case. Then show that

the kurtosis

$$\text{kurt} = (1/n)[1/\{p(1-p)\} - 6],$$

for the binomial, and comment.

Ex. 1.4 *Trinomial probabilities.* Consider the so-called *trinomial distribution* for a random pair (X, Y) , with probability mass function

$$f(x, y) = \frac{n!}{x! y! (n-x-y)!} p^x q^y (1-p-q)^{n-x-y} \quad \text{for } x \geq 0, y \geq 0, x+y \leq n.$$

Here n is the total count number, p, q the probabilities of events of type One and Two in repeated experiments, with $p+q < 1$. With $Z = n - X - Y$ representing the number of events of type Three (not One, not Two), this is a model for the numbers of events One, Two, Three in n independent experiments; hence the trinomial name. See also Ex. 1.5.

(a) Verify that what is here called the probability mass function is the same as the density of the distribution with respect to counting measure on the set of (x, y) with $x \geq 0, y \geq 0, x+y \leq n$ – or, for that matter, with respect to counting measure on the set of all pairs (x, y) with $x \geq 0, y \geq 0$.

(b) Show by summing over the y that the distribution of $X \sim \text{binom}(n, p)$.

(c) Show that $Y | (X = x) \sim \text{binom}(n-x, q/(1-p))$. Give a formula for $E(Y | X = x)$, and deduce the formula for EX from this. Find also the covariance between X and Y , using this scheme of conditioning with respect to $X = x$ first. Deduce that the correlation between them is $-\{p/(1-p)\}^{1/2}\{q/(1-q)\}^{1/2}$.

(d) Find a formula for $\Pr(X \leq x_0, Y \leq y_0)$, expressed as a sum over $x \in \{0, 1, \dots, x_0\}$ (as opposed to a double sum over lots of (x, y) pairs). For a setup with $n = 50$, $(p, q) = (0.22, 0.33)$, compute the probability $\Pr(X \leq 15, Y \leq 20)$.

Ex. 1.5 *The multinomial model.* The binomial model, with basic properties treated in Ex. 1.3, is about sorting and counting events in two categories; if $Y \sim \text{binom}(n, p)$, then also $n - Y \sim \text{binom}(n, 1 - p)$. The *multinomial model* is the natural extension of also Ex. 1.4 to more than two and three categories. Suppose there are n independent experiments, where each time one (and only one) of the events $A_1, \dots, A_k$ takes place, with the same probabilities $p_1, \dots, p_k$ for each experiment. Let then $Y = (Y_1, \dots, Y_k)$, with Y_j counting the number of times A_j occurred, for $j = 1, \dots, k$. Of course $Y_1 + \dots + Y_k = n$, and $p_1 + \dots + p_k = 1$, so there are $k - 1$ free parameters in the model.

(a) Show that $Y_j \sim \text{binom}(n, p_j)$, and deduce that we already know $EY_j = np_j$ and $\text{Var } Y_j = np_j(1 - p_j)$, even before we start working on the joint distribution of $(Y_1, \dots, Y_k)$.

the multinomial
model

(b) Show that the joint probability distribution becomes

$$f(y_1, \dots, y_k) = \Pr(Y_1 = y_1, \dots, Y_k = y_k) = \frac{n!}{y_1! \dots y_k!} p_1^{y_1} \dots p_k^{y_k}$$

for nonnegative $(y_1, \dots, y_k)$ with sum n . The first factor $n!/(y_1! \dots y_k!)$ is a combinatorial one, the number of different ways one may place ‘1’ in y_1 positions, ‘2’ in y_2 positions, etc., up to ‘ k ’ in y_k positions. Note that this generalises the classic $n!/(y_1! y_2!) = \binom{n}{y_1}$ for the binomial case, the number of ways one may place ‘1’ in y_1 ways (and hence ‘2’ in $n - y_1$ ways) in a list $1, \dots, n$.

(c) Show that each pair has a trinomial distribution, e.g.

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \frac{n!}{y_1! y_2! (n - y_1 - y_2)!} p_1^{y_1} p_2^{y_2} (1 - p_1 - p_2)^{n - y_1 - y_2}$$

for $y_1 \geq 0, y_2 \geq 0, y_1 + y_2 \leq n$. Explain that formulae from Ex. 1.4 therefore apply to pairs (Y_i, Y_j) here. For $i \neq j$, show that $\text{cov}(Y_i, Y_j) = -np_i p_j$, and find the correlation between Y_i and Y_j .

(d) Among the most used acronyms of statistical parlance is *i.i.d.*, for independent and identically distributed. Explain in the present setup that $Y = Z_1 + \dots + Z_n$, where $Z_1, \dots, Z_n$ are i.i.d., with Z_i taking values $(1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$ with probabilities $p_1, \dots, p_k$. Derive again the formulae for means, variances, covariances, starting with this representation.

i.i.d.

Ex. 1.6 *Histograms.* Suppose data $Y_1, \dots, Y_n$ are i.i.d. from some continuous density f over some interval $[a, b]$. Create disjoint cells $C_1, \dots, C_k$, with $C_j = (a_{j-1}, a_j]$, for $a = a_0 < \dots < a_k = b$. Let N_j count the number of data points in cell j . The *histogram*, associated with the chosen cells, is then

histogram

$$\hat{f}(x) = \hat{p}_j / h_j \quad \text{for } x \in C_j, \text{ with lengths } h_j = a_j - a_{j-1},$$

where $\hat{p}_j = N_j/n$ estimates $p_j = \Pr(Y_i \in C_j)$. In most automatic histogram algorithms, the width is taken constant across cells. The notation $\hat{f}(x)$ indicates that beyond being an effective way of showing the essential spread and shape of the data, it is *an estimate* of the underlying $f(x)$.

for better
methods, check
Ex. 3.42

(a) Carry out some simulations, sampling n datapoints from the standard normal, creating histograms with k cells. For small and big n , play with k small, big, and about right. In R, you may use `hist(y, breaks=20, prob=T)`, etc.

(b) Show that $(N_1, \dots, N_k)$ is multinomial. Taking a constant cell width h , for simplicity, explain that we may write $p_j = \int_{C_j} f \, dy = f(\zeta_j)h$, for a suitable ζ_j inside

cell C_j . From this derive that for $x \in C_j$, we have $E\hat{f}(x) = f(\zeta_j)$ and $\text{Var}\hat{f}(x) = f(\zeta_j)\{1 - f(\zeta_j)h\}/(nh)$. Argue that for the histogram to achieve $E\hat{f}(x) \rightarrow f(x)$ and $\text{Var}\hat{f}(x) \rightarrow 0$, as n increases, we need $h \rightarrow 0$ and $nh \rightarrow \infty$.

Ex. 1.7 Hazard rates and survival functions. Here and below we shall partly follow the implied tradition of using say T and $f(t)$ and $F(t)$, for random variables with their densities and c.d.f.s, rather than say Y and $f(y)$ and $F(y)$, when these relate to *time*. – Consider a random variable T on the halflife $[0, \infty)$, with density f and c.d.f. F . Classes of such distributions are sometimes most conveniently or fruitfully defined and discussed in terms of their hazard or cumulative hazard functions, as opposed to their densities and c.d.f.s, as we outline here; see also Ch. 11.

(a) For any nonnegative random variable T , the survival curve, or function, is $S(t) = \Pr(T \geq t)$. For T continuous, $S(t) = 1 - F(t)$, but explain that in the general case, we need $S(t) = 1 - F(t-)$, with $F(t-)$ the left hand limit of F at position t . Show that integrated survival is equal to mean survival:

$$ET = \int_0^\infty S(t) dt = \int_0^\infty t dF(t). \quad (1.3)$$

(You may write the latter as $\int_0^\infty \int_0^\infty I(s < t) ds dF(t)$ and then use the Fubini theorem.) Show from this that $ET^p = \int_0^\infty \{1 - F(t^{1/p})\} dt$, for any positive p .

(b) Show that

hazard rate
function

$$\Pr(T \in [t, t + \varepsilon] | T \geq t) = h(t)\varepsilon + O(\varepsilon^2) \quad \text{for small } \varepsilon, \quad (1.4)$$

in terms of the so-called *hazard rate* function $h(t) = f(t)/\{1 - F(t)\}$. With T interpreted as the time to a certain event, the function $h(t)$ describes the chance of this event taking place in the next instance, among those having survived up to t .

(c) So we may deduce the hazard rate from the density. Starting instead with $h(t)$, define first *the cumulative hazard* $H(t) = \int_0^t h(s) ds$, and show that $F(t) = 1 - \exp\{-H(t)\}$. The function $S(t) = \Pr(T \geq t) = \exp\{-H(t)\}$ is important in its own right, and is called *the survival function*.

(d) Suppose an individual has survived up to time t_0 . Show that

$$\Pr(T \geq t | T \geq t_0) = \frac{S(t)}{S(t_0)} = \exp[-\{H(t) - H(t_0)\}] \quad \text{for } t \geq t_0.$$

Show that the median lifetime, for such an individual having lived up to t_0 , is $t^* = H^{-1}(H(t_0) + \log 2)$.

(e) Suppose an event takes place at T as soon as one of two independent continuously distributed T_1 and T_2 lifetimes occur. With these having cumulative hazards H_1 and H_2 , show that $\Pr(T \geq t) = \exp\{-H_1(t) - H_2(t)\}$. Explain that this entails additive hazards, with $h(t) = h_1(t) + h_2(t)$.

additive
hazards

Ex. 1.8 The exponential distribution. The *exponential distribution* is a simple but important one, in probability theory and statistics, which with positive parameter θ has the density $f(t, \theta) = \theta \exp(-\theta t)$ for $t > 0$. We write $T \sim \text{Expo}(\theta)$ to indicate this.

the exponential
distribution

(a) Show that the c.d.f. becomes $F(t, \theta) = 1 - \exp(-\theta t)$, and find the median. Show also that we may write $T = T_0/\theta$, where T_0 has the unit exponential distribution with density $\exp(-t_0)$. Show that T has mean and variance $1/\theta$ and $1/\theta^2$.

(b) Using Ex. 1.7, show that the hazard rate is constant, $h(t) = \theta$, and that the cumulative hazard rate is $H(t) = \theta t$. Show also that the exponential distribution is the only one where the hazard rate is constant.

(c) Show that the median survival time is $(\log 2)/\theta$. If an individual has survived up to time t_0 , what is the median survival time?

the memoryless property

(d) Assume certain light bulbs have a longevity distribution with the property that $\Pr(T \geq t_0 + t | T \geq t_0)$ does not depend on t_0 . Argue that such light bulbs may be sold as if they were brand new, as long as they are still alive. Show that their distribution must be exponential.

(e) Consider $T = \min(T_1, \dots, T_k)$, with these being independent exponentials with parameters $\theta_1, \dots, \theta_k$. Show that T is also exponential, with hazard $\theta_1 + \dots + \theta_k$.

the gamma function

Ex. 1.9 The Gamma distribution. The gamma function is important in various branches in mathematics, probability theory, and statistics, and is defined as $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x) dx$ for a positive. Verify that $\Gamma(1) = 1$, that $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$, and that $\Gamma(m) = (m-1)!$ for $m = 1, 2, \dots$. We may hence define a family of probability densities via $g_0(t, a) = \Gamma(a)^{-1} t^{a-1} \exp(-t)$ for $t > 0$. This is called the *Gamma distribution* with shape parameter a .

(a) With T_0 having this density, and b a positive scale parameter, show that $T = T_0/b$ has density

$$g(t, a, b) = \{b^a / \Gamma(a)\} t^{a-1} \exp(-bt) \quad \text{for } t > 0.$$

the Gamma distribution

This is the two-parameter $\text{Gam}(a, b)$ distribution.

(b) When T has the $\text{Gam}(a, b)$ distribution, show that the mean and variance are a/b and a/b^2 . Find also that $E T^p = \{\Gamma(a+p)/\Gamma(a)\}/b^p$, valid for any p , as long as $p > -a$. Use this to show that the skewness and kurtosis become equal to skew = $2/a^{1/2}$ and kurt = $6/a$. Finally, regarding moments, find that the inverse gamma distributed variable $1/T$ has mean $b/(a-1)$ and finite variance $b^2/\{(a-1)^2(a-2)\}$, as long as $a > 2$.

(c) Verify that for $a = 1$ we have the exponential distribution, with density $b \exp(-bt)$ and cumulative $1 - \exp(-bt)$. Show that $a = 2$ gives density $b^2 t \exp(-bt)$ and cumulative $1 - \exp(-ty)(1 + bt)$. More generally, show that the cumulative is

$$\int_0^t g(s, a, b) ds = 1 - \exp(-bt) \left\{ 1 + bt + \frac{(bt)^2}{2!} + \dots + \frac{(bt)^{a-1}}{(a-1)!} \right\}$$

for the case of a being an integer.

(d) For a an integer, give an explicit expression for the hazard function $h(t, a, b)$, as per (1.4), and show that it converges to b as time increases. Show that this is the case also for any a , i.e. not only for integers; it increases from zero to b , if $a > 1$, and decreases from infinity to b , if $a < 1$.

(e) Let T_1, T_2 be independent and exponential with the same θ . Show that $T_1 + T_2 \sim \text{Gam}(2, \theta)$. With $T_1, \dots, T_k$ seen as the independent waiting times between events, show that the time to event k is a $\text{Gam}(k, \theta)$.

(f) With $T_1 \sim \text{Gam}(a_1, b)$ and $T_2 \sim \text{Gam}(a_2, b)$ independent, show that $T_1 + T_2 \sim \text{Gam}(a_1 + a_2, b)$. Generalise. This may indeed be accomplished via the convolution formulae from Ex. C.7, but as for other instances it becomes easier to prove such statements via m.g.f.s; see Ex. 1.32–1.33.

Ex. 1.10 *Mixing the exponential.* Sometimes waiting time type data do not follow an exact exponential distribution, but rather one characterised as a mixture of such; T given θ has the $\text{Expo}(\theta)$ distribution, but the values of θ vary from occasion to occasion, or from individual to individual.

(a) Suppose indeed that $T|\theta \sim \text{Expo}(\theta)$ but that θ has some density $g(\theta)$. Show that the density of T then becomes $f(t) = \int_0^\infty \theta \exp(-\theta t) g(\theta) d\theta$.

(b) Suppose the distribution of θ is such that $1/\theta$ has mean value $1/\theta_0$ and a positive standard deviation τ . Show, starting with $E(T|\theta) = 1/\theta$ and $\text{Var}(T|\theta) = 1/\theta^2$, that $ET = 1/\theta_0$ and $\text{Var} T = 1/\theta_0^2 + 2\tau^2$; see (1.2). The case of a very tight distribution for the θ corresponds to τ small, which again means the case of a constant rate θ_0 for all.

(c) A convenient class of distributions for θ is the Gamma, with parameters (a, b) , from Ex. 1.9. Its mean and variance are a/b and a/b^2 ; now find also the mean and variance of $1/\theta$. Show that the density of T can be written

$$f(t, a, b) = \int_0^\infty \theta \exp(-\theta t) \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) d\theta = \frac{ab^a}{(b+t)^{a+1}},$$

and also that its c.d.f. is

$$F(t, a, b) = 1 - \left(\frac{b}{b+t} \right)^a = 1 - \frac{1}{(1+t/b)^a}.$$

(d) Find an expression for the hazard rate function $h(t, a, b) = f(t, a, b)/\{1 - F(t, a, b)\}$, and comment on its form, compared to the exponential case.

(e) Find the mean and variance of T , for the $g(t, a, b)$ distribution. This may be used to estimate (a, b) from data.

Ex. 1.11 *Gamma-mixing the gamma.* A given parametric distribution may sometimes be fruitfully extended by placing a separate distribution on one of its parameters. The following is an illustration.

(a) Consider a distribution which for given individuals is a gamma, but where the scale parameter varies between individuals. Specifically, suppose $Y|b \sim \text{Gam}(a_0, b)$ and that b has a distribution with $E1/b = 1/b_0$ and $\text{Var} 1/b = \tau^2$. Show that Y has mean a_0/b_0 and variance $a_0/b_0^2 + (a_0 + a_0^2)\tau^2$.

(b) For the special case of $b \sim \text{Gam}(c, d)$, thus leading to a 3-parameter model, find the density $f(y, a, c, d)$ for Y :

$$\bar{f}(y) = \int_0^\infty \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by) \frac{d^c}{\Gamma(c)} b^{c-1} \exp(-db) db = \frac{d^c}{\Gamma(c)} \frac{\Gamma(a+c)}{\Gamma(a)} \frac{y^{a-1}}{(d+y)^{a+c}}.$$

Transformations, uniform, Pareto, Cauchy, Beta, Dirichlet

Ex. 1.12 *Transformation from X to Y .* We often encounter transformations, from one variable X to another Y , also in the vector case. We need formulae for how the density $g(y)$ of the Y can be found in terms of the density $f(x)$ for X .

(a) In the one-dimensional case, suppose $X = h(Y)$, equivalently $Y = h^{-1}(X)$, where h is smooth and increasing. Show that $\Pr(Y \leq y) = \Pr(X \leq h(Y))$, with density formula

$$g(y) = f(h(y))h'(y).$$

Show also that if $x = h(y)$ is continuous and decreasing, the formula becomes $g(y) = f(h(y))|h'(y)|$. Write down density formulae for the variables $Y_1 = \exp(X)$, $Y_2 = 3.33 - 2.22X$, $Y_3 = \log X$ (assuming for that case that X is positive).

(b) Show that if X is normal, then a linearly transformed $Y = a + bX$ is also normal. Show that if $X \sim \text{Gam}(a, b)$, with density proportional to $x^{a-1} \exp(-bx)$, then $Y = bX \sim \text{Gam}(a, 1)$.

(c) Suppose then that $X = (X_1, \dots, X_p)^t$ and $Y = (Y_1, \dots, Y_p)^t$ are vectors, with transformations binding them together,

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} h_1(Y_1, \dots, Y_p) \\ \vdots \\ h_p(Y_1, \dots, Y_p) \end{pmatrix}, \quad \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} h_1^{-1}(X_1, \dots, X_p) \\ \vdots \\ h_p^{-1}(X_1, \dots, X_p) \end{pmatrix}.$$

We write this as $X = h(Y)$ and $Y = h^{-1}(X)$, for short. It is assumed that these systems of equations have unique solutions, and that the transformations are smooth, with continuous partial derivatives. In particular, the so-called Jacobi matrix

$$J(y) = \frac{\partial h(y)}{\partial y} = \frac{\partial h(y_1, \dots, y_p)}{\partial y_1 \cdots \partial y_p},$$

having $\partial h_i(y)/\partial y_j$ as its (i, j) component, exists, and is continuous, with a non-zero determinant $\det(J(y))$. – Now, if X has density $f(x)$, show that

$$\Pr(Y \in B) = \int_{h(B)} f(x) dx = \int_B f(h(y)) |\det(J(y))| dy.$$

This shows that Y has density $g(y) = f(h(y))|\det(J(y))|$. This is essentially the multi-dimensional ‘integration by substitution’ formula of calculus.

(d) Let X, Y be independent standard exponentials, and consider $(U, V) = (X+Y, X/Y)$. Show that these become independent, having densities $u \exp(-u)$ and $1/(v+1)^2$. More generally, suppose in this construction of (U, V) that X, Y are independent from the $\text{Gam}(a, 1)$. Show that $U \sim \text{Gam}(2a, 1)$, that V has density $\{\Gamma(2a)/\Gamma(a)^2\} v^{a-1}/(v+1)^{2a}$, and that they are independent.

Ex. 1.13 Transformations to the uniform. We say that a variable U is uniform on the interval $[a, b]$ if its density is constant over that interval, i.e. $1/(b-a)$, and zero outside. In particular, we write $U \sim \text{unif}(0, 1)$ to indicate a variable with the uniform distribution on the unit interval.

(a) For such a $U \sim \text{unif}(0, 1)$, find its mean and variance. Find the probabilities that U lands in $[0.03, 0.04]$, or in $[0.77, 0.78]$.

(b) Let F be a continuous and increasing c.d.f. for a variable X . Show that $U = F(X)$ is uniform on the unit interval. Conversely, we may start with $U \sim \text{unif}(0, 1)$ and map to $X' = F^{-1}(U)$. Show that X' has distribution F . This is hence a simulation recipe, generating realisations from a given and perhaps not standard distribution.

(c) Consider the c.d.f. $F(x) = (x/10)^{3.33}$ on $[0, 10]$. Simulate 10^4 independent X_i from this distribution, by transforming uniforms. Make a fine histogram, with the density $f(x)$ plotted alongside.

Ex. 1.14 Polar and spherical coordinates. Points in the plane may be transformed from Cartesian to polar coordinates, and vice versa; similarly, but with more delicate details, points in the three-dimensional space can be represented in spherical coordinates.

Jacobi matrix
and
determinant

polar
coordinates

(a) Suppose X and Y are independent and standard normal, and transform to polar coordinates, $X = R \cos A$ and $Y = R \sin A$. Find the density $g(r, a)$ for (R, A) , with R positive and $A \in [0, 2\pi]$. Show in particular that length R and angle A become independent, with A having a uniform distribution on $[0, 2\pi]$ (i.e. the flat density $1/(2\pi)$ over that interval). Find also the distribution of $Z = Y/X = \tan A$; see also Ex. 1.18.

(b) Explain that a somewhat idiosyncratic but nevertheless feasible way of simulating pairs of standard normals is to draw R from the density $h_0(r) = r \exp(-\frac{1}{2}r^2)$ and A from the uniform over $[0, 2\pi]$, followed by the polar transformation. Do this in your computer, generating say 10^4 pairs (X_i, Y_i) . Then tweak your code, drawing R from some density h different from h_0 . Show that this amounts to the joint density

$$f(x, y) = h((x^2 + y^2)^{1/2})(x^2 + y^2)^{-1/2} / (2\pi).$$

Take e.g. $h(r) = \frac{1}{2}\gamma r^{\gamma-1} \exp(-\frac{1}{2}r^\gamma)$, with γ close to but not equal to 2. Generate and plot 10^4 points (x, y) via this recipe, perhaps via the inverse c.d.f. trick of Ex. 1.13; how can you spot that these data are *not* independent normal pairs?

(c) Then go from plane to space. Start with (X, Y, Z) a triple of independent standard normals, and transform to (R, A, B) via

$$X = R \sin A \cos B, \quad Y = R \sin A \sin B, \quad Z = R \cos A.$$

Set up the Jacobi matrix $\partial(x, y, z)/\partial(r, a, b)$ of partial derivatives, and show that its determinant becomes $r^2 \sin a$. Use this to show that the joint density of (R, A, B) can be written in product form $g_1(r)g_2(a)g_3(b)$, with g_1 the density proportional to $r^2 \exp(-\frac{1}{2}r^2)$, $g_2(a) = \frac{1}{2} \sin(a)$ over $[0, \pi]$, and $g_3(b) = 1/(2\pi)$ over $[-\pi, \pi]$.

(d) Just to see operationally that transformations go two ways, simulate say 10^5 standard normal triples (X, Y, Z) via simulating triples (R, A, B) from the densities found above. Check that the resulting (X, Y, Z) look adequate. As for the plane question, tweak your code to draw R from a somewhat different distribution than proportional to $r^2 \exp(-\frac{1}{2}r^2)$, and explain that you have invented a generalisation of the triple independent normal.

Ex. 1.15 Ordering exponentials. Let Y_1, Y_2, Y_3 be independent unit exponentials, with density $f(y) \exp(-y)$ for y positive, and order them, to $Y_{(1)} < Y_{(2)} < Y_{(3)}$. Then define the so-called spacings between them, $Z_1 = Y_{(1)}$, $Z_2 = Y_{(2)} - Y_{(1)}$, $Z_3 = Y_{(3)} - Y_{(2)}$.

(a) Show first that the joint density of $(Y_{(1)}, Y_{(2)}, Y_{(3)})$ is $3! f(y_{(1)})f(y_{(2)})f(y_{(3)})$ on the set $y_{(1)} < y_{(2)} < y_{(3)}$. Find then the joint density for (Z_1, Z_2, Z_3) , and show that they are independent.

(b) Then generalise, considering i.i.d. unit exponentials $Y_1, \dots, Y_n$, ordered into $Y_{(1)} < \dots < Y_{(n)}$. Work with the scaled spacings $V_1 = nY_{(1)}$, $V_2 = (n-1)(Y_{(2)} - Y_{(1)})$, up to $V_{n-1} = 2(Y_{(n-1)} - Y_{(n-2)})$, $V_n = Y_{(n)} - Y_{(n-1)}$. Reordering the V_i to W_{n+1-i} , for notational convenience, convince yourself that

$$Y_{(1)} = \frac{W_n}{n}, \quad Y_{(2)} = \frac{W_n}{n} + \frac{W_{n-1}}{n-1}, \quad \dots, \quad Y_{(n)} = \frac{W_n}{n} + \frac{W_{n-1}}{n-1} + \dots + \frac{W_2}{2} + \frac{W_1}{1},$$

with the W_i in fact being i.i.d. unit exponentials.

(c) The Euler constant $\gamma_e = 0.5772\dots$ is defined as the limit of $1 + 1/2 + \dots + 1/n - \log n$. Use the above to show that $Y_{(n)} = \max_{i \leq n} Y_i$ has mean close to $\log n + \gamma_e$, and variance converging to $\pi^2/6$.

the Pareto
distribution

Ex. 1.16 *The Pareto distribution.* Consider a variable T defined on the range $T \geq t_0$, for some positive t_0 , with c.d.f. $F(t) = 1 - (t_0/t)^\theta$ for $t \geq t_0$, for some positive parameter θ . This is the *Pareto distribution*, and we may write $T \sim \text{pareto}(t_0, \theta)$ to indicate this.

(a) Use the mean formula of (1.3) to show that the mean of a Pareto is $t_0\theta/(\theta - 1)$, for $\theta > 1$. Show furthermore that the variance is $t_0^2\theta/\{(\theta - 1)^2(\theta - 2)\}$, for $\theta > 2$, and that the median is $\text{med}(T) = t_0 2^{1/\theta}$.

(b) Show that the density is $f(t, \theta) = \theta t_0^\theta/t^{\theta+1}$ for $t \geq t_0$; you may use this to find the mean formula again. Furthermore, explain that the hazard rate is $h(t) = \theta/t$.

(c) Show that $\Pr(T \geq t | T \geq t_1) = (t_1/t)^\theta$, for $t \geq t_1$. So $T | (T \geq t_1)$ is Pareto (t_1, θ) . Deduce from this that

$$\text{med}(T - t | T \geq t) = ct = (2^{1/\theta} - 1)t \quad \text{for all } t.$$

the Lindy Effect

This is linked to the so-called Lindy Effect: the longer the life is observed to be (of a company, an idea, a party), the longer is the remaining lifetime expected to be (note that $c > 1$). Show in fact that the Pareto distribution is the only one with the property that $\text{med}(T - t | T \geq t) = ct$ for all t .

(d) Show that $Y = \log(T/t_0)$ is exponential with parameter θ . So a representation of the Pareto is $T = t_0 \exp(Y)$, with $Y \sim \text{Expo}(\theta)$. You may use this to find the mean formula once more.

Ex. 1.17 *Maxima of i.i.d. samples.* To illustrate one of the many ways in which suitable start models may be generalised and extended, consider $Y_1, \dots, Y_n$ i.i.d. from some distribution with c.d.f. F and density f , and from these define $M_n = \max_{i \leq n} Y_i$.

(a) Show that M_n has c.d.f. $G_n(z) = F(z)^n$ with density $g_n(z) = nF(z)^{n-1}f(z)$. Show that its median is $\mu_m = F^{-1}((\frac{1}{2})^{1/n})$. Compute and draw these, for say $n = 1, \dots, 10$, for the case of the Y_i being standard normal, tracking also the medians μ_m .

(b) Before pursuing some maxima, consider a situation with X_1 and X_2 being independent, with c.d.f.s H_1, H_2 and densities h_1, h_2 . Show that

$$\Pr(X_1 \leq X_2) = \int H_1(x_2)h_2(x_2)dx_2 = \int \{1 - H_2(x_1)\}h_1(x_1)dx_1.$$

(c) Long jumpers A and B have jumps being $N(7.90, 0.10^2)$ and $N(7.80, 0.15^2)$, respectively (on the metres scale). What is the probability that A jumps longer than B, with one jump each? Consider also p_m , the probability that B after m jumps is better than A after $m = 1, 2, 3, 4, 5, 6$ jumps (in the sense of ‘best jump so far’), assuming that all jumps are independent. Show that $p_m = \int F_1(z)^m m F_2(z)^{m-1} f_2(z) dz$, in terms of the two distributions. Compute p_m both via numerical integration and via simulations.

Ex. 1.18 *Ratios and the Cauchy.* If (X, Y) has a certain distribution, what happens to the ratio $V = Y/X$?

(a) Suppose that X and Y are independent with the same density f on $(0, \infty)$. Show that $V = Y/X$ has density $g(v) = \int_0^\infty xf(x)f(vx)dx$. With X and Y independent from the same exponential distribution, show that $g(v) = 1/(1+v)^2$.

(b) With X and Y independent from the same Gamma (a, b) , show that $V = Y/X$ has density $\{\Gamma(2a)/\Gamma(a)^2\} v^{a-1}/(1+v)^{2a}$.

(c) Suppose now that X and Y are independent from the same density f , symmetric around zero. Show that V has density $g(v) = 2 \int_0^\infty xf(x)f(vx)dx$. For the special case of a ratio of two independent standard normals, show that

the Cauchy

$$g(v) = (1/\pi)/(1+v^2) \quad \text{with c.d.f.} \quad G(v) = \frac{1}{2} + (1/\pi) \arctan v.$$

This is the *Cauchy distribution* (in its standard form). Show that it has no mean. Find its interquartile range. Deduce the Y/X distribution also via Ex. 1.14(a).

(d) There is an infinitely long straight line ahead of you, with distance r to the nearest point. You kick a ball towards the line, with some angle $A \in (-\pi/2, \pi/2)$. Show that it crosses the line at position $X = r \tan(A)$. Show that with A having density g and c.d.f. G , over $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$, then X has c.d.f. $F(x) = G(\arctan(x/r))$. When A is uniform over that interval, show that X/r is standard Cauchy. Investigate what happens if A is uniform over a tighter interval, like $(-\pi/4, \pi/4)$.

(e) What distribution does the random angle A need to have, in order for X to be normal? For $r = 1$ unit away from the line, show that the peculiar density $g(a) = \phi(\tan(a))/\cos^2(a)$ leads to a standard normal X . Graph g to see that it is symmetric and bimodal, with modes at ± 45 degrees.

a peculiar
density to find
the normal
again

Ex. 1.19 *The Beta distribution.* An important class of distributions, over the unit interval $(0, 1)$, is the *Beta distribution*, with two positive parameters. We write $p \sim \text{Beta}(a, b)$ if its density is

the Beta
distribution

$$\text{be}(p, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1} \quad \text{for } p \in (0, 1).$$

(a) As an introductory exercise, of separate interest, suppose X and Y are independent Gamma variables with parameters $(a, 1)$ and $(b, 1)$. Construct from these the sum $Z = X + Y$ and ratio $P = X/(X + Y)$. Finding the joint density for (P, Z) , demonstrate that Z and P are independent, with $Z \sim \text{Gam}(a+b, 1)$ and P having precisely the $\text{be}(p, a, b)$ density. This in particular shows that the integration constant is the correct one, i.e. that $\int_0^1 p^{a-1} (1-p)^{b-1} dp = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

(b) Compute and display a few of these densities, for (a, b) of your choice. Note that the uniform is the special case of $(a, b) = (1, 1)$.

(c) Show that $E p = p_0 = a/(a+b)$ and that $\text{Var } p = p_0(1-p_0)/(a+b+1)$.

(d) Find a formula for $E p^m$, for $m = 1, 2, \dots$, in terms of (a, b) , and in terms of the reparametrisation $(c p_0, c(1-p_0))$. Use this to find

$$E(p - p_0)^3 = \frac{2p_0(1-p_0)(1-2p_0)}{(c+1)(c+2)},$$

with a consequent formula for the skewness. For fixed mean p_0 , show that the skewness tends to zero with increasing c .

(e) Examine the particular $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ distribution. Show that its density and c.d.f. become $f(p) = (1/\pi)/\{p(1-p)\}^{1/2}$ and $F(p) = (2/\pi) \arcsin(\sqrt{p})$. Find its quantile $F^{-1}(q)$. Plot these functions.

Ex. 1.20 *The Dirichlet distribution.* Let $G_1, \dots, G_k$ be independent and Gamma distributed, with parameters $(a_1, 1), \dots, (a_k, 1)$. With $G = G_1 + \dots + G_k$ their sum, consider the random ratios

$$(X_1, \dots, X_{k-1}) = (G_1/G, \dots, G_{k-1}/G).$$

It inherits a distribution, with density $h(x_1, \dots, x_{k-1})$, worked with below, in the simplex where each $x_i \geq 0$ and $x_1 + \dots + x_{k-1} < 1$. Taking also $X_k = G_k/G = 1 - (X_1 + \dots + X_{k-1})$

the Dirichlet
distribution

on board, we have a vector $(X_1, \dots, X_k)$ of random probabilities summing to 1 over its k categories. Its distribution has a name: it's the *Dirichlet distribution*, with k categories, and parameters $(a_1, \dots, a_k)$, which we write as $X \sim \text{Dir}(a_1, \dots, a_k)$.

(a) Suppose $(X_1, X_2, X_3, X_4, X_5, X_6) \sim \text{Dir}(a_1, a_2, a_3, a_4, a_5, a_6)$. Show that $(X_1 + X_4 + X_6, X_2, X_3 + X_5) \sim \text{Dir}(a_1 + a_4 + a_6, a_2, a_3 + a_5)$. Generalise and formalise this summing-over-cells property of the Dirichlet distribution.

(b) With $X \sim \text{Dir}(a_1, \dots, a_k)$, show that each $X_i \sim \text{Dir}(a_i, a - a_i)$, with $a = a_1 + \dots + a_k$, and that this is the same as a $\text{Beta}(a_i, a - a_i)$. Show from this that

$$\mathbb{E} D_i = \xi_i = a_i/a, \quad \text{Var } D_i = \xi_i(1 - \xi_i)/(a + 1),$$

Show also the $\text{cov}(D_i, D_j) = -\xi_i \xi_j / (a + 1)$ for $i \neq j$.

(c) We have been able to derive certain basic properties above, without really needing an expression for the density of a Dirichlet vector. We tend to this now, using the transformation machinery of Ex. 1.12. Show in fact, starting with $(G_1, \dots, G_k)$ and transforming to $(X_1, \dots, X_{k-1}, G)$, (i) that $(X_1, \dots, X_{k-1})$ has the density

$$h(x_1, \dots, x_{k-1}) = \frac{\Gamma(a)}{\Gamma(a_1) \dots \Gamma(a_k)} x_1^{a_1-1} \dots x_{k-1}^{a_{k-1}-1} (1 - x_1 - \dots - x_{k-1})^{a_k-1}$$

over the simplex; (ii) that $G \sim \text{Gam}(a, 1)$; (iii) that these are independent; and (iv) that this also verifies the implied formula for integrating $x_1^{a_1-1} \dots x_{k-1}^{a_{k-1}-1} (1 - x_1 - \dots - x_{k-1})^{a_k-1}$ over the simplex. Note that these efforts and results generalise the findings of Ex. 1.19(a).

Ex. 1.21 *Dirichlets inside Dirichlets.* The summing-over-cells property of the Dirichlet, see Ex. 1.20, has other consequences and angles, and we shall learn here that long Dirichlet vectors might be split into Dirichlet parts via Dirichlet cuts.

(a) Start with A, B, C independent gammas with parameters $(a, 1)$, $(b, 1)$, $(c, 1)$. Form from these $X = A/(A+B)$, $Y = (A+B)/(A+B+C)$, $Z = A+B+C$. Explain that we already know that $X \sim \text{Beta}(a, b)$, $Y \sim \text{Beta}(a+b, c)$, $Z \sim \text{Gam}(a+b+c, 1)$. Show in fact that X, Y, Z also are independent. This means working out their joint distribution; establish first that the inverse transform is $A = XYZ$, $B = (1-X)YZ$, $C = (1-Y)Z$, and that the associated Jacobi determinant becomes yz^2 . Explain that all of this leads to the product Beta representation

$$\frac{A}{A+B+C} = \frac{A}{A+B} \frac{A+B}{A+B+C} = \text{Beta}(a, b) \text{Beta}(a+b, c) = \text{Beta}(a, b+c).$$

(b) Suppose $(X_1, \dots, X_k, Y_1, \dots, Y_\ell)$ is $\text{Dir}(a_1, \dots, a_k, b_1, \dots, b_\ell)$, which we may represent as $X_i = G_i/(S+T)$, $Y_j = H_j/(S+T)$, with the G_i and H_j being independent gamma variables with the appropriate parameters, and sums $S = \sum_{i=1}^k G_i$ and $T = \sum_{j=1}^\ell H_j$. Explain that $W = \sum_{i=1}^k X_i = S/(S+T)$ is a $\text{Beta}(a, b)$, with $a = \sum_{i=1}^k a_i$ and $b = \sum_{j=1}^\ell b_j$, and show that

$$X_i = \frac{G_i}{S+T} = W \frac{G_i}{S} = W X'_i \quad \text{and} \quad Y_j = \frac{H_j}{S+T} = (1-W) \frac{H_j}{T} = (1-W) Y'_j,$$

where $X' = (X'_1, \dots, X'_k) \sim \text{Dir}(a_1, \dots, a_k)$ and $Y' = (Y'_1, \dots, Y'_\ell) \sim \text{Dir}(b_1, \dots, b_\ell)$, independent of W . We learn that the long Dirichlet vector (X, Y) may be split into two separate Dirichlet vectors $X' = X/w$ and $Y' = Y/(1-w)$, by conditioning on $W = w$.

(c) Now generalise to longer vectors. Let $X = (X_1, \dots, X_k)$ be a $\text{Dir}(a_1, \dots, a_k)$, where each X_i is a vector $(X_{i,1}, \dots, X_{i,m_i})$, with corresponding $a_i = (a_{i,1}, \dots, a_{i,m_i})$. Write $W_i = \sum_{j=1}^{m_i} X_{i,j}$ for the sum over X_i . Explain first that $(W_1, \dots, W_k) \sim \text{Dir}(b_1, \dots, b_k)$, where $b_i = \sum_{j=1}^{m_i} a_{i,j}$ is the sum over a_i . Show next that with $X'_i = X_i/W_i$, then $X'_i \sim \text{Dir}(a_i)$, and that $X'_1, \dots, X'_k$ are independent of $(W_1, \dots, W_k)$.

(d) The previous point implies that a long Dirichlet vector may be split into independent Dirichlet components, in several ways. In the setting above, starting with the long $(X_1, \dots, X_k)$, suppose we condition on $(W_1, \dots, W_k) = (w_1, \dots, w_k)$, a given probability vector. With $X'_i = X_i/w_i$, show that $X'_1, \dots, X'_k$ are independent, with $X'_i \sim \text{Dir}(a_i)$. This is the sorcerer's apprentice property of the Dirichlet distribution; pin it down, into subsets with given sums, and see that each part is a scaled Dirichlet again.

Ex. 1.22 *The Beta-binomial distribution.* Sometimes comparable binomial experiments may be modelled and analysed jointly, but where the success probability is not the same across studies. The $p = \text{Pr}(\text{girl})$ may e.g. vary from family to family, see Story #2.

(a) Suppose in general terms that $Y|p \sim \text{binom}(n, p)$, and that p has a distribution with mean p_0 and standard deviation τ_0 . Using the double expectation rule (1.2), show that

$$EY = np_0 \quad \text{and} \quad \text{Var } Y = np_0(1 - p_0) + n(n-1)\tau_0^2.$$

Hence the extra-binomial component of the variance, the $n(n-1)\tau_0^2$, becomes more noticeable with increasing n . The case of $\tau_0 = 0$ corresponds to the usual binomial.

(b) Suppose $Y|p \sim \text{binom}(n, p)$ and that $p \sim \text{Beta}(a, b)$. Show that this leads to the distribution

$$\begin{aligned} \text{Pr}(Y = y) &= \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} g(p, a, b) \, dp \\ &= \binom{n}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(n+a+b)} \quad \text{for } y = 0, 1, \dots, n. \end{aligned}$$

Give formulae for the mean and variance of Y . For the special case of the uniform for p , show that all outcomes for Y are equally likely.

Ex. 1.23 *The Dirichlet-multinomial distribution.* Here we deal with the natural extension of the Beta-binomial setup, from the case of two categories to more than two.

(a) Let $Y = (Y_1, \dots, Y_k)$, for given probability vector $p = (p_1, \dots, p_k)$, have a multinomial $(n, p_1, \dots, p_k)$ model, as per Ex. 1.5. Assume then that the p is not fixed, but with p_i variances $\tau_{0,i}^2$ around mean $p_{0,i}$. Show that Y_i , marginally, has mean $np_{0,i}$ and variance $np_{0,i}(1 - p_{0,i}) + n(n-1)\tau_{0,i}^2$.

(b) Let in particular $p \sim \text{Dir}(cp_0)$, with parameters $cp_0 = (cp_{0,1}, \dots, cp_{0,k})$. Show that

$$\text{Var } Y_i = \{n + n(n-1)/(c+1)\}p_{0,i}(1 - p_{0,i}) = \frac{c+n}{c+1}np_{0,i}(1 - p_{0,i}),$$

with a clear overdispersion factor with respect to multinomial variation.

(c) Show that the marginal distribution of $(Y_1, \dots, Y_k)$, now overdispersed compared to

the multinomial, becomes

$$\begin{aligned}\bar{f}(y_1, \dots, y_k) &= \int \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_{k-1}^{y_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{y_k} \\ &\quad g(p_1, \dots, p_{k-1}) \, dp_1 \cdots dp_{k-1} \\ &= \frac{n!}{y_1! \cdots y_k!} \frac{\Gamma(c)}{\Gamma(cp_{0,1}) \cdots \Gamma(cp_{0,k})} \frac{\Gamma(cp_{0,1} + y_1) \cdots \Gamma(cp_{0,k} + y_k)}{\Gamma(c + n)}.\end{aligned}$$

(d) For the case of Dirichlet parameters $cp_0 = (1, \dots, 1)$, show that all outcomes $(y_1, \dots, y_k)$ have the same probability. How many different outcomes can there be?

Laws of small numbers, Poisson, geometric, negative binomial

Ex. 1.24 *The Poisson distribution.* Counting a high number of events with small probabilities leads to the Poisson distribution, which we define here. A count variable Y is said to have the *Poisson distribution* with parameter $\theta > 0$ if

$$\Pr(Y = y) = \exp(-\theta)\theta^y/y! \quad \text{for } y = 0, 1, 2, \dots$$

We write $Y \sim \text{Pois}(\theta)$ to indicate this.

(a) Show that the probabilities indeed sum to 1. Verify next that $EY = \theta$, $EY(Y-1) = \theta^2$, and show from this that the variance is equal to the mean. Show further that $EY(Y-1)(Y-2) = \theta^3$, $EY(Y-1)(Y-2)(Y-3) = \theta^4$, and with further algebra that for $W = (Y - \theta)/\sqrt{\theta}$, we have skew = $EW^3 = 1/\sqrt{\theta}$, kurt = $EW^4 - 3 = 1/\theta$. Show also that $\text{Var}(Y - \theta)^2 = 2\theta^2 + \theta$.

(b) With $Y \sim \text{Pois}(\theta)$, what is the most probable outcome? What is the probability that Y is odd?

(c) Show that the sum of two independent Poisson variables is Poisson, with parameter equal to the sum of the two parameters. Generalise.

(d) Consider $Y \sim \text{binom}(n, p)$, and assume that n grows, while p becomes small, in the fashion of $np \rightarrow \theta$. Show that Y then tends to the $\text{Pois}(\theta)$ distribution, in the sense that the point probabilities converge. See also Ex. 2.8 for a fuller picture.

(e) In some event counting applications there are more zeros than predicted by the Poisson, leading naturally to a more general model with

$$\Pr(Y = 0) = p_0, \quad \Pr(Y = y) = (1 - p_0) \exp(-\theta)\theta^y/y! / \{1 - \exp(-\theta)\} \quad \text{for } y \geq 1.$$

the Law of
Small Numbers

the ZIP

Verify that these probabilities sum to 1, and find expressions for the mean and variance. This model is sometimes called the zero-inflated Poisson, since situations with $p_0 > \exp(-\theta)$ are prevalent, but also cases with $p_0 < \exp(-\theta)$ are allowed. Simulate say 1000 datapoints from the model with $\theta = 3.00$ and $p_0 = 0.25$, and check the histogram.

Ex. 1.25 *The geometric distribution.* Suppose Y has the distribution with point probabilities $f(y) = (1 - p)^{y-1}p$ for $y = 1, 2, \dots$. This is the *geometric distribution*, and we write $Y \sim \text{geom}(p)$ to indicate this.

(a) Show that the probabilities $f(y)$ indeed sum to 1. Suppose independent experiments are carried out, each time with probability p that a certain event A takes place. With Y the first time A happens, show that $Y \sim \text{geom}(p)$.

(b) Show that Y has mean $1/p$ and variance $(1 - p)/p^2$, via direct summation of $\sum_{y=1}^{\infty} yf(y)$ etc. If Y is the number of times you need to roll a six-sided die until it shows a '6', find the mean and the standard deviation.

(c) Another way of finding the mean and variance is as follows. With probability p , $Y = 1$; with complementary probability $1 - p$, $Y = 1 + Y'$, with Y' having the same distribution as Y . Show that this leads to $EY = p + (1 - p)(1 + EY)$ and solve. Use this representation to also find the variance. Show also that $E(Y - 1/p)^3 = (1 - p)(2 - p)/p^3$.

(d) Show that $\Pr(Y \geq y) = (1 - p)^{y-1}$, derive a formula for $\Pr(Y \geq y_0 + y | Y \geq y_0)$, and comment.

(e) A simple related distribution is when one starts counting at 0, not at 1, so to speak. Show that with $Y \sim \text{geom}(p)$, as defined above, the variable $Y_0 = Y - 1$ has point probabilities $\Pr(Y_0 = y) = q^y p$ for $y = 0, 1, \dots$, writing $q = 1 - p$. Show that Y_0 has mean $(1 - p)/p$ and variance $(1 - p)/p^2$.

(f) Suppose Y given p has geometric probabilities $(1 - p)^{y-1}p$ for $y = 1, 2, \dots$, but that p itself stems from a uniform distribution. Find the distribution for Y ; note that this implies $1/2 + 1/6 + 1/12 + 1/20 + \dots = 1$.

Ex. 1.26 *Time to last event.* Suppose independent geometric experiments are carried out by m individuals, say players throwing dice until they get the '6'. How long time does it take until the event in question has taken place, for all individuals?

(a) Let $Y_1, \dots, Y_m$ be the time needed for the m individuals to see the event in question. Show that $F(y) = \Pr(Y_i \leq y) = 1 - (1 - p)^y$ for $y = 1, 2, \dots$. For $Z_m = \max(Y_1, \dots, Y_m)$, show that its c.d.f. is $H_m(z) = F(z)^m$. Find a formula for the median time to final event: $z_m^* = \log(1 - (\frac{1}{2})^{1/m}) / \log(1 - p)$ (which may be rounded to the nearest integer).

(b) For $p = 1/6$, compute also the mean $E Z_m$, for say $m = 1, \dots, 100$. Plot this, along with the median times z_m^* .

Ex. 1.27 *Mixing the Poisson.* Suppose observations come from Poisson mechanisms, but with different parameters, forming their own distribution. There are several versions and uses of such Poisson overdispersion models.

(a) Suppose $Y | \theta \sim \text{Pois}(\theta)$ but that θ has a distribution with mean θ_0 and variance τ_0^2 . Show that Y has mean θ_0 and variance $\theta_0 + \tau_0^2$.

(b) Specialise to the case of $\theta \sim \text{Gam}(a, b)$, see Ex. 1.9. Show that $EY = \theta_0 = a/b$ and that $\text{Var } Y = \theta_0(1 + 1/b)$. Argue that with a large b we come back to pure Poisson. Show also that the marginal distribution of Y becomes

$$f(y, a, b) = \frac{\Gamma(a + y)}{\Gamma(a) y!} \frac{b^a}{(b + 1)^{a+y}} = \frac{\Gamma(a + y)}{\Gamma(a) y!} \left(\frac{b}{b + 1}\right)^a \left(\frac{1}{b + 1}\right)^y \quad \text{for } y = 0, 1, 2, \dots$$

We are discovering the general *negative binomial* distribution in the process, of the form negative
binomial

$$g(y, a, p) = \frac{\Gamma(a + y)}{\Gamma(a) y!} (1 - p)^y p^a \quad \text{for } y = 0, 1, \dots, \quad (1.5)$$

for parameters $a > 0$, $p \in (0, 1)$; see Ex. 1.28 for more details.

Ex. 1.28 *The negative binomial.* We met the negative binomial distribution in Ex. 1.27 and now point to other features and constructions.

(a) Let X_1, X_2 be independent from the geometric distribution $q^x p$ for $x = 0, 1, \dots$, with $q = 1 - p$. Show that $Y = X_1 + X_2$ has distribution $\Pr(Y = y) = (y + 1)q^y p^2$, for $y = 0, 1, \dots$. For $Y = X_1 + X_2 + X_3$ a sum of three such independent geometric variables, show that $\Pr(Y = y) = \binom{y+2}{2} q^y p^3$ for $y = 0, 1, \dots$.

(b) Generalise to the case of $Y = X_1 + \dots + X_a$, the sum of a independent geometric variables, each with $q^x p$ for $x = 0, 1, \dots$. Show that

$$\Pr(Y = y) = \binom{y+a-1}{a-1} q^y p^a = \frac{\Gamma(y+a)}{\Gamma(a) y!} (1-p)^y p^a \quad \text{for } y = 0, 1, \dots,$$

i.e. the *negative binomial distribution* with parameters (a, p) . Deduce that the number of ways in which one may find nonnegative numbers $x_1, \dots, x_a$ with a given sum y is $\binom{y+a-1}{a-1} = (y+a-1)!/\{(a-1)! y!\}$. In how many ways may one find 5 nonnegative numbers with sum 100? And with 10 nonnegative numbers with sum 100?

(c) How do we know that the negative binomial probabilities (1.5) sum to one, also when the a is a non-integer? Deduce from this that

$$\sum_{y=0}^{\infty} \frac{\Gamma(y+a)}{\Gamma(a)} \frac{u^y}{y!} = \frac{1}{(1-u)^a} \quad \text{for } u \in (0, 1).$$

Show that $EY = aq/p$, $\text{Var } Y = aq/p^2$.

(d) In one of the episodes of the television series *Siffer* (NRK, 2011), programme leader Jo Røislien flipped a coin and then announced he would do this nine more times to get the very same outcome ten times in a row – which he then proceeded to do. He looked a bit tired, though; he had just kept on doing this, complete with his opening statement, until he had achieved the ten in a row event, and then showed only this crowning minute on tv. About how many times did he need to flip his coin, in total, before he (and his camera man) could show that final string of crowns? Simulate the process, and give a histogram of say 1,000 realisations.

Ex. 1.29 *Conditioning on Poisson sums and a generalised binomial.* We start with Poisson sums and see a connection to the binomial. Using generalised Poissons then lead to generalisations of the binomial.

(a) Let X and Y be independent Poissons with parameters θ_1, θ_2 . Show that X given $X+Y = n$ is a binomial (n, p) , with $p = \theta_1/(\theta_1 + \theta_2)$. Generalise to the case of $X_1, \dots, X_m$ being independent Poissons, with parameters $\theta_1, \dots, \theta_m$. Show that their conditional distribution given $X_1 + \dots + X_m = n$ is a multinomial with count n and probabilities $(p_1, \dots, p_m)$, where $p_j = \theta_j/(\theta_1 + \dots + \theta_m)$.

(b) To generalise the above, and in its turn also the binomial distribution, consider the gamma-mixed Poissons of 1.27. Specifically, let $X | \theta_1 \sim \text{Pois}(\theta_1)$ and $Y | \theta_2 \sim \text{Pois}(\theta_2)$, with gamma distributions $(c\theta_{0,1}, c)$ and $(c\theta_{0,2}, c)$ for θ_1 and θ_2 ; when c is large, this means tight concentration around $\theta_{1,0}, \theta_{2,0}$, and we're back to Poisson. Show that the conditional distribution of X given $X+Y = n$ may be written

$$f(x | n) \propto \binom{n}{x} \Gamma(c\theta_{0,1} + x) \Gamma(c\theta_{0,2} + n - x) \quad \text{for } x = 0, 1, \dots, n,$$

where ' $\propto$ ' means 'proportional to'. For a particular case, explain that with $c\theta_{0,1} = c\theta_{0,2} = 1$, X has the uniform distribution on $0, 1, \dots, n$.

(c) It may not be easy to sum the terms above directly, to find the normalisation constant, but show via expressions in Ex. 1.22 that we in fact must have

$$f(x | n) = \binom{n}{x} \frac{\Gamma(c\theta_{0,1} + c\theta_{0,2})}{\Gamma(c\theta_{0,1}) \Gamma(c\theta_{0,2})} \frac{\Gamma(c\theta_{0,1} + x) \Gamma(c\theta_{0,2} + n - x)}{\Gamma(c\theta_{0,1} + c\theta_{0,2} + n)}$$

for $x = 0, 1, \dots, n$. In yet other words, we have reinvented the Beta-binomial distribution with gamma-mixing of Poisson parameters. Also, for large c , we are back to plain binomial (n, p) , with $p = \theta_{0,1}/(\theta_{0,1} + \theta_{0,2})$.

(d) To invent other generalisations of the binomial, therefore, we might attempt other extensions of the Poisson. One such is to let X and Y have point probabilities proportional to $\theta_1^x/(x!)^\gamma$ and $\theta_2^y/(y!)^\gamma$, for some $\gamma \neq 1$; we work with such models in Ex. 4.35 and Story #28. Show here that X given $X + Y = n$ has the distribution

$$f(x|n) \propto \binom{n}{x}^\gamma p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n,$$

where $p = \theta_1/(\theta_1 + \theta_2)$. Compute these probabilities, for say $n = 50$, $p = 0.33$, and some values of γ around 1. Check then numerically $np_0 = EX$ and the dispersion ratio $\rho = \text{Var } X / \{np_0(1-p_0)\}$, to learn that there is overdispersion and underdispersion, compared to the binomial case, for $\gamma < 1$ and $\gamma > 1$, respectively.

Moments, moment-generating functions, characteristic functions

Ex. 1.30 *Jensen, Hölder, and Cauchy-Schwarz.* The three inequalities named after the four mathematicians in the title of this exercise are used time and again in probability and statistics.

(a) Recall that the function $g: \mathbb{R} \rightarrow \mathbb{R}$ is convex if $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$ for all x, y and all $\lambda \in [0, 1]$. Show that if g is convex, then g is continuous, and that

$$\lambda_1 g(x_1) + \dots + \lambda_k g(x_k) \geq g(\lambda_1 x_1 + \dots + \lambda_k x_k),$$

for any $\lambda_1, \dots, \lambda_k \in [0, 1]$ such that $\lambda_1 + \dots + \lambda_k = 1$. Use this to show that for any simple random variable $X_m = \sum_{j=1}^m a_j I_{A_j}$ (see Ex. B.11 for the precise definition), $Eg(X_m) \geq g(EX_m)$ for any convex function g . This is Jensen's inequality for simple random variables.

(b) For a convex function $g: \mathbb{R} \rightarrow \mathbb{R}$ we can for any x_0 choose a and b such that $a + bx_0 = g(x_0)$ and $a + bx \leq g(x)$ for all x (make a drawing!). Use this to show Jensen's inequality for general random variables X , namely that for any convex function $g: \mathbb{R} \rightarrow \mathbb{R}$

Jensen's
inequality

$$Eg(X) \geq g(EX).$$

To remember what way the inequality goes, recall the variance is always nonnegative, so $\text{Var}(X) = EX^2 - (EX)^2 \geq 0$, and that $g(x) = x^2$ is a convex function.

(c) Consider a variable Y with mean μ and median m . Show that $|\mu - m| \leq \sigma$, the standard deviation.

(d) For a random (Z_1, Z_2) , with finite variances σ_1^2 and σ_2^2 , and correlation ρ , show from

Cauchy-Schwarz
inequality

$$\text{Var}(Z_1/\sigma_1 + Z_2/\sigma_2) = 2 + 2\rho \quad \text{and} \quad \text{Var}(Z_1/\sigma_1 - Z_2/\sigma_2) = 2 - 2\rho$$

that the correlation must be in $[-1, 1]$. If $\rho = 1$, what can be said about Z_1, Z_2 ?

(e) Let X and Y be random variables such that EX^p and $E|Y|^q$ are finite, where $p, q > 1$ and $1/p + 1/q = 1$. Set $K = EX^p$, so that $Q(A) = K^{-1}EI_A|X|^p$ is a probability measure (see Ex. B.15(e)), and let $E_Q(\cdot)$ be the expectation with respect to Q . Use Jensen's inequality to show that

$$\frac{1}{K^q} (E|XY|)^q = (E_Q \frac{|Y|}{|X|^{p-1}})^q \leq \frac{1}{K} E|Y|^q,$$

and from this, deduce Hölder's inequality,

Hölder's
inequality

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}.$$

Cauchy–
Schwarz
inequality

From the above, deduce the Cauchy–Schwarz inequality $|\mathbb{E}XY|^2 \leq \mathbb{E}X^2 \mathbb{E}Y^2$.

(f) Show that for any random variables X and Y $\text{cov}(X, Y)^2 \leq \text{Var}(X) \text{Var}(Y)$ entailing, again, that the correlation of two random variables lies between -1 and 1 .

Ex. 1.31 Moments. Consider a random variable X with c.d.f. F . Its mean is $\mathbb{E}X = \int x \, dF(x)$, and we may of course define higher moments.

(a) Use results of Ex. C.3 to show that $\mathbb{E}X^k$, first seen as $\int y \, dG_k(y)$, the mean of the variable $Y = X^k$ with distribution G_k inherited from F , is also the same as $\int x^k \, dF(x)$; thus there is no ambiguity there.

(b) For $r < s$, show that $(\mathbb{E}|X|^r)^{1/r} \leq (\mathbb{E}|X|^s)^{1/s}$, i.e. $h(r) = (\mathbb{E}|X|^r)^{1/r}$ is a non-decreasing function in r . In particular, note that if $|X|$ has a finite s -moment, then all moments of smaller order are also finite. Illustrate by computing and graphing $h(r)$ and $\log h(r)$ for the case of $X \sim \text{Expo}(1)$.

(c) For X a standard normal, show that $\mathbb{E}|X|^p = 2^{p/2} \Gamma(\frac{1}{2}(p+1)) / \sqrt{\pi}$ for $p \geq 0$, a formula that also can be written $(\frac{1}{2})^{p/2} \Gamma(p+1) / \Gamma(\frac{1}{2}p+1)$. Compute and graph the function $h(r)$ for this case. For p an even integer, the formula simplifies to $(\frac{1}{2})^{p/2} p! / (\frac{1}{2}p)!$.

(d) For a random variable X with finite fourth moment, we have defined its skewness and kurtosis in Ex. 1.3. Give expressions for skew $= h_3(\mu_1, \mu_2, \mu_3)$ and kurt $= h_4(\mu_1, \mu_2, \mu_3, \mu_4)$ in terms of the moments $\mu_j = \mathbb{E}X^j$, and also for skew $= h_3^*(\mu_2^*, \mu_3^*)$ and kurt $= h_4^*(\mu_2^*, \mu_3^*, \mu_4^*)$ in terms of the centralised moments $\mu_j^* = \mathbb{E}(X - \mu_1)^j$.

Ex. 1.32 Moment-generating functions. For a random variable Y , with distribution P , its m.g.f. is

$$M(t) = \mathbb{E} \exp(tY) = \int \exp(ty) \, dP(y),$$

defined for each t at which the expectation exists. The moment-generating function is useful for finding and characterising distributions, for finding their moments, for handling the distributions of sums of variables, and in connection with distributional limits. When Y has a density $f(y)$ (with respect to Lebesgue measure), we have $M(t) = \int \exp(ty) f(y) \, dy$, and if it is discrete with pointmasses $f(y)$ for sample space S , say, then $M(t) = \sum_{y \in S} \exp(ty) f(y)$. The expectation operator is more general, however, and $M(t)$ is perfectly defined also for intermediate cases where Y can have both discrete and continuous parts; see Ex. C.3.

(a) For a standard normal $Y \sim N(0, 1)$, show that $M(t) = \exp(\frac{1}{2}t^2)$. When $Y \sim N(\xi, \sigma^2)$, derive $M(t) = \exp(\xi t + \frac{1}{2}\sigma^2 t^2)$.

(b) A moment-generating function has that name since it generates moments; we indeed have $M'(0) = \mathbb{E}Y$, $M''(0) = \mathbb{E}Y^2$, etc., under mild conditions, see Ex. C.21 for details. Use this to find the first four moments of the standard normal.

(c) For $Y \sim \text{Expo}(\theta)$, show that $M(t) = 1/(1 - t/\theta)$, for $t < \theta$.

(d) For $Y \sim \text{Gam}(a, b)$, with density $\{b^a/\Gamma(a)\}y^{a-1} \exp(-by)$, show that $M(t) = \{b/(b-t)\}^a$, for $t < b$. In particular, $M(t) = 1/(1-t)^a$ for $\text{Gam}(a, 1)$.

(e) Suppose Y is equal to zero with probability 0.90, but a standard normal with probability 0.10. Find the $M(t)$, and generalise.

(f) For the binomial (n, p) , show that $M(t) = \{1 - p + p \exp(t)\}^n$.

(g) For $Y \sim \text{Pois}(\theta)$, find $M(t) = \exp\{\theta(e^t - 1)\}$. Use this, with Ex. 1.27, to find $M(t)$ also for the negative binomial (a, p) :

$$M(t) = \left(\frac{p}{1 - (1-p)e^t} \right)^a \quad \text{for } t < -\log(1-p).$$

(h) Let $Y = \pm 1$ with probabilities $\frac{1}{2}, \frac{1}{2}$. Show that

$$M(t) = \cosh(t) = \frac{1}{2}(e^t + e^{-t}) = 1 + (1/2)t^2 + (1/4!)t^4 + (1/6!)t^6 + \dots$$

(i) For the uniform distribution on the unit interval, show that $M(t) = \{\exp(t) - 1\}/t$, for $t \neq 0$, and with $M(0) = 1$. For Y having the uniform distribution on the $[-1, 1]$ interval, show that

$$M(t) = \frac{\exp(t) - \exp(-t)}{2t} = \frac{\sinh t}{t},$$

and that this function may be written as the infinite sum $1 + (1/3!)t^2 + (1/5!)t^4 + \dots$.

Ex. 1.33 *Distribution of sums via moment-generating functions.* Importantly, the m.g.f. $M(t) = E \exp(tY)$, if it exists in a neighbourhood around zero, characterises the distribution; variables whose m.g.f.s are identical in such a neighbourhood have identical distributions. See details in Ex. C.21.

(a) Suppose $M(t) = (\frac{1}{2})^{25} \{1 + \exp(t)\}^{25}$. What is the underlying distribution? Similarly, if $M(t) = 0.99 + 0.01 \exp(\frac{1}{2}t^2)$, what is the distribution?

(b) With X and Y being independent, with m.g.f.s M_1 and M_2 , show that the sum $Z = X + Y$ has m.g.f. $M_X(t)M_Y(t)$.

(c) With X and Y independent and standard normal, show that $X + Y$ is $N(0, 2)$. Redo the questions of Ex. 1.2.

(d) Redo a part of Ex. 1.24, showing that sums of independent Poissons are Poisson. Similarly, if X, Y, Z are independent binomials $(n_1, p), (n_2, p), (n_3, p)$, identify the distribution of $W = X + Y + Z$.

(e) If $X_1, \dots, X_n$ are independent gamma variables, with parameters $(a_1, b), \dots, (a_n, b)$, find the distribution of $X_1 + \dots + X_n$.

Ex. 1.34 *The Laplace distribution.* The Laplace or *double exponential distribution*, in its simplest form, has density $f_0(y) = \frac{1}{2} \exp(-|y|)$, on the real line; note the cusp at its centre point zero. the Laplace distribution

(a) Let V_1 and V_2 be independent standard exponentials. Show that $Y = V_1 - V_2$ has this density $f_0(y)$. Deduce from this that its m.g.f. is $M_0(t) = 1/(1 - t^2)$, for $|t| < 1$.

(b) More generally, consider $Y = V_1 - V_2$ where these two are independent and $\text{Expo}(\theta)$. Show that Y has density $f(y) = \frac{1}{2}\theta \exp(-\theta|y|)$, with zero mean and variance $2/\theta^2$. Also, show that its m.g.f. is $M(t) = 1/\{1 - (t/\theta)^2\}$ for $|t| < \theta$. The Laplace with variance 1 is hence that with $\theta = \sqrt{2}$.

(c) Suppose X for given σ is a $N(0, \sigma^2)$, but that the variance $V = \sigma^2$ has some distribution. Show that the m.g.f. for such a normal scale mixture becomes $M(t) = E \exp(tX) = M_V(\frac{1}{2}t^2)$, where $M_V(s)$ is the m.g.f. for V . In particular, show that if $X | \sigma \sim N(0, \sigma^2)$ and $\sigma^2 \sim \text{Expo}(1)$, then X has the Laplace distribution with variance 1.

(d) If $X|V \sim N(0, V)$, and V has density $g(v)$, show that X has density $f(x) = \int_0^\infty \phi(x/v^{1/2})(1/v^{1/2})g(v) dv$. Translate the result above to the interesting formula

$$\begin{aligned} \int_0^\infty \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2} \frac{x^2}{v}\right) \frac{\exp(-v)}{v^{1/2}} dv &= 2 \int_0^\infty \frac{1}{(2\pi)^{1/2}} \exp\left\{-\left(\frac{1}{2}x^2/w^2 + w^2\right)\right\} dw \\ &= \frac{1}{2}\sqrt{2} \exp(-\sqrt{2}|x|). \end{aligned}$$

Use this to find a formula for the integral $\int_0^\infty \exp\{-(av^2 + b/v^2)\} dv$.

Ex. 1.35 *Summing uniforms.* Consider independent uniforms $U_1, U_2, \dots$ on the unit interval. Let us sum them.

(a) Let $U_1, U_2, \dots$ be i.i.d. uniforms on the unit interval. Show that the densities for $U_1 + U_2$ and $U_1 + U_2 + U_3$ may be written

$$\begin{aligned} f_2(x) &= \begin{cases} x & \text{when } 0 \leq x \leq 1, \\ 2 - x & \text{when } 1 \leq x \leq 2, \end{cases} \\ f_3(x) &= \begin{cases} \frac{1}{2}x^2 & \text{when } 0 \leq x \leq 1, \\ \frac{1}{2}(-2x^2 + 6x - 3) & \text{when } 1 \leq x \leq 2, \\ \frac{1}{2}(3 - x)^2 & \text{when } 2 \leq x \leq 3. \end{cases} \end{aligned}$$

Show that f_2 is continuous, but that its derivative has a jump at position $x = 1$. Show however that f_3 is smoother, with a continuous derivative.

(b) Generalise the above to the case of $X = U_1 + \dots + U_n$; show that its density may be written

$$f_n(x) = \frac{1}{(n-1)!} \sum_{j=0}^{\lfloor x \rfloor} (-1)^j \binom{n}{j} (x-j)^{n-1}$$

for $0 \leq x \leq n$, where $\lfloor x \rfloor$ is the so-called floor function, the largest integer equal to or to the left of x (so $\lfloor 2.99 \rfloor = 2$, $\lfloor 3.00 \rfloor = 3$, $\lfloor 3.01 \rfloor = 3$). In particular, show that $f_n(x) = x^{n-1}/(n-1)!$ for $0 \leq x \leq 1$.

(c) Draw the densities f_1, f_2, f_3, f_4, f_5 in a diagram, and comment on their forms.

(d) Let now N be the number of uniforms U_j needed in order for their sum to exceed 1. Show that

$$\Pr(N > n) = \Pr(U_1 + \dots + U_n < 1) = F_n(1),$$

with F_n the cumulative for f_n . Also show that in fact $F_n(1) = 1/n!$. Use this to deduce that $\Pr(N = n) = 1/\{n(n-2)!\}$ and that $EN = e$. As a follow-up computational exercise, simulate a high number of such N to calculate the mathematical constant e to three decimal places. Check if you get Ibsen's and Tolstoy's birth years right.

Ex. 1.36 *Characteristic functions.* Above we have found multiple uses for moment-generating functions, as in Ex. 1.32 and 1.33. Another very useful transform is the *characteristic function*, worked with in App. B, see Ex. C.24, which is also a crucial technical tool in the development of the large-sample theory of Ch. 2. For a variable with distribution P , its definition is $\varphi(t) = E \exp(itX) = \int \{\cos(tx) + i \sin(tx)\} dP(x)$, with $dP(x)$ to be read as $f(x) dx$ when P has a density f . Notably, this $\varphi(t)$ always exists, also for distributions without means, etc. We note here the inversion theorem from Ex. C.34, that if $\int |\varphi(t)| dt$ is finite, then there is a density $f(x) = (2\pi)^{-1} \int \exp(-itx) \varphi(t) dt$, see Ex. C.34, which is also seen to be continuous.

the Irwin–Haal
distribution,
from the related
and
concurrently
published
papers Irwin
(1927), Hall
(1927)

(a) For the standard normal, show that $\varphi(t) = \exp(-\frac{1}{2}t^2)$, and that $\varphi(t) = \exp(-\frac{1}{2}\sigma^2 t^2)$ for the $N(0, \sigma^2)$.

(b) Show that $\varphi(t)$ is real, and then equal to $E \cos(tX)$, if and only if the distribution is symmetric around zero.

(c) For U a unit exponential, show that $\varphi(t) = 1/(1 - it)$. Deduce that $X = U - V$, with its Laplace distribution, has $\varphi(t) = 1/(1 + t^2)$, and that

$$\frac{1}{2} \exp(-|x|) = (2\pi)^{-1} \int \cos(tx) \frac{1}{1 + t^2} dt.$$

(d) Show from this, essentially by changing from (x, t) to (t, x) , that the standard Cauchy has $\varphi(t) = \exp(-|t|)$.

(e) In Ex. 1.18(d) we saw that if a ball is kicked from distance r to an infinitely long straight line, with random angle A being uniform on $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$, then the position where the ball crosses the line is $X = rX_0$, where X_0 is Cauchy. Now assume r is not fixed, but comes from a unit exponential distribution. Show that X , a scale mixture of Cauchys, has characteristic function $\phi(t) = 1/(1 + |t|)$.

(f) For V a uniform on $[-1, 1]$, show $\varphi(t) = (\sin t)/t$. Prove from this that even though $\int (\sin t)/t dt = \pi$, the integral $\int |(\sin t)/t| dt$ must be infinite. Show that

$$f(x) = (2\pi)^{-1} \int \cos(tx) \left(\frac{\sin t}{t}\right)^2 dt$$

is equal to the triangular density on $[-2, 2]$.

(g) Just as for m.g.f.s, sums of independent components are associated with products of characteristic functions. Show that if X, Y, Z are independent, with $\varphi_1(t), \varphi_2(t), \varphi_3(t)$, then $X + Y + Z$ has $\varphi_1(t)\varphi_2(t)\varphi_3(t)$ as its characteristic function. When $X_1, \dots, X_n$ are i.i.d. with characteristic function $\varphi(t)$, show that $(X_1 + \dots + X_n)/\sqrt{n}$ has $\varphi_n(t) = \varphi(t/\sqrt{n})^n$. We learn in Ch. 2 that if the X_i have mean zero and finite variance σ^2 , then $\varphi_n(t) \rightarrow \exp(-\frac{1}{2}\sigma^2 t^2)$. In this connection, you may numerically compute

$$f_n(x) = (2\pi)^{-1} \int \cos(tx) \left(\frac{\sin t/\sqrt{n}}{t/\sqrt{n}}\right)^n dt$$

and verify that it tends to the $N(0, 1/12)$ density; cf. Ex. 1.35. Give an approximation to $I_n = \int \{(\sin t)/t\}^n$ as n grows.

(h) Show that there cannot be two i.i.d. variables with sum being unif $[0, 1]$.

Ex. 1.37 *Formulae for the double Cauchy.* In Ex. 1.36 we have learned, with the density f_1 for the standard Cauchy, that

$$f_1(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \text{ is a density with } \varphi_1(t) = \int \frac{1}{\pi} \frac{\cos(tx)}{1 + x^2} dx = \exp(-|t|).$$

(a) With X_0 the unit Cauchy, work with the scaled $X = \sigma X_0$, with density $(1/\pi)(1/\sigma)/\{1 + (x/\sigma)^2\}$. Explain that

$$\int \frac{\cos(tx)}{1 + (x/\sigma^2)} dx = \pi\sigma \exp(-\sigma|t|).$$

Take the derivative w.r.t. σ , at position 1, to show that

$$f_2(x) = \frac{2}{\pi} \frac{1}{(1 + x^2)^2} \text{ is a density with } \varphi_2(t) = \int \frac{2}{\pi} \frac{\cos(tx)}{(1 + x^2)^2} dx = \exp(-|t|)(1 + |t|).$$

(b) Prove the impressive looking formula $\int \cos x/(1+x^2)^3 dx = (7/8)\pi/e$.

Ex. 1.38 *Generating functions.* Moment-generating functions, studying distributions via the transformation $M(t) = E \exp(tX)$, have several close relatives, which might be more convenient for certain classes of distributions. It is e.g. common to use *Laplace transformations* $L(s) = E \exp(-sX)$ for distributions on $[0, \infty)$, then studied for $s \geq 0$. Here we work through the basic properties of *generating functions*, primarily used for distributions on the nonnegative integers. If $\Pr(Y = j) = p_j$, for $j = 0, 1, 2, \dots$, define $G(s) = E s^Y = \sum_{j=0}^{\infty} p_j s^j = p_0 + p_1 s + p_2 s^2 + \dots$, called the generating function for that distribution, or for variables having that distribution.

(a) Show that $G(s) = M(\log s)$, for s such that the latter exists. Demonstrate that $G(s)$ is finite, for $|s| < 1$, and also for $s = 1$. Find the generating functions for (i) the binomial (n, p) ; (ii) the Poisson with parameter θ ; (iii) the geometric with $\Pr(Y = j) = q^{j-1}p$ for $j \geq 1$, with $q = 1 - p$; answers are

$$G_1(s) = (1 - p + ps)^n, \quad G_2(s) = \exp\{-\theta(1 - s)\}, \quad G_3(s) = \frac{ps}{1 - qs},$$

the latter valid for $|s| < 1/q$.

(b) Returning to the general case, give an expression for $G'(s)$, show that $G'(1) = EY$, and that $G''(1) = EY(Y - 1)$. Find the mean and variance for the Poisson using generating functions.

(c) Suppose X and Y are random variables taking on values in $\{0, 1, 2, \dots\}$, and that their generating functions are equal, on an interval around zero. Show that X and Y must have identical distributions.

(d) Show that if X, Y, Z are independent, with generating functions G_1, G_2, G_3 , then the generating function for $X + Y + Z$ is $G_1(s)G_2(s)G_3(s)$. Show from this, and the previous point, yet again, that a sum of independent Poissons is a Poisson.

Ex. 1.39 *Getting Y from two copies of X .* Let X_1, X_2 be independently drawn from some distribution on the nonnegative integers, and from these draw Y from the binomial $(X_1 + X_2, \frac{1}{2})$.

(a) Writing ξ and σ^2 for the mean and variance for the X distribution, set up formulae for the conditional mean and variance of Y , given X_1, X_2 . From these show that $EY = \xi$, $\text{Var } Y = \frac{1}{2}\xi + \frac{1}{2}\sigma^2$.

(b) Write $G_0(s) = E s^{X_i}$ for the generating function of the X_i . Show that the generating function for Y may be written $G(s) = G_0(\frac{1}{2} + \frac{1}{2}s)^2$. When $X_i \sim \text{binom}(m, p)$, show from this that $Y \sim \text{binom}(2m, \frac{1}{2}p)$. Then show that if the X_i are Poisson, then Y reproduces the same distribution.

Ex. 1.40 *Sums of random lengths and the compound Poisson.* Let $X_1, X_2, \dots$ be i.i.d., from a distribution with mean ξ , variance σ^2 , and m.g.f. $M_0(t)$. Consider then a random sum of these random elements; $Z = \sum_{i=1}^N X_i$, where N has some distribution with mean λ , variance τ^2 , and generating function $G(s) = E s^N$. We define Z as zero if $N = 0$.

(a) Show that Z has m.g.f. $M(t) = G(M_0(t))$. Show that Z has mean $\lambda\xi$ and variance $\lambda\sigma^2 + \xi^2\tau^2$.

(b) Consider the so-called compound Poisson variable $Z = \sum_{i=1}^N X_i$, where the X_i are i.i.d. with m.g.f. $M_0(s)$ and $N \sim \text{Pois}(\lambda)$. Show that the m.g.f. of Z may be written

$$E \exp(tZ) = E M_0(s)^N = \exp[\lambda\{M_0(t) - 1\}],$$

Laplace
transforms

generating
functions

compound
Poisson

and that this leads to mean $\lambda\xi$ and variance $\lambda(\xi^2 + \sigma^2)$. Find an expression for $E \exp(tZ)$ for the particular case of $X_i \sim \text{Gam}(a, b)$.

(c) Consider now $Z = \sum_{i=1}^N X_i$, where the X_i are i.i.d. unit exponentials. For given N , the Z is a $\text{Gam}(N, 1)$. With $N \sim \text{geom}(p)$, show that $Z \sim \text{Expo}(p)$. Generalise.

(d) Show via patient derivatives of the $M(t)$ that $\text{skew}(Z) = (1/\lambda)^{1/2}\gamma_3$ and $\text{kurt}(Z) = (1/\lambda)\gamma_4$, with $\gamma_3 = E X^3 / (E X^2)^{3/2}$ and $\gamma_4 = E X^4 / (E X^2)^2$. For a Poisson sum $Z = \sum_{i=1}^N X_i$ of i.i.d. standard normals, so that $Z | (N = n) \sim N(0, n)$, find the variance and kurtosis.

Ex. 1.41 *The logarithmic distribution.* Consider a variable X with point probabilities $\Pr(X = x) = c(p)^{-1}p^x/x$ for $x = 1, 2, \dots$, with p a parameter in $(0, 1)$. This distribution is sometimes called the *logarithmic distribution*.

(a) Show that we must have $c(p) = -\log(1 - p)$. Find expressions for the m.g.f. $M(t)$, its mean, and its variance. Comment on the cases where p is close to zero, or close to one. Show also for its generating function that $G(s) = E s^X = c(ps)/c(p)$.

(b) Consider $Z = \sum_{i=1}^N X_i$, with the X_i being i.i.d. with this logarithmic distribution, and N is Poisson, with parameter expressed as $\lambda c(p)$. Find the mean and variance of Z , and show that its distribution is a negative binomial. We learn that the negative binomial is inside the class of compound Poissons.

The multinormal, the t, the chi-squared, the F

Ex. 1.42 *Mean and variance matrix for a random vector.* Consider a random vector of length p , say $Y = (Y_1, \dots, Y_p)^t$. Its mean vector is defined as the vector of means, i.e. $\xi = EY = (EY_1, \dots, EY_p)^t$, and its *variance matrix* Σ (also often called the covariance matrix), of dimension $p \times p$, has the variances on the diagonal and the covariances outside.

(a) Show that the elements of Σ are $\text{cov}(Y_i, Y_j)$, for $i, j = 1, \dots, p$, and that we may write $\Sigma = E(Y - \xi)(Y - \xi)^t$. If we transform Y to $Z = AY + b$, with A a $m \times p$ matrix and b a vector of length m , show that $EZ = A\xi + b$ and that $\text{Var } Z = A\Sigma A^t$. Explain that this generalises the usual rule $\text{Var}(aY) = a^2 \text{Var } Y$ for one-dimensional variables.

(b) Let X_1, X_2, X_3, X_4 be i.i.d. standard normals. Set up the variance matrix for $(X_1, X_1 + X_2, X_1 + X_2 + X_3, X_1 + X_2 + X_3 + X_4)^t$.

(c) For the multinomial model studied in Ex. 1.5, with $Y = (Y_1, \dots, Y_k)^t$ counting the number of n events that fall in categories $1, \dots, k$, with probabilities $p_1, \dots, p_k$, show that $EY = np$, where p is the vector of $p_1, \dots, p_k$. Show also that its variance matrix can be written $\Sigma = D - pp^t$, where D is diagonal with elements p .

(d) For the multinomial model, with $\mathbf{1}$ the vector $(1, \dots, 1)^t$, show that $\mathbf{1}^t \Sigma \mathbf{1} = 0$. Explain that this is related to the linear relationships $\sum_{j=1}^k p_j = 1$ and $\sum_{j=1}^k Y_j = n$. The matrix is hence not of full rank, and not invertible. Work however with the shorter vector $Y_0 = (Y_1, \dots, Y_{k-1})^t$, and show that its variance matrix may be written $\Sigma_0 = D_0 - p_0 p_0^t$, where $p_0 = (p_1, \dots, p_{k-1})^t$ and D_0 is diagonal with these elements. Show that $\Sigma_0^{-1} = D_0^{-1} + (1/p_k) \mathbf{1}_0 \mathbf{1}_0^t$, where $\mathbf{1}_0$ is the vector of $k - 1$ 1s. For any vector $x = (x_1, \dots, x_k)$ with sum zero, and with x_0 being the shortened version $(x_1, \dots, x_{k-1})$, demonstrate that $x_0^t \Sigma_0^{-1} x_0 = \sum_{j=1}^k x_j^2 / p_j$.

Ex. 1.43 *The multinormal distribution.* Let $Y = (Y_1, \dots, Y_p)^t$ be a random vector of length p . We say that it is multinormally distributed, with mean vector ξ and variance

matrix Σ , which needs to be positive definite, provided its joint density can be written

$$f(y) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(y - \xi)^t \Sigma^{-1} (y - \xi)\},$$

where the domain for y is all of $\mathbb{R}^p$. We write $Y \sim N_p(\xi, \Sigma)$ to indicate this distribution.

(a) Show that $\int y f(y) dy$ indeed is equal to ξ , so calling it the mean vector is appropriate. Show also that $E(Y - \xi)(Y - \xi)^t$, calculated from the density, is equal to Σ .

(b) Show that if $Y \sim N_p(\xi, \Sigma)$, then $Y - \xi \sim N_p(0, \Sigma)$.

(c) Assume now that A is an invertible $p \times p$ matrix, and consider the transformation $Z = AY$. Show that if $Y \sim N_p(\xi, \Sigma)$, then $Z = AY \sim N_p(A\xi, A\Sigma A^t)$.

(d) By the spectral decomposition theorem of linear algebra, there is an orthonormal matrix P , with $PP^t = I = P^tP$, such that $P\Sigma P^t = D = \text{diag}(\lambda_1, \dots, \lambda_p)$, with these values being the eigenvalues of Σ . Show that $Z = P(Y - \xi)$ has components $Z_1, \dots, Z_p$ which are independent, with $Z_j \sim N(0, \lambda_j)$.

(e) Show that a vector Y is multinormal if and only if all linear combinations are normal. In particular, if $Y \sim N_p(\xi, \Sigma)$, then $V = c^t Y = c_1 Y_1 + \dots + c_p Y_p$ is normal $N(c^t \xi, c^t \Sigma c)$. We allow ourselves treating constants as normals, i.e. $N(\xi_0, \sigma^2)$ with $\sigma_0 = 0$.

(f) Generalise point (c) to state that with any matrix A , of size say $q \times p$, the transformed $Z = AY$ is a multinormal $N_q(A\xi, A\Sigma A^t)$.

(g) For the binormal case, with means ξ_1, ξ_2 , standard deviations σ_1, σ_2 , and correlation ρ , show that the density may be written

$$f(x, y) = \frac{1}{2\pi} \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)^{1/2}} \exp \left[-\frac{1}{2} \frac{1}{1 - \rho^2} \left\{ \left(\frac{x - \xi_1}{\sigma_1} \right)^2 + \left(\frac{y - \xi_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x - \xi_1}{\sigma_1} \right) \left(\frac{y - \xi_2}{\sigma_2} \right) \right\} \right].$$

Show that X and Y are independent if and only if the correlation is zero.

(h) We learn that the situation is easy and clean for the multinormal case, where independence is equivalent to zero correlation. This is in general more complicated. (i) Let $X \sim N(0, 1)$ and $Y = DX$, with D a random sign with equal probabilities for $-1, 1$. Show that Y also is standard normal, that the correlation is zero, but that they are dependent. (ii) For another example, again with $X \sim N(0, 1)$, let $Y = X$ if $|X| \leq a$ but $Y = -X$ if $|X| > a$. Show that Y is standard normal. Show that the correlation $\rho(a)$ goes from -1 to 1 as a goes from zero to infinity; thus there is an a_0 for which the correlation is zero. Also, find this a_0 numerically (answer: 1.537). In these zero-correlation constructions, with normal marginals, the point is that there is not joint binormality.

Ex. 1.44 *The multinormal and conditional distributions.* Consider a multinormally distributed vector, of length $p + q$, blocked into subvectors of sizes p and q . Let us write this as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_{p+q} \left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

(a) By carrying out a linear transformation, and using results from Ex. 1.43, show that

$$Z = Y_1 - \Sigma_{12} \Sigma_{22}^{-1} Y_2 \sim N_p(\xi_1 - \Sigma_{12} \Sigma_{22}^{-1} \xi_2, \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}),$$

and that this Z is independent of Y_2 .

(b) Show that the distribution of Y_1 given $Y_2 = y_2$ must be multinormal. Derive the important formulae for the conditional mean and variance,

$$E(Y_1 | y_2) = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \xi_2), \quad \text{Var}(Y_1 | y_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Note that the conditional mean is a linear function in y_2 and that the conditional variance matrix is constant, not depending on y_2 .

(c) Now study the simplest two-dimensional prototype case, with

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Show that $Y|x \sim N(\rho x, 1 - \rho^2)$ and that $X|y \sim N(\rho y, 1 - \rho^2)$. Discuss implications for situations where an easy to measure X might be a proxy for a harder to come by Y . How can you estimate Y from X , and with what precision?

(d) In generalisation of this special binormal case, consider a binormal

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} \xi \\ \xi_0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\sigma_0 \\ \rho\sigma\sigma_0 & \sigma_0^2 \end{pmatrix}\right).$$

In particular, the X is seen as stemming from a $N(\xi_0, \sigma_0^2)$ distribution. Show that $Y|(X = x)$ is normal, with constant variance $\sigma^2(1 - \rho^2)$ and linear mean function $E(Y|x) = \xi + \rho(\sigma/\sigma_0)x$. This is essentially the linear regression model, for $Y|x$, here derived as a consequence of binormality. Without pretensions of having precise numbers, guess the binormal parameters in your population for (X, Y) being height and weight, and deduce from this a model predicting a person's weight from his or her height.

(e) Generalise this to situations with a $(p + 1)$ -dimensional normal distribution for $(Y, X_1, \dots, X_p)$, with Y seen as the main outcome, influenced by covariates $X_1, \dots, X_p$. Show that Y given these covariates is normal, with a constant variance, and find the linear conditional mean function $E(Y | X_1, \dots, X_p)$. This is essentially the linear multiple regression model; see Ex. 3.33.

Ex. 1.45 *How tall is Nils?* Assume that the heights of Norwegian men above the age of twenty follow the normal distribution $N(\xi, \sigma^2)$ with $\xi = 180$ cm and $\sigma = 9$ cm.

(a) Given this information only, what is your point estimate of his height, and what is your 95 percent prediction interval?

(b) Assume now that you learn that his four brothers are actually 195 cm, 207 cm, 196 cm, 200 cm tall, and furthermore that correlations between brothers' heights in the population of Norwegian men is equal to $\rho = 0.80$. Use this information about his four brothers to revise your initial point estimate of his height, and provide the updated 95 percent prediction interval. Is Nils a statistical outlier in his family?

(c) Suppose that Nils has n brothers and that you learn their heights. Give formulae for the updated normal parameters ξ_n and σ_n , in the conditional distribution of his height given these extra pieces of information. Use this to clarify the following statistical point: Even if you get to know all facts concerning 99 brothers, there should be a limit to your confidence in what you may infer about Nils.

Ex. 1.46 *X is bigger than Y and Y is bigger than X.* Starting with (U, V) be binormal with zero means, unit variances, and correlation $\rho \in (0, 1)$, define $X = \exp(\frac{1}{2}U^2)$ and $Y = \exp(\frac{1}{2}V^2)$.

- (a) With $X = x$, let u be the positive solution to $\frac{1}{2}u^2 = \log x$. Given $X = x$, explain that V is either a $N(\rho u, 1 - \rho^2)$ or a $N(-\rho u, 1 - \rho^2)$, with 50-50 probabilities.
- (b) For the first case, write $V = \rho u + (1 - \rho^2)^{1/2}N$, with N a standard normal, and work out that

$$E \exp(\frac{1}{2}V^2 | u) = \exp(\frac{1}{2}u^2)/\rho.$$

Since that answer is the same, for $\pm u$, explain that $E(Y | X) = X/\rho$, and by symmetry that $E(X | Y) = Y/\rho$. So, apparently, X is bigger than Y on average, and Y is bigger than X on average. Discuss why this does not need to be a paradox, and supply some plots to illustrate what goes on.

Ex. 1.47 *Simulating from the multinormal.* For given ξ and Σ , how can we simulate realisations y from the $N_p(\xi, \Sigma)$? We take it for granted here that yes, we can simulate from the one-dimensional $N(\xi, \sigma^2)$, e.g. via `xi + sigma*rnorm(100)` in R. For simplicity we take $\xi = 0$ here, without loss of generality.

- (a) From details of Ex. 1.43, there is a spectral decomposition, with $P\Sigma P^t = D$, with the diagonal D having the eigenvalues of Σ and P orthonormal, i.e. $PP^t = I_n = P^tP$, with I_n the identity matrix. Show that with $D^{1/2}$ the diagonal with the square-roots of the eigenvalues, and with $\Sigma^{1/2} = P^t D^{1/2} P$, this is a matrix whose square is Σ .
- (b) Explain from this that with $N = (N_1, \dots, N_p)^t$ a string of i.i.d. standard normals, $Y = \Sigma^{1/2}N \sim N_p(0, \Sigma)$.

(c) Explain from Ex. 1.44 that a perhaps long multinormal vector (Y_1, Y_2) , with lengths p and q , can be simulated in blocks; first Y_1 , then $Y_2 | Y_1$. This is sometimes practical in high dimensions.

(d) To attempt this in practice, simulate say $n = 50$ copies $Y_i = (Y_{i,1}, \dots, Y_{i,p})^t$, in dimension say $p = 10$, with all variables being i.i.d. standard normal. From these compute the empirical variance matrix Σ_n , with empirical variances on the diagonal and covariances outside, i.e. $\sigma_{j,k} = (n-1)^{-1} \sum_{i=1}^n (Y_{i,j} - \bar{Y}_j)(Y_{i,k} - \bar{Y}_k)$ for $j, k = 1, \dots, p$. Then simulate from the multinormal $N_p(0, \Sigma_n)$. Check to see if their empirical variance matrix is close to Σ_n again.

Ex. 1.48 *The chi-squared.* This exercise goes through some basic properties of the chi-squared; see also Ex. 1.52 for its eccentric cousin, the noncentral or eccentric chi-squared.

- (a) We say that a nonnegative variable X has the *chi-squared distribution*, with degrees of freedom m , and write $X \sim \chi_m^2$ for this, when its density takes the form

$$g_m(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} \exp(-\frac{1}{2}x) \quad \text{for } x > 0.$$

Show that its m.g.f. becomes $M(t) = (1-2t)^{-m/2}$, for $t < \frac{1}{2}$. Show further that $E X = m$, $\text{Var } X = 2m$, and that the skewness, i.e. $E W^3$ with $W = (X - E X)/(\text{Var } X)^{1/2} = (X - m)/(2m)^{1/2}$, is $(8/m)^{1/2}$. From the density, show $E(\chi_m^2)^p = 2^p \Gamma(m/2 + p)/\Gamma(m/2)$.

(b) Via the m.g.f., show the simple and basic convolution property for the chi-squared, that if $X_1, \dots, X_n$ are independent and chi-squared distributed with degrees of freedom $m_1, \dots, m_n$, then the sum $Z = \sum_{i=1}^n X_i$ is chi-squared too, with degrees of freedom $\sum_{i=1}^n m_i$. A generalisation is given in Ex. 1.52.

(c) If N is standard normal, show that $X = N^2 \sim \chi_1^2$. Establish that if $X \sim \chi_m^2$, with m a natural number, then it may be represented as $X = N_1^2 + \dots + N_m^2$, in terms of

there are
simulation
algorithms in
packages, like
`mvrnorm` om R,
but the recipe
here works
better for large
dimension

independent standard normals $N_1, \dots, N_m$. Note, however, that the χ_m^2 with the density $g_m(x)$ above may be used also when m is not a natural number.

(d) There are connections between the chi-squared and the Gamma distribution (see Ex. 1.9). Show that the $\text{Gam}(\frac{1}{2}m, \frac{1}{2})$ is the χ_m^2 ; that if $Y \sim \text{Gam}(a, b)$, then $X = 2bY \sim \chi_{2a}^2$; and that if $Z \sim \chi_m^2$, then $\frac{1}{2}Z \sim \text{Gam}(\frac{1}{2}m, 1)$.

(e) Consider independent $X \sim \chi_a^2$ and $Y \sim \chi_b^2$. Show, perhaps via Gamma distribution ratios in Ex. 1.12, that $R = X/(X + Y) \sim \text{Beta}(\frac{1}{2}a, \frac{1}{2}b)$.

(f) When $X \sim \chi_m^2$, show that $E \log X = \log 2 + \psi(\frac{1}{2}m)$, where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function.

(g) Consider $W = AB$, a product of independent standard normals. Use (1.2) to show that its m.g.f. is $M(t) = 1/(1-t^2)^{1/2}$, for $|t| < 1$. Deduce from this that W has the same distribution as $\frac{1}{2}(K - L)$, where K and L are independent with χ_1^2 distributions.

(h) Then consider $W = AB + CD$, where A, B, C, D are independent standard normals. Show that W has density $\frac{1}{2} \exp(-|w|)$, i.e. the Laplace distribution worked with in Ex. 1.34.

(i) Consider the second degree equation $x^2 + Bx + C = 0$ from school. If there are many such equations, with B and C being independent standard normal, how many of these equations will have both roots real?

Ex. 1.49 Orthonormal transformations. We have seen in Ex. 1.43 that a multinormal vector can be sent via a linear transformation to independent one-dimensional normal components, and vice versa. This also leads to useful characterisation and representation theorems involving independence. In the present exercise we shall e.g. find a proof that the sample mean $\bar{Y}$ and the sample variance statistic $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2$ are independent; this classic fact, which does not hold outside the normal family, is actively used in Ex. 3.4, Ex. 3.6, and in many other situations in our book.

(a) Suppose $X = (X_1, \dots, X_n)$ is a vector with i.i.d. and standard normal components, and let A be an orthonormal matrix, which means $AA^t = I = A^t A$. In yet other words, each row of A and each column of A has length 1, rows are orthogonal, as well as columns. Show that $Y = AX$ must have components $Y_1, \dots, Y_n$ which are also i.i.d. and standard normal. – Here you may also use the general transformation formula of Ex. 1.12.

(b) To exemplify the above, show that if X_1, X_2 are independent and standard normal, then also Y_1, Y_2 , where

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} (X_1 + X_2)/\sqrt{2} \\ (X_1 - X_2)/\sqrt{2} \end{pmatrix},$$

must be independent and standard normal.

(c) When A is orthonormal, show that it preserves length, so $\|Au\| = \|u\|$, for any vector u ; here $\|u\|$ is Euclidean length, so $\|u\|^2 = u_1^2 + \dots + u_n^2$.

(d) Let again $X_1, \dots, X_n$ be i.i.d. standard normals. Construct an orthogonal matrix A by letting its first row be $(1/\sqrt{n}, \dots, 1/\sqrt{n})$, and define $Y = AX$. Then show that $Y_1 = \sqrt{n}\bar{X} = \sum_{i=1}^n X_i/\sqrt{n}$, and that $Z_0 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=2}^n Y_i^2$.

(e) Conclude from this that (i) $\sqrt{n}\bar{X} \sim N(0, 1)$, (ii) $Z_0 \sim \chi_{n-1}^2$, and (iii) $\bar{X}$ and Z_0 are independent. This is proven more directly in Ex. 1.54 below, via decompositions of chi-squares.

(f) Show that this implies the following classical and important properties, starting with an independent sample $Y_1, \dots, Y_n$ from the $N(\xi, \sigma^2)$: The statistics $\bar{Y}$ and $Z =$

$\sum_{i=1}^n (Y_i - \bar{Y})^2$ are independent, with $\bar{Y} \sim N(\xi, \sigma^2/n)$ and $Z \sim \sigma^2 \chi_{n-1}^2$. Show also from this that the classical empirical variance

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (1.6)$$

is unbiased for the population variance, i.e. $E \hat{\sigma}^2 = \sigma^2$. Construct an unbiased estimator for σ , of the type $c_n \hat{\sigma}$.

(g) Consider the general multinormal distribution $Y \sim N_p(\xi, \Sigma)$, with invertible Σ . Show that $K = (Y - \xi)^t \Sigma^{-1} (Y - \xi) \sim \chi_p^2$. Suppose $Y = y_{\text{obs}}$ is observed, that Σ is known, but ξ unknown. Give a confidence region R such that $\xi \in R$ with probability 90 percent. How does this region shrink, if you observe 100 vectors from the multinormal, rather than merely 1?

Ex. 1.50 *The t distribution.* Consider independent variables $X \sim N(0, 1)$ and $K \sim \chi_m^2$. The ratio $t = X/(K/m)^{1/2}$ is then said to have *the t distribution*, with m degrees of freedom. We write $t \sim t_m$ to indicate this.

(a) Looking at the denominator first, show that K/m has mean 1 and variance $2/m$, indicating that it will be close to 1 with high probability when m is moderate or big. Explain that this implies that t is not far from having a standard normal distribution for growing m . Finer details regarding this are in Ch. 2.

(b) Find the mean and variance of t .

(c) Show that its density can be written

$$g_m(x) = \frac{\Gamma((m+1)/2)}{\Gamma(m/2)} \frac{1}{\sqrt{m}\pi} \frac{1}{(1+x^2/m)^{(m+1)/2}}.$$

Show that $g_m(x)$ tends to the standard normal density $\phi(x)$ as m increases, matching the intuition above. Show also that the Cauchy distribution, see Ex. 1.18, is the t_1 distribution.

(d) Find also the skewness and kurtosis for the t_m distribution. In particular, show that the latter is $\text{kurt} = 6/(m-4)$ for $m > 4$.

(e) Assume $Y_1, \dots, Y_n$ are i.i.d. $N(\xi, \sigma^2)$. With $\hat{\sigma}$ the empirical standard deviation, from (1.6), show that

$$t = (\bar{Y} - \xi)/(\hat{\sigma}/\sqrt{n})$$

has the t-distribution with $n-1$ degrees of freedom. This is the classic t-statistic dating all the way back to [Student \(1908\)](#),

Ex. 1.51 *The noncentral t distribution.* Consider $X \sim N(\xi, 1)$, independent of $K \sim \chi_m^2$. The ratio $t = X/(K/m)^{1/2} = (\xi + N(0, 1))/(\chi_m^2/m)^{1/2}$ is then said to follow the *noncentral t distribution*, written $t \sim t_m(\xi)$, with degrees of freedom m and noncentrality parameter ξ . Its distribution is a complex one, but both the density $g_m(t, \xi)$ and c.d.f. $G_m(t, \xi)$ are implemented in various packages, as e.g. `dt(t,m,ksi)` and `pt(t,m,ksi)` in R.

(a) Explain that the classical t distribution of Ex. 1.50 corresponds to the special case of $\xi = 0$. Via arguments in that exercise, explain also that the $t_m(\xi)$ becomes close to a $N(\xi, 1)$ for growing m .

(b) Find expressions for the mean and variance of the $t_m(\xi)$.

(c) Consider again an i.i.d. sample $Y_1, \dots, Y_n$ from $N(\xi, \sigma^2)$, with sample mean $\bar{Y}$ empirical standard deviation $\hat{\sigma}$, as in Ex. 1.50. Show that $t = \sqrt{n}\bar{Y}/\hat{\sigma}$ has the $t_{n-1}(\xi/\sigma)$ distribution. We shall need this for studying the power of the t test in Ex. 4.13.

Ex. 1.52 *The noncentral chi-squared.* Consider also the so-called *noncentral chi-squared distribution*, say $K \sim \chi_m^2(\lambda)$, with λ the excentre or eccentricity parameter; the case of $\lambda = 0$ corresponds to the ordinary $K \sim \chi_m^2$. It is the distribution of $Y_1^2 + \dots + Y_m^2$, where the Y_i are independent normals, with $Y_i \sim N(\xi_i, 1)$, and $\lambda = \sum_{i=1}^m \xi_i^2$.

(a) Show that the m.g.f. of $K \sim \chi_m^2(\lambda)$ may be written

$$M(t) = E \exp(tK) = \frac{\exp\{\lambda t/(1-2t)\}}{(1-2t)^{m/2}} \quad \text{for } t < \frac{1}{2}.$$

(b) Also the noncentral chi-squared distributions have convolution properties, generalising those of Ex. 1.48. If $K_i \sim \chi_{m_i}^2(\lambda_i)$, and these are independent, for $i = 1, \dots, n$, show that $\sum_{i=1}^n K_i$ is another noncentral chi-squared, with degrees of freedom $\sum_{i=1}^n m_i$ and excentre parameter $\sum_{i=1}^n \lambda_i$.

(c) Its density can be expressed in several ways; show that this is one such valid formula:

$$f(k, m, \lambda) = \sum_{j=0}^{\infty} \left\{ \exp(-\frac{1}{2}\lambda) (\frac{1}{2}\lambda)^j / j! \right\} g_{m+2j}(k),$$

where $g_{m+2j}(k)$ is the χ_{m+2j}^2 density. In other words, the noncentral chi-squared is a Poisson mixture of central chi-squared distributions. Show that this entails the representation $K | (J = j) \sim \chi_{m+2j}^2$, where $J \sim \text{Pois}(\frac{1}{2}\lambda)$. Also non-integer values of m are allowed here.

(d) Establish that for $K \sim \chi_m^2(\lambda)$, we have $E K = m + \lambda$ and $\text{Var } K = 2m + 4\lambda$. Show also that the skewness of K becomes $2^{3/2}(m + 3\lambda)/(m + 2\lambda)^{3/2}$. What is required in order for this skewness to tend to zero?

(e) With N a standard normal, let $K = (\lambda^{1/2} + N)^2$, which has the $\chi_1^2(\lambda)$ distribution. Consider the normalised variable

$$\frac{K - (1 + \lambda)}{(2 + 4\lambda)^{1/2}} = \frac{N^2 + 2\lambda^{1/2}N - 1}{(2 + 4\lambda)^{1/2}}.$$

Work out its m.g.f. and show that it tends to $\exp(\frac{1}{2}t^2)$ for growing λ .

(f) More generally, with $K \sim \chi_m^2(\lambda)$, work out a formula for the m.g.f. $M(t)$ for $Z = \{K - (m + \lambda)\}/(2m + 4\lambda)^{1/2}$. For any fixed m , show that $M(t) \rightarrow \exp(\frac{1}{2}t^2)$ as λ grows, and comment on this finding.

Ex. 1.53 *Noncentral chi-squared for empirical variances.* We saw in Ex. 1.49 that if $X_1, \dots, X_n$ are i.i.d. $N(a, 1)$, with common a , then $Z = \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2$, with consequences for the empirical variance estimator. Here are some fruitful generalisations.

(a) Let the X_i have non-identical means, $X_i \sim N(a_i, 1)$. Show that $Z \sim \chi_{n-1}^2(\lambda)$, with noncentrality parameter $\lambda = \sum_{i=1}^n (a_i - \bar{a})^2$.

(b) Assume now that $X_i \sim N(a_i, 1/m_i)$ for $i = 1, \dots, n$, perhaps reflecting sample sizes m_i for different groups, and with $M = \sum_{i=1}^n m_i$. Consider $Z = \sum_{i=1}^n m_i (X_i - \tilde{X})^2$, with $\tilde{X} = \sum_{i=1}^n (m_i/M) X_i$. Show that $Z = \sum_{i=1}^n m_i X_i^2 - M \tilde{X}^2$, and that its distribution is a $\chi_{n-1}^2(\lambda)$, with $\lambda = \sum_{i=1}^n m_i (a_i - \tilde{a})^2$, where $\tilde{a} = \sum_{i=1}^n (m_i/M) a_i$.

(c) Let $X \sim N_p(\xi, \Sigma)$. Show that $Z = X^t \Sigma^{-1} X \sim \chi_p^2(\xi^t \Sigma^{-1} \xi)$.

Ex. 1.54 *Tell me about X and $X + Y$.* In exercises above we have seen that sums of independent normals, Poissons, chi-squares are respectively normal, Poisson, chi-square. Sometimes we meet questions going the other way.

(a) Suppose X and Y are independent, and that you learn the distributions of X and $Z = X + Y$. Show that Y must have m.g.f. $M_Y(t) = M_Z(t)/M_X(t)$. As an illustration, suppose $X \sim N(0, 1)$ and $X + Y \sim N(0, 2)$. Show that $Y \sim N(0, 1)$.

(b) Explain similarly that if X and Y are independent, with X and $X + Y$ being Poisson, then Y is also Poisson.

(c) Suppose X and Y are independent, that $X \sim \chi_a^2$ and that $Z = X + Y \sim \chi_{a+b}^2$. Show that the only possibility is then the expected one, that $Y \sim \chi_b^2$. More generally, if $X \sim \chi_a^2(\lambda_1)$ and $X + Y \sim \chi_{a+b}^2(\lambda_1 + \lambda_2)$, with independence, then $Y \sim \chi_b^2(\lambda_2)$.

(d) Consider $X_1, \dots, X_n$ i.i.d. standard normal. Writing as usual $\bar{X}$ for the sample mean, we know that $\sqrt{n}\bar{X}$ is also standard normal. Show that the vector of $X_i - \bar{X}$ is independent of $\bar{X}$; for that detail, argue via multinormality and Ex. 1.43. Writing $Z_0 = \sum_{i=1}^n (X_i - \bar{X})^2$, show from $\sum_{i=1}^n X_i^2 = Z_0 + n\bar{X}^2$ that $Z_0 \sim \chi_{n-1}^2$. (This is perhaps a simpler proof than the one found in Ex. 1.49.) Generalise: suppose $Y_i \sim N(\mu_i, 1)$, and let $\bar{\mu} = n^{-1} \sum_{i=1}^n \mu_i$. Show from

$$Q = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n \{Y_i - \bar{\mu} - (\bar{Y} - \bar{\mu})\}^2 = \sum_{i=1}^n (Y_i - \bar{\mu})^2 - n(\bar{Y} - \bar{\mu})^2$$

that we must have $Q \sim \chi_{n-1}^2(\sum_{i=1}^n (\mu_i - \bar{\mu})^2)$.

(e) A rather more general multidimensional extension of the above is as follows. Consider independent $Y_j \sim N_p(\theta, \Sigma_j)$ for $j = 1, \dots, k$. Show that the minimiser of $Q(\theta) = \sum_{j=1}^k (Y_j - \theta)^t \Sigma_j^{-1} (Y_j - \theta)$ is $\hat{\theta} = A^{-1} \sum_{j=1}^k \Sigma_j^{-1} y_j$, with $A = \sum_{j=1}^k \Sigma_j^{-1}$, and use this to infer that $\hat{\theta} \sim N_p(\theta, A^{-1})$. Furthermore, show

$$Q(\theta) = Q_{\min} + (\hat{\theta} - \theta)^t A (\hat{\theta} - \theta),$$

with $Q_{\min} = Q(\hat{\theta})$, and explain that its distribution is $\chi_{(k-1)p}^2$.

Ex. 1.55 *The F distribution.* As we have seen Ex. 1.49, with a normal sample $X_1, \dots, X_n$, the distribution of the classical empirical variance estimator $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is governed by $\hat{\sigma}^2 \sim \sigma^2 \chi_m^2/m$, where $m = n-1$ is the degrees of freedom. Suppose there are two independent samples, from normal distributions $N(\xi_1, \sigma_1^2)$ and $N(\xi_2, \sigma_2^2)$, of sample sizes n_1 and n_2 , with estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.

(a) Let $\rho = \sigma_1/\sigma_2$, the ratio of standard deviations. For the ratio of the two empirical variances, show that

$$R^2 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \rho^2 F, \quad \text{where } F \sim \frac{\chi_{m_1}^2/m_1}{\chi_{m_2}^2/m_2},$$

with degrees of freedom $m_1 = n_1 - 1$ and $m_2 = n_2 - 1$, and with the two chi-squareds being independent. We say that F has the *F distribution*, or Fisher distribution, with degrees of freedom (m_1, m_2) , and write $F \sim F(m_1, m_2)$.

(b) Show that when $F \sim F(m_1, m_2)$, then $1/F \sim F(m_2, m_1)$. Show furthermore that

$$E F = \frac{m_2}{m_2 - 2}, \quad E F^2 = \frac{m_1 + 2}{m_1} \frac{m_2^2}{(m_2 - 2)(m_2 - 4)},$$

independence is
equivalent to
zero correlation
for the
multinormal
world

these expressions being finite when $m_2 > 2$ and $m_2 > 4$, respectively. Find also an expression for the variance. Verify that both $E F$ and $E F^2$ tend to 1 as the degrees of freedom increase.

(c) The main aspects of the F distribution have been worked out, above, from the constructive definition given in (a), without actually needing any formula for its density; also, probabilities are found using software packages, like `pf(x,m1,m2)` in R. Once in a while one needs the density function, however. Show first that the cumulative function can be written

$$\Pr(F \leq x) = \Pr(\chi_{m_1}^2/m_1 \leq x\chi_{m_2}^2/m_2) = \int_0^\infty G(xy(m_1/m_2), m_1)g(y, m_2) dy,$$

in terms of the cumulative $G(\cdot, m)$ and density $g(\cdot, m)$ of the χ_m^2 . Take the derivative,

$$h(x, m_1, m_2) = \int_0^\infty g(xy(m_1/m_2), m_1)y(m_1/m_2)g(y, m_2) dy.$$

Complete the math to land at

$$h(x, m_1, m_2) = \frac{\Gamma(\frac{1}{2}(m_1 + m_2))}{\Gamma(\frac{1}{2}m_1)\Gamma(\frac{1}{2}m_2)} (m_1/m_2)^{m_1/2} \frac{x^{m_1/2-1}}{\{1 + (m_1/m_2)x\}^{(m_1+m_2)/2}}.$$

Ex. 1.56 Which of the two is closer to zero? Suppose (A, B) is binormal with zero means, standard deviations τ_1, τ_2 , and correlation ρ . In the context of later chapters, the (A, B) could be the limiting distribution associated with two estimators. There is no closed-form formula for the general $\text{pr} = \Pr(|B| \leq |A|)$, but we shall find a one-dimensional integral form which will then be easy to evaluate numerically.

(a) Write $A = \tau_1 X$ and $B = \tau_2 Y$, for standard normals (X, Y) with correlation ρ . Looking first at the zero-correlation case, show that

$$\text{pr} = \Pr(Y^2 \leq \lambda^2 X^2) = F_{1,1}(\lambda^2), \quad \text{with } \lambda = \tau_1/\tau_2,$$

in terms of the c.d.f. for the $F_{1,1}$ distribution. Check that this is $\frac{1}{2}$ for $\tau_1 = \tau_2$.

(b) For the general case, show via conditioning on X that $\text{pr} = \Pr(|Y| \leq \lambda|X|) = \int g(x)\phi(x) dx$, in which

$$g(x) = \Pr(|\rho X + (1 - \rho^2)^{1/2} Z| \leq \lambda|x|) = \Phi\left(\frac{\lambda|x| - \rho x}{(1 - \rho^2)^{1/2}}\right) - \Phi\left(\frac{-\lambda|x| - \rho x}{(1 - \rho^2)^{1/2}}\right),$$

with Z a standard normal. Verify that $\text{pr} = \frac{1}{2}$ for the case of equal variances, irrespective of ρ . For an illustration of the general formula, check with Story #97.

The exponential family class

Ex. 1.57 *The exponential family class.* Many parametric models fall under the wide umbrella of the exponential family class, which we treat in this and the following exercises. This will be properly generalised and extended down the road, but we start with this definition: Suppose Y has model density of the form

$$\begin{aligned} f(y, \theta) &= \exp\{\theta_1 T_1(y) + \cdots \theta_p T_p(y) - k(\theta_1, \dots, \theta_p)\} h(y) \\ &= \exp\{\theta^t T(y) - k(\theta)\} h(y), \end{aligned} \tag{1.7}$$

exponential
family

for appropriate functions $T_1(y), \dots, T_p(y)$ and $h(y)$; the $k(\theta)$ function is there to secure integration to one. We require that there is no fixed linear relationship among $T_1(y), \dots, T_p(y)$, and also that the support of the distribution, the smallest closed set having probability 1, is the same, for all parameter values. We then say Y is of the exponential family class, with *data functions* $T(y) = (T_1(y), \dots, T_p(y))^t$ and *natural parameters* $\theta = (\theta_1, \dots, \theta_p)^t$.

(a) Before we start developing the general theory for the full class, we verify that a few classic models are under its umbrella. For the following models, write the model density in a form matching (1.7). (i) $Y \sim \text{binom}(n, p)$. (ii) $Y \sim \text{Pois}(\theta)$. (iii) $Y \sim \text{Beta}(a, b)$. (iv) $Y \sim \text{Gam}(a, b)$. (v) $Y \sim N(\xi, \sigma^2)$, with known σ (and see Ex. 1.58). (vi) Let (X, Y, Z) be trinomial (n, p, q, r) , with $r = 1 - p - q$. First work with $f(x, y, p, q)$, and then with the submodel where $p = a^2$, $q = 2a(1 - a)$, $r = (1 - a)^2$.

(b) Show that we must have

$$k(\theta) = \log \left(\int \exp\{\theta^t T(y)\} h(y) d\nu(y) \right),$$

assumed to be finite for at least some θ , and with $d\nu(y)$ the dominating measure (often enough simply dy , for models with continuous densities). Let in fact Θ be the set of θ such that $k(\theta)$ is finite, called the natural parameter region. Show that Θ is convex.

(c) The score function of a model is $u(y, \theta) = \partial \log f(y, \theta) / \partial \theta$; see Ex. 5.6 for more on this. For the class of models studied here, show that $u(y, \theta) = T(y) - \xi(\theta)$, where

$$\xi(\theta) = \frac{\partial k(\theta)}{\partial \theta} = \frac{\int T(y) \exp\{\theta^t T(y)\} h(y) dy}{\int \exp\{\theta^t T(y)\} h(y) dy}.$$

Show that the score function must have mean zero, which here means $E_\theta T(Y) = \xi(\theta)$. Show also that $\text{Var}_\theta T(Y) = \partial^2 k(\theta) / \partial \theta \partial \theta^t$, giving variances and covariances of the $T_j(Y)$ in one matrix formula.

(d) We come back to general likelihood theory in Ch. 5, but for now show that if $Y_1, \dots, Y_n$ are i.i.d. from the exponential family density, then the logarithm of the joint density can be written

$$\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta) = n\{\theta^t \bar{T} - k(\theta)\},$$

with $\bar{T} = (1/n) \sum_{i=1}^n T(Y_i)$ the vector of averages $\bar{T}_j = (1/n) \sum_{i=1}^n T_j(y_i)$. If this is a family with say $p = 3$ parameters, and $n = 10000$, then the full relevant information required for computing the ℓ_n function is captured in the 3 averages $\bar{T}_1, \bar{T}_2, \bar{T}_3$. This is related to the sufficiency concept, returned to in Ch. 4. Show also that this $\ell_n(\theta)$ is a concave function.

Ex. 1.58 *The normal and binormal exponential family members.* The normal and binormal (and multinormal) classes of distributions belong under the exponential family umbrella.

(a) Consider first the normal (ξ, σ^2) model. Explain that its density may be written

$$f(y) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2} \frac{y^2}{\sigma^2} + \frac{\xi y}{\sigma^2} - \frac{1}{2} \frac{\xi^2}{\sigma^2} - \log \sigma\right),$$

and argue from this that the normal is of exponential family type, with natural parameters ξ/σ^2 and $1/\sigma^2$, and data functions equivalent to (y, y^2) .

natural
parameter
region

(b) Consider then (X, Y) binormal, with its five parameters $\xi_1, \xi_2, \sigma_1, \sigma_2, \rho$, see Ex. 1.43. Show that the distribution is of the exponential family type, with natural parameters

$$\frac{1}{1-\rho^2} \left(\frac{\xi_1}{\sigma_1^2} - \frac{\rho \xi_2}{\sigma_1 \sigma_2} \right), \frac{1}{1-\rho^2} \left(\frac{\xi_2}{\sigma_2^2} - \frac{\rho \xi_1}{\sigma_1 \sigma_2} \right), \frac{1}{(1-\rho^2)\sigma_1^2}, \frac{1}{(1-\rho^2)\sigma_2^2}, \frac{\rho}{(1-\rho^2)\sigma_1 \sigma_2}.$$

In the terminology of exponential families, identify also the associated $T_1(x, y), \dots, T_5(x, y)$.

(c) Show that $(\xi_1, \xi_2, \sigma_1, \sigma_2, \rho)$ is in one-to-one correspondence with the five exponential class natural parameters above.

(d) Suppose now that $\xi_1 = \xi_2$ and $\sigma_1 = \sigma_2$, so this distribution for (X, Y) has now three parameters. Show that it is inside the exponential family, and identify its natural parameters.

Ex. 1.59 *A log-linear density model on the unit interval.* The machinery of the exponential family class makes it easy to construct new models, starting with relevant data functions T_j , in the language of Ex. 1.57.

(a) For densities on the unit interval, start with data function $T(y) = y - \frac{1}{2}$, to define $f(y, \theta) = \exp\{\theta T(y) - k(\theta)\}$ for $y \in [0, 1]$. Find the $k(\theta)$, a formula for the c.d.f. $F(y, \theta)$, and the mean and variance for Y having this density.

(b) Then go on to

$$f(y, \theta_1, \theta_2) = \exp\{\theta_1(y - \frac{1}{2}) + \theta_2(y - \frac{1}{2})^2 - k(\theta_1, \theta_2)\},$$

i.e. using $T_j(y) = (y - \frac{1}{2})^j$ as data functions, for $j = 1, 2$. Set up an integral function for $k(\theta_1, \theta_2)$ and some algorithms for determining the means, variances, and covariance, for $T_1(Y), T_2(Y)$, for any given (θ_1, θ_2) .

(c) Spell out the necessary details for the third order log-linear density model, which uses $T_j(y) = (y - \frac{1}{2})^j$ for $j = 1, 2, 3$. Again, set up algorithms so that you may compute the means, variances, covariances for T_1, T_2, T_3 , for any $(\theta_1, \theta_2, \theta_3)$.

Yet other models: log-normal, Weibull, Gompertz, Gumbel, et al.

Ex. 1.60 *The log-normal distribution.* Starting with $X \sim N(\xi, \sigma^2)$, the variable $Y = \exp(X)$ is said to be a *log-normal*, and we write $Y \sim \log N(\xi, \sigma^2)$ to indicate this.

the log-normal
distribution

(a) Consider the view that the distribution should or could have been named the exponential instead – would you agree? Show that with $Y \sim \log N(\xi, \sigma^2)$, its mean and variance are $\exp(\xi + \frac{1}{2}\sigma^2)$ and $\{\exp(2\sigma^2) - \exp(\sigma^2)\} \exp(2\xi)$.

(b) Show that its density may be written $\phi_\sigma(\log y - \xi)/y = \sigma^{-1} \phi(\sigma^{-1}(\log y - \xi))/y$, for $y > 0$. Find its mode.

(c) Assume that $Y | \xi \sim \log N(\xi, \sigma^2)$, and that $\xi \sim N(\xi_0, \tau^2)$. Show that marginally, $Y \sim \log N(\xi_0, \sigma^2 + \tau^2)$.

(d) Show that a product of independent log-normals is log-normal. Suppose $Y_1, \dots, Y_n$ are i.i.d. from the $\log N(\xi, \sigma^2)$ distribution. Explain what happens to their harmonic mean, $Z_n = (Y_1 \cdots Y_n)^{1/n}$.

(e) Assume a random time variable T has the $\log N(0, 1)$ distribution. Find a formula for its hazard rate $h(t)$, and show that $h(t) \doteq (\log t)/t$ for growing t . Plot the exact hazard rate, along with its approximation, and comment.

Ernst Hjalmar
Waloddi
Weibull

Ex. 1.61 *The Weibull distribution.* The Weibull distribution, with positive parameters (a, b) , has c.d.f. $F(t) = 1 - \exp\{-(t/a)^b\}$ for $t \geq 0$. The b is called the shape parameter, with a a scale parameter. The Weibull generalises the exponential distribution, which is the special case of $b = 1$. Other parametrisations are sometimes convenient, as with $1 - \exp(-ct^b)$.

(a) Find a formula for the median, and more generally for the q -quantile $F^{-1}(q)$. Show that the density can be written $f(t) = \exp\{-(t/a)^b\}bt^{b-1}/a^b$ for $t > 0$. Find also a formula for the hazard rate, and draw this in a diagram, for $b = 0.9, 1.0, 1.1$, say for $a = 1$.

(b) Use the general mean formula of (1.3) to work through the details of

$$ET^p = \int_0^\infty \Pr(T^p \geq u) du = \int_0^\infty \exp\{-(u^{1/p}/a)^b\} du = a^p \Gamma(1 + p/b).$$

Show that this leads to mean $a\Gamma(1 + 1/b)$ and variance $a^2\{\Gamma(1 + 2/b) - \Gamma(1 + 1/b)\}^2$. With T from the Weibull (a, b) , plot the function $\text{sd}(T)/ET$ as a function of b .

(c) Show that $V = (T/a)^b \sim \text{Expo}(1)$, and use this to give a recipe for simulating outcomes from any Weibull.

Ex. 1.62 *The Gompertz distribution.* The *Gompertz distribution*, with positive parameters (a, b) , has hazard rate $h(t) = a \exp(bt)$.

(a) Find the cumulative hazard rate, the c.d.f., and the density. Find also a formula for the median, and more generally for the q quantile, expressed via (a, b) . You may check with Story #51 where this is used.

(b) Suppose an individual has survived up to time t_0 . Show that her cumulative hazard rate, for the remaining lifetime, is $H(t) - H(t_0) = (a/b)\{\exp(bt) - \exp(bt_0)\}$. Give a formula for $t^*(t_0)$, her median survival time.

Ex. 1.63 *Cumulants and the cumulant-generating function.* For a variable X with m.g.f. $M(t)$, assumed finite in an interval around zero, it is sometimes fruitful to work with the *cumulant-generating function* $K(t) = \log M(t)$. When expanded in a power series around zero, with

cumulants

$$K(t) = K'(0)t + \frac{1}{2}K''(0)t^2 + \cdots = \sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j,$$

we call the coefficients $\kappa_j = K^{(j)}(0)$ the *cumulants* of the distribution.

(a) For the $N(\xi, \sigma^2)$ distribution, show that $K(t) = \xi t + \frac{1}{2}\sigma^2 t^2$. For the unit exponential distribution, show $K(t) = t + t^2/2 + t^3/3 + \cdots$, with $\kappa_j = (j-1)!$ for $j \geq 1$.

(b) If X has mean ξ , write $X = \xi + X_0$. Show that $K(t) = \xi t + K_0(t)$, where $K_0(t) = \log M_0(t)$, with M_0 the m.g.f. of the zero-mean variable X_0 . Hence the cumulants κ_j for X are the same as the cumulants $\kappa_{j,0}$ for X_0 , for $j \geq 2$.

(c) Via successive derivatives of $M(t) = \exp\{K(t)\}$, show that $M' = MK'$, $M'' = M\{K'' + (K')^2\}$, $M''' = M\{K''' + 3K''K' + (K')^3\}$, and one more. Write ξ and σ^2 for the mean and variance of X . From the equations, show first that $K'(0) = \kappa_1 = \xi$ and that $K''(0) = \kappa_2 = \sigma^2$, and next that

$$EX^3 = \kappa_3 + 3\xi\kappa_2 + \xi^3, \quad EX^4 = \kappa_4 + 4\xi\kappa_3 + 6\xi^2\kappa_2 + 3\kappa_2^2 + \xi^4.$$

(d) Skewness and kurtosis were defined in Ex. 1.3. We may now find useful expressions for these in terms of the cumulants. Show first via formulae above that

$$\mathbb{E}(X - \xi)^3 = \kappa_3, \quad \mathbb{E}(X - \xi)^4 = \kappa_4 + 3\sigma^4.$$

Explain that this leads to $\text{skew}(X) = \kappa_3/\sigma^3$, $\text{kurt}(X) = \kappa_4/\sigma^4$.

Ex. 1.64 *The Gumbel distribution.* Here we work with the Gumbel distribution, useful e.g. in models for extreme values.

the Gumbel
distribution

(a) Let $X_1, \dots, X_n$ be i.i.d. from the standard exponential distribution. Show that their maximum value M_n has c.d.f. $\{1 - \exp(-m)\}^n$. Deduce that $M_n - \log n$ has c.d.f. $G_n(u) = \{1 - (1/n)\exp(-u)\}^n$, for all $u \geq -\log n$.

(b) Show that the limit c.d.f. for $M_n - \log n$ becomes $G(u) = \exp\{-\exp(-u)\}$, and that this defines a c.d.f. on the full line. This is called the *Gumbel distribution*. Find its median and interquartile range.

you may read
*Vier Jahre
Lüge*, 1919;
*Vier Jahre
politischer
Mord*, 1922;
*Statistics of
Extremes*, 1958

(c) Find its density $g(u) = \exp\{-u - \exp(-u)\}$, and draw it in a diagram, along with the densities g_n for say $n = 10, 20, 30$, for $M_n - \log n$.

(d) With U having the Gumbel distribution, show that its m.g.f. becomes $M(t) = \Gamma(1 - t)$, for $t < 1$. To find the moments, show by taking derivatives of $\Gamma(a) = \int_0^\infty v^{a-1} \exp(-v) dv$ at position $a = 1$ that

$$\mathbb{E} U^p = (-1)^p \Gamma^{(p)}(1) = (-1)^p \int_0^\infty (\log v)^p \exp(-v) dv.$$

These may be found numerically. You may also work with the cumulant-generating function $K = \log M$, where a certain connection can be made to the Riemann zeta function, namely

$$K(t) = \log \Gamma(1 - t) = \gamma_e t + \sum_{j=2}^\infty \frac{\zeta(j)}{j} t^j$$

for $|t| < 1$. Here $\gamma_e = 0.5772\dots$ is the Euler constant, see Ex. 1.15, and $\zeta(j) = \sum_{n=1}^\infty 1/n^j$. In the terminology of Ex. 1.63, show that $\kappa_1 = \gamma_e$, $\kappa_2 = \zeta(2)$, $\kappa_3 = 2\zeta(3)$, $\kappa_4 = 6\zeta(4)$. With $\zeta(2) = \pi^2/6$, $\zeta(3) = \pi^3/25.7943$, $\zeta(4) = \pi^4/90$, establish that $\mathbb{E} U = \gamma_e$, $\text{Var } U = \sigma^2 = \pi^2/6$, $\text{skew}(U) = 2\zeta(3)/\sigma^3 \doteq 1.1395$, $\text{kurt}(U) = 12/5 = 2.4$.

(e) To appreciate the perhaps strange-looking connection from Gumbel moments to the zeta function, show from efforts of Ex. 1.15 that

$$M_n - \log n = \sum_{i=1}^n W_i/i + a_n$$

where $W_i = V_i - 1$, for i.i.d. unit exponentials V_i , and $a_n \rightarrow \gamma_e$. Find formulae for the mean, variance, skewness, kurtosis of $M_n - \log n$ based on this, and take their limits; these will agree with formulae found above. In particular, you have now found a probability proof of the famous $\sum_{n=1}^\infty 1/n^2 = \pi^2/6$.

the Basel
problem, solved
by Euler 1734,
with instant
world fame

Ex. 1.65 *The logistic distribution.* Consider the logistic distribution (in its standard form), with c.d.f. $H(x) = \exp(x)/\{1 + \exp(x)\}$, over the real line.

the logistic
distribution

(a) Show that H indeed is a proper c.d.f., and that its density is $h(x) = \exp(x)/\{1 + \exp(x)\}^2 = H(x)\{1 - H(x)\}$, symmetric around zero. Find its interquartile range.

(b) With X having the logistic distribution, show that its m.g.f. $M(t)$ becomes

$$\int \frac{\exp\{(t+1)x\}}{\{1 + \exp(x)\}^2} dx = \int_0^\infty \frac{u^t}{(1+u)^2} du = \int_0^1 v^t (1-v)^{1-t} dt = \Gamma(1+t)\Gamma(1-t),$$

for $|t| < 1$. Show from this that actually $X = U - V$, in terms of independent Gumbel distributed U and V , see Ex. 1.64. Demonstrate this also directly, via convolution. Use this representation to show that $\text{Var } X = \pi^2/3$ and that its kurtosis is $\text{kurt} = 6/5 = 1.2$.

(c) Here we managed to find moments of the logistic via the Gumbel difference representation. We may also use the intriguing formula $\Gamma(1+t)\Gamma(1-t) = \pi t / \sin(\pi t)$, a consequence of Euler's reflection formula. Take derivatives of this function to find the variance and kurtosis formulae.

(d) A scaled version of the logistic has c.d.f. $H(x, \tau) = H(x/\tau) = \exp(x/\tau) / \{1 + \exp(x/\tau)\}^2$, for a suitable τ . Show that the variance is $\tau^2 \pi^2/3$, which is 1 for $\tau = \sqrt{3}/\pi$. Draw a figure with the density, alongside the standard normal, and comment.

Ex. 1.66 *Alice and Bob correlate their binomials.* Here we show how correlated binomials may be constructed, leading also to correlated random walks.

(a) Alice flips her fair coin n times, with i.i.d. 0-1 outcomes $A_1, \dots, A_n$. Bob has two coins, and mixes between them depending on Alice's outcomes: if $A_i = 1$, he uses the plus-coin with probability $\frac{1}{2} + a$ for heads; and if $A_i = 0$ he uses his minus-coin with $\frac{1}{2} - a$ for heads. With B_i his outcome, show that $\Pr(B_i = 1) = \frac{1}{2}$, and argue therefore that both X_n and Y_n are binomial $(n, \frac{1}{2})$ variables, where $X_n = \sum_{i=1}^n A_i$ and $Y_n = \sum_{i=1}^n B_i$ are the number of heads for Alice and for Bob. Show that the correlation between these two binomials is $2a$.

(b) In the little story-telling above, Bob observes Alice's outcomes, one by one, which then influence his choice between two coins; Alice doesn't even need to be aware of Bob's existence. Explain however that we from observed pairs of coin flips (A_i, B_i) never can see the difference between that scenario and the alternative one, that Bob is the one flipping his fair coin, without caring for Alice, before she chooses between two biased ones. This is arguably an instance of what Breiman (2001) alludes to as the Rashomon Effect (from a Japanese movie in which different persons report very differently about something they have all observed): data alone cannot help us uncover which of the chains of action have been at work. See Story #9 for further discussion. Show indeed that as long as Alice and Bob have a joint scheme of producing outcomes $(0, 0), (0, 1), (1, 0), (1, 1)$, with probabilities respectively $\frac{1}{4}(1+a), \frac{1}{4}(1-a), \frac{1}{4}(1-a), \frac{1}{4}(1+a)$, then (X_n, Y_n) have the correlated binomial distribution.

(c) Find a way to compute $f(x, y) = \Pr(X_n = x, Y_n = y)$, for $x, y = 0, 1, \dots, n$.

(d) Leaving the Rashomon aspects to the side, generalise the first setup to the case of two correlated binomials (n, p) , where p is not necessarily $\frac{1}{2}$. Take indeed $\Pr(A_i = 1) = p$ and then $\Pr(B_i = 1 | A_i = 1) = p + a$, $\Pr(B_i = 0 | A_i = 0) = 1 - p + ap/(1 - p)$, for $a < \min(p, 1 - p)$, and show that this works properly. What is the correlation between X_n and Y_n ?

Ex. 1.67 *The hypergeometric distribution.* You draw a sample of n items from a bag of N , which has A of Type One and $B = N - A$ of type Two. Consider X , the number among the sampled n which are of Type One.

the Rashomon
Effect: different
models may
offer equally
good
explanations

(a) Show that X has distribution

$$f(x) = \Pr(X = x) = \binom{A}{x} \binom{B}{n-x} / \binom{N}{n}.$$

For which x is this positive? Explain the identity $\sum_{x=0}^A \binom{A}{x} \binom{B}{n-x} = \binom{A+B}{n}$.

(b) Show that $EX = nA/N = np$, with $p = A/N$ the proportion of Type One in the bag. This may be done working with $\sum_{x=0}^n xf(x)$, or by writing $X = J_1 + \dots + J_n$, with J_i and indicator for selected item i being a Type One or not.

(c) Explain that if one samples one item at the time, followed by replacing the item, then the $J_1, \dots, J_n$ above are independent Bernoulli variables with probability $p = A/N$, leading in that case to binomial variance $np(1-p)$. For the present hypergeometric setting, where n items sampled in one go without replacement, the J_i are dependent; show that $\text{cov}(J_1, J_2) = p(A-1)/(N-1) - p^2 = -p(1-p)/(N-1)$. Deduce that the variance formula becomes $\text{Var } X = c_n np(1-p)$, with c_n being the shrinking factor $(N-n)/(N-1)$. This may be accomplished working algebraically with $EX(X-1)$, or via the representation above.

(d) Explain that with n small compared to N , the X is close to a binomial. Attempt to put up a precise statement regarding this.

Ex. 1.68 *Zero degrees of freedom: the purely noncentral excentric chi-squared.* The noncentral chi-squared, with its somewhat complicated density found in Ex. 1.52, admits a simple representation, as a pure chi-squared plus a pure noncentral part.

(a) Consider a variable K from the noncentral chi-squared distribution, say $K \sim \chi_m^2(\lambda)$. We found its m.g.f. in the exercise pointed to. Deduce that if there is a representation, of the form $K = K_0 + U$, with $K_0 \sim \chi_m^2$ and a certain purely eccentric variable U , then we must have $M(t) = E \exp(tU) = \exp\{\lambda t/(1-2t)\}$. In particular, such a U will then have a distribution depending only on λ , the same for each m . so $\chi_m^2(\lambda) = \chi_m^2 + \chi_0^2(\lambda)$

(b) Let us use the occasion to demonstrate the moment-generating aspect of moment-generating functions. Still under the assumption that there actually is a purely eccentric U , as above, find the four first moments of U , using $EU^j = M^{(j)}(0)$, the derivatives of M at zero. Show indeed that $EU = \lambda$, $\text{Var } U = 4\lambda$, and that its skewness is $3/\lambda^{1/2}$.

(c) Indeed, there is such a U . You may check Hjort (1988a) to find constructive arguments to deduce the distribution of U , as opposed to taking the answer and then verifying that it is correct. Here we are content with the verification: show that with U having the distribution

$$\Pr(U \leq u) = \sum_{j=1}^{\infty} p(j) \Gamma_{2j}(u), \quad \text{with} \quad p(j) = \exp(-\tfrac{1}{2}\lambda) (\tfrac{1}{2}\lambda)^j / j!,$$

where $\Gamma_{2j}(u) = \Pr(\chi_{2j}^2 \leq u)$ is the cumulative for the χ_{2j}^2 , its m.g.f. indeed becomes the desired $\exp\{\lambda t/(1-2t)\}$. We allow also Γ_0 here, for the chi-squared with zero degrees of freedom, for the variable which is equal to zero with probability one. This is a Poisson mixture of χ_{2j}^2 distributions. In particular, U has a positive pointmass at zero, $\Pr(U = 0) = \exp(-\tfrac{1}{2}\lambda)$.

(d) Use the representation $U | (J = j) \sim \chi_{2j}^2$, with $J \sim \text{Pois}(\tfrac{1}{2}\lambda)$, to find the mean, the variance, the skewness, the kurtosis of U . Use these formulae again to find the mean, variance, skewness, kurtosis for $K = K_0 + U$, the noncentral $\chi_m^2(\lambda)$.

(e) What is required, of (m, λ) , for $K \sim \chi_m^2(\lambda)$ to have a normal limit distribution (when properly normalised)?

Notes and pointers

(xx to come. a bit old literature, but crisply, and not systematic. brief genesis of the normal, a few sentences on the chi-squared, [Pearson \(1900\)](#), the t, [Student \(1908\)](#), the F, the Dirichlet, more. we also point to essential things in later chapters. point out that the normal is famous and useful also because of a host of approximation methods. xx)

(xx mention things for [Irwin \(1927\)](#) and [Hall \(1927\)](#) and before. xx)

(xx application of zero degrees model, Ex. 1.68: point to [Groß and Möller \(2024\)](#) for ensemble postprocessing in meteorology, with days with no rain match the model's pointmass at zero. xx)

(xx where do we have a precise theorem on $M_X = M_Y$ implying $F = G$? inversion formula? Jensen is Johan Jensen. Vector form too?)

(xx find the little fun fact from [Hjort \(1994a\)](#), regarding t and midway between cauchy and normal. xx)

The chi-squared is on the list over deservedly famous distributions in probability theory and statistics, and stems from Karl Pearson's famous 1900 paper, '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'. [xx a bit more: he establishes the chi-square distribution, the test carrying the chi-squared name, and sets up a rigorous conceptual framework for hypothesis testing. xx]

(xx point to uses of expofamily in both Chs. 4 and 5. xx)

I.2

Large-sample theory

The broad themes of this chapter are the concepts, details, methods, results, applications pertaining to three modes of convergence for random variables: convergence in probability, convergence almost surely, convergence in distribution. The first two have to do with random variables X_n coming close to some limit X , with increasing n , typically indexed by sample size; often the limit is merely a constant. The chief result here, with various extensions and uses, is the Law of Large Numbers, that the empirical average of a sequence of observations tends to the expected value of the underlying distribution. The third mode of convergence rather involves the distribution of X_n coming close to the distribution of some limit X , with the Central Limit Theorem being a prime statement. These machineries also lead to practically useful approximations; the idea is that a complicated distribution may be approximated by something much simpler. This is in particular helped by a collection of approximation methods called the delta method. The theory is developed first for functions of i.i.d. sequences, involving tools of moment-generating functions and characteristic functions, along with various probability inequalities. It is then extended to cover cases of independent variables from non-equal distributions, culminating in the Lindeberg theorem, giving precise conditions under which a sum of independent components approaches normality. Methods and results from this chapter are crucial for developing the likelihood theory of Ch. 5, and also for several later chapters.

Key words: central limit theorems, characteristic functions, delta method, large-sample approximations, laws of large numbers, Lindeberg conditions, modes of convergence

In this chapter we study various modes of convergence of sequences $(X_n)_{n \geq 1}$ of random variables and random vectors, with several results also applicable to more general random elements, such as the stochastic processes studied in Ch. 10. The index n typically refers to the sample size, with X_n being some function of n data points, for example an estimator. The modes of convergence we study take place as n grows without bounds; hence the name large-sample theory.

Applications of large-sample theory are plentiful in probability and statistics, partly to understand crucial phenomena better, and partly to provide fruitful and practical approximations; an estimator or a statistic might have a very complicated exact distribution, but have a simple to use and sometimes accurate large-sample approximation. The key convergence concepts, with ensuing approximations and applications, are as follows.

First, if $(X_n)_{n \geq 1}$ is a sequence of random vectors, and X is some random vector, we say that X_n *converges to X in probability*, written $X_n \rightarrow_{\text{pr}} X$, if for each positive ε ,

$$\Pr(\|X_n - X\| \geq \varepsilon) \rightarrow 0, \quad \text{as } n \text{ tends to infinity.} \quad (2.1)$$

Here $\|x - y\| = (\sum_{j=1}^k |x_j - y_j|^2)^{1/2}$ is the Euclidean distance, and hence ordinary distance in the one-dimensional case. We will most often encounter situations where the limit X is not a random variable, but simply a constant, for example a fixed parameter value. If $\hat{\theta}_n$ is an estimator of θ , we say that the estimator is *consistent* if $\hat{\theta}_n \rightarrow_{\text{pr}} \theta$. Various results from the present chapter will indeed have immediate applications in Ch. 3.

consistency of
an estimator

Second, a stronger demand is that $(X_n)_{n \geq 1}$ converges to X *almost surely*, that is, the convergence occurs with probability one. This means that

$$\Pr(\lim_n \|X_n - X\| = 0) = 1, \quad (2.2)$$

and we write $X_n \rightarrow_{\text{a.s.}} X$. Again, the most commonly encountered state of affairs is that the limit is a constant. We say that an estimator $\hat{\theta}_n$ is *strongly consistent* for θ if $\hat{\theta}_n \rightarrow_{\text{a.s.}} \theta$. A strong achievement indeed is the *strong Law of Large Numbers* (LLN), which says that

the LLN

$$\text{if } X_1, X_2, \dots \text{ are i.i.d. with finite mean } \xi, \text{ then } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow_{\text{a.s.}} \xi, \quad (2.3)$$

with no further assumptions required. One may readily prove weaker versions of the LLN, as in Ex. 2.10(a) and Ex. 2.15, but with the development of sharper tools, and separate valuable results along the way, we reach the strong LLN in Ex. 2.57–2.58. Note that convergence in probability and almost surely can easily be defined for more general random elements taking values in some metric space, by, in (2.1) and (2.2), substituting the Euclidean norm $\|\cdot\|$ with the relevant metric d .

The third and statistically speaking most important concept is that of *convergence in distribution*. If the random vectors X_n and X of dimension $k \geq 1$ have c.d.f.s F_n and F , we say that X_n converges in distribution to X if

$X_n \rightarrow_d X$
 $F_n \Rightarrow F$

$$F_n(x) \rightarrow F(x), \quad \text{for all continuity points } x = (x_1, \dots, x_k) \text{ of } F. \quad (2.4)$$

We write $X_n \rightarrow_d X$ or $F_n \Rightarrow F$ to indicate convergence in distribution, allowing for simplicity also statements like $X_n \rightarrow_d N(0, 1)$ and $X_n \rightarrow_d F$, and the like. This form of convergence is also called *weak convergence* and *convergence in law*. The bigger sibling of the LLN is the Central Limit Theorem (CLT): if $X_1, X_2, \dots$ are i.i.d. random variables with variance σ^2 , then

the CLT

$$\sqrt{n}(\bar{X}_n - E X_1) \rightarrow_d N(0, \sigma^2), \quad (2.5)$$

again with no further assumptions needed beyond a finite variance. Below we go considerably further, however, in detail, in extensions, in applications. In particular, with additional tools and efforts we reach the Lindeberg theorem, with precise necessary conditions for a sum of independent variables from different distributions to approach normality. Such results are, for example, used to establish approximate normality for estimators in regression models.

Reaching the CLTs, with variations, requires hard mathematical work, with various technical details to sort out. When the main theorems have been established, however,

along with further tools, their actual use for statistical applications might be relatively straightforward. In particular, functions of approximately normal variables are also approximately normal, as we learn in the subsection on delta methods.

Modes of convergence

Ex. 2.1 Implications. Of the three modes of convergence studied in this chapter, convergence almost surely is the strongest, and convergence in distribution the weakest, with convergence in probability lying in the middle. Let $(X_n)_{n \geq 1}$ and X be random vectors.

(a) Show that $X_n \rightarrow_{\text{a.s.}} X$ if and only if the event $\cap_{n \geq 1} \cup_{k \geq n} \{\|X_k - X\| \geq \varepsilon\}$ has probability zero for all $\varepsilon > 0$.

(b) Show that if $X_n \rightarrow_{\text{a.s.}} X$, then $X_n \rightarrow_{\text{pr}} X$.

(c) Show that if $X_n \rightarrow_{\text{pr}} X$ then $X_n \rightarrow_d X$.

(d) Let Y be a nonnegative random variable. Show that

$$\Pr(Y \geq a) \leq \frac{\mathbb{E} Y}{a}, \quad \text{for any } a > 0.$$

This is Markov's inequality. Deduce that if $\mathbb{E} \|X_n - X\| \rightarrow 0$, then $X_n \rightarrow_{\text{pr}} X$.

(e) Show that $\|X_n - X\| \rightarrow_{\text{a.s.}} 0$ is equivalent to $|X_{n,j} - X_j| \rightarrow_{\text{a.s.}} 0$ for each j , and that the same is true for convergence in probability.

Ex. 2.2 The Borel-Cantelli lemma. Let $A_1, A_2, \dots$ be a sequence of events. Consider $\cap_{n \geq 1} \cup_{m \geq n} A_m$, the full-sequence event corresponding to the A_n occurring infinitely often.

(a) Let N be the total number of occurrences of the A_i . Show that $\mathbb{E} N = \sum_{i=1}^{\infty} \Pr(A_i)$.

(b) Show that if $\sum_{n=1}^{\infty} \Pr(A_n)$ is convergent, then $\Pr(\cap_{n \geq 1} \cup_{m \geq n} A_m) = 0$. So sooner or later there will be a finite (but random) n , such that none of the A_m will ever occur, for $m > n$.

(c) Assume in addition that the $A_1, A_2, \dots$ are independent. Show that if $\sum_{n=1}^{\infty} \Pr(A_n)$ is divergent, then $\Pr(\cap_{n \geq 1} \cup_{m \geq n} A_m) = 1$. In particular, for the case of independent events, there can't be say a 50 percent chance that there will be infinitely many occurrences. To show this, you may prove and use that $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$.

(d) Let $X_1, X_2, \dots$ be i.i.d. from the unit exponential distribution. Will there be infinitely many cases with $X_i \geq 0.99 \log i$, with $X_i \geq \log i$, with $X_i \geq 1.01 \log i$?

(e) Let $X_1, X_2, \dots$ be i.i.d. standard normal. Establish the Mills ratio $\Pr(X_i \geq a) = a^{-1} \phi(a) \{1 + o(1)\}$, as $a \rightarrow \infty$, and use this to show that there will be infinitely many cases with $|X_i| \geq (2 \log i)^{1/2}$, but only finitely many cases with $|X_i| \geq (2.001 \log i)^{1/2}$.

Ex. 2.3 The converses do not hold. The implications arrows in Ex. 2.1(b)&(c) only point in one direction, as we will now see.

(a) Let $X_1, X_2, \dots$ be Bernoulli random variables with success probabilities $p_1, p_2, \dots$. Show that $X_n \rightarrow_{\text{pr}} 0$ as long as $p_n \rightarrow 0$, but that $X_n \rightarrow_{\text{a.s.}} 0$ takes more. Will there be infinitely many $X_n = 1$, for $p_n = 1/n^{0.99}$, for $p_n = 1/n$, for $p_n = 1/n^{1.01}$? Conclude that we can have convergence in probability, but not almost surely.

(b) Let U be unif $(0, 1)$, and divide the unit interval in 2 pieces, then 3 pieces, and so on: $A_1 = [0, 1/2]$, $A_2 = (1/2, 1]$, $A_3 = [0, 1/4]$, $A_4 = (1/4, 1/2]$, $A_5 = (1/2, 3/4]$, etc. Define $X_n = I(U \in A_n)$ for $n = 1, 2, \dots$, and show that $X_n \rightarrow_{\text{pr}} 0$, but not almost surely.

(c) Find an example where $X_n \rightarrow_d X$, but there is not convergence in probability.

Markov's
inequality

The
Borel-Cantelli
lemma

(d) Let X_n be a sequence of binary random variable with distributions $\Pr(X_n = a_n) = p_n$ and $\Pr(X_n = 0) = 1 - p_n$. Construct the a_n and p_n sequences such that $X_n \rightarrow_{\text{pr}} 0$, but $E|X_n|$ does not converge to zero.

(e) Similarly, let $X_n = \theta + Z_n$ for $Z_1, Z_2, \dots$ independent with distributions $\Pr(Z_n = 2^n) = \Pr(Z_n = -2^n) = 1/(3n)$ and $\Pr(Z_n = 0) = 1 - 2/(3n)$. Show that $X_n \rightarrow_{\text{pr}} \theta$, but that $E|X_n - \theta|$ does not converge to zero.

Ex. 2.4 Partial converses. The implication arrows of Ex. 2.1 only point in one direction. We do, however, have certain partial converses.

(a) Show that if $X_n \rightarrow_{\text{pr}} X$, then there is a subsequence $(n_k)_{k \geq 1}$ such that $X_{n_k} \rightarrow_{\text{a.s.}} X$. To construct such a subsequence, you may use the Borel–Cantelli lemma of Ex. 2.2. This gives a partial converse to Ex. 2.1(b).

(b) Show that $X_n \rightarrow_{\text{pr}} X$ if and only if any subsequence $(X_{n_k})_k$ has a further subsequence $(X_{n_{k_j}})_j$ such that $X_{n_{k_j}} \rightarrow_{\text{a.s.}} X$.

(c) Show that if $X_n \rightarrow_d a$ for some constant $a \in \mathbb{R}$, then $X_n \rightarrow_{\text{pr}} a$; providing a partial converse to Ex. 2.1(c). Generalise to higher dimensions.

(d) As a partial converse to Ex. 2.1(d), suppose that $X_n \rightarrow_{\text{pr}} X$ and $\|X_n\| \leq Y$ such that $|Y|^p$ is integrable for some $p \geq 1$. Show that then $E\|X_n - X\|^p \rightarrow 0$. This is a version of the Dominated convergence theorem (DCT), see Ex. B.14, and you ought to consult that exercise before doing this one.

Dominated
Convergence
Theorem

(e) Another partial converse to Ex. 2.1(d) is provided by the following lemma. If X_n is a sequence of nonnegative random vectors such that $X_n \rightarrow_{\text{pr}} X$ and $EX_n \rightarrow EX$, with X integrable, then $E\|X_n - X\| \rightarrow 0$. Prove it first in one dimension, where you may use that $|y| = y + 2\max(-y, 0)$, then generalise.

Ex. 2.5 Uniform integrability. Yet another partial converse to Ex. 2.1(d) is obtained by introducing the concept of *uniform integrability*. A sequence $X_1, X_2, \dots$ of random variables is uniformly integrable if

$$\lim_{M \rightarrow \infty} \sup_{n \geq 1} E|X_n|I_{|X_n| > M} = 0. \quad (2.6)$$

As we show in this exercise, if $(X_n)_{n \geq 1}$ is uniformly integrable and $X_n \rightarrow_{\text{pr}} X$, then $E|X_n - X| \rightarrow 0$. Thus, compared to the DCT (see Ex. 2.4(d) and Ex. B.14) the uniform integrability assumption takes the place of the domination assumption of the giving the DCT its name. Moreover, as we will see in (c) and (i) below, uniform integrability is a bit weaker than boundedness, and turns out to be necessary for convergence in mean.

(a) Show that $E|X| < \infty$ is equivalent to $\lim_{M \rightarrow \infty} E|X|I_{|X| > M} = 0$. This explains why the property in (2.6) is called *uniform integrability*.

(b) Show that if $(X_n)_{n \geq 1}$ is uniformly integrable, then $\sup_n E|X_n| < \infty$. Exhibit a sequence of random variables for which the converse does not hold.

(c) Show that if $(X_n)_{n \geq 1}$ is dominated by an integrable Y , then it is uniformly integrable. Construct a sequence of random variables that is uniformly integrable, but that is not dominated by an integrable random variable.

(d) Let $a, b \in \mathbb{R}$ and suppose that $(X_n)_{n \geq 1}$ and $(Y_n)_{n \geq 1}$ are uniformly integrable. Show that $aX_n + bY_n$ is uniformly integrable.

(e) Let $(X_n)_{n \geq 1}$ be uniformly integrable such that $X_n \rightarrow_{\text{a.s.}} X$. Show that the limit X must be integrable, and that $E|X_n - X| \rightarrow 0$. To show the integrability part, consult

Fatou's lemma Ex. B.14(b). For the convergence in mean part, the trick is to truncate $Y_n = |X_n - X|$ and take it from there.

(f) Show that in (e), it is sufficient that X_n converges in probability to X .

(g) Show that if $(X_n)_{n \geq 1}$ is such that $\sup_n E|X_n|^{1+\delta} < \infty$ for some $\delta > 0$, then it is uniformly integrable.

(h) Show that $(X_n)_{n \geq 1}$ is uniformly integrable if and only if

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} E|X_n|I_{|X_n| > M} = 0,$$

and X_n is integrable for each n .

(i) The claim in (e) has a converse. Show that if $(X_n)_{n \geq 1}$ and X are nonnegative and integrable, $X_n \rightarrow_{\text{a.s.}} X$ and $E X_n \rightarrow E X$, then $(X_n)_{n \geq 1}$ is uniformly integrable.

(j) Deduce from (i) that if $(X_n)_{n \geq 1}$ and X are integrable (but not necessarily nonnegative), $X_n \rightarrow_{\text{a.s.}} X$ and $E|X_n - X| \rightarrow 0$, then $(X_n)_{n \geq 1}$ is uniformly integrable.

(k) Show that if X_n^2 and Y_n^2 are uniformly integrable, then $X_n Y_n$ is uniformly integrable. Show also that if X_n^p is uniformly integrable, then so is X_n^r for all $1 \leq r \leq p$.

Ex. 2.6 Scheffé's lemma. If we replace the random vectors in Ex. 2.4(e) with the densities $(f_n)_{n \geq 1}$ and f , then the limit is automatically integrable, and we obtain the lemma typically known by the name of Scheffé.

Scheffé's lemma

(a) Let Y_n and Y be random vectors with densities f_n and f with respect to a dominating measure μ . Show that if $f_n(y) \rightarrow f(y)$ for μ -almost all y , then $\int |f_n - f| d\mu \rightarrow 0$. Show also that this implies uniform convergence $\sup_B |\Pr(Y_n \in B) - \Pr(Y \in B)| \rightarrow 0$, the supremum being over all Borel sets B ; and, in particular, that $Y_n \rightarrow_d Y$.

(b) Suppose X_n has the Gamma(a_n, b_n) distribution, and that $(a_n, b_n) \rightarrow (1, 1)$. Show that X_n tends in distribution to a unit exponential. What is the general principle here?

(c) There are many situations where the Scheffé lemma does not apply. Consider for example the sequence $Y_n = 1/n$. Clearly, Y_n converges in distribution to zero, but show that $\sup_B |\Pr(Y_n \in B) - \Pr(Y \in B)| = 1$.

(d) If Y_n and Y have densities f_n and f , and $Y_n \rightarrow_d Y$, we should expect $f_n \rightarrow f$. This is not always happening, however. Consider the case of $F_n(y) = y + (n\pi)^{-1} \sin(n\pi y)$. Plot the F_n and its density f_n , for some n . Show that $Y_n \rightarrow_d \text{unif}(0, 1)$, but that $f_n(y)$ does not converge to 1 for all y . It is also instructive to transform to the approximate normal scale, via $X = \Phi^{-1}(Y)$. Show that X then has density $g_n(x) = \phi(x)\{1 + \cos(n\pi\Phi(x))\}$, with very notable oscillations, but where the c.d.f. $G_n(x)$ nevertheless tends to $\Phi(x)$.

Ex. 2.7 From discrete to continuous. Discrete distributions may have continuous limits.

(a) Let X_n have distribution $\Pr(X_n = j/n) = 1/(n+1)$ for $j = 0, 1, \dots, n$. Show that $X_n \rightarrow_d X$, where X has the uniform distribution on the unit interval. Show also that $\sup_B |\Pr(X_n \in B) - \Pr(X \in B)|$ does not converge to zero (cf. Ex. 2.6(c)).

(b) With perhaps similar techniques as in (a), consider X_n with distribution $\Pr(X_n = j/n) = j/\{n(n+1)/2\}$ for $j = 1, \dots, n$. Find its limit distribution.

(c) Let $X_1, X_2, \dots$ be independent Bernoulli random variables with success probability p . Only using (2.4), show that $\sqrt{n}(\bar{X}_n - p) \rightarrow_d N(0, p(1-p))$. You ought to use Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$ here; a formula we prove via probabilistic tricks in Story #103.

Ex. 2.8 Many small probabilities give a Poisson. The Law of Small Numbers says that if we sum a high number of 0-1 variables, with each having a small probability of 1, then

we're close to a Poisson. This was touched upon in Ex. 1.24, but here we give a clearer picture, with more to come in Ex. 2.29, with natural setups yielding Poisson limits.

der Gesetz der
kleinen Zahlen,
1898

(a) Suppose $Y_n \sim \text{binom}(n, p_n)$, with p_n becoming small with growing n in a way which has $np_n \rightarrow \lambda$. Show that $Y_n \rightarrow_d \text{Pois}(\lambda)$. As a small demonstration, take $n = 100$ and $p = 0.05$. Compare the exact and the limiting probabilities.

(b) Suppose (X_n, Y_n) has the trinomial distribution, with parameters (n, p_n, q_n) , see Ex. 1.4, where we also found the correlation between them. Assume now that p_n, q_n become small with n , such that $np_n \rightarrow \lambda_1, nq_n \rightarrow \lambda_2$. Show that the correlation between X_n and Y_n tends to zero, and that $(X_n, Y_n) \rightarrow_d (X, Y)$, where X and Y are independent and Poisson with parameters λ_1, λ_2 . Generalise to the multinomial situation.

(c) Many small Poissons become a Poisson: consider $Y_1, \dots, Y_m$, i.i.d. $\text{Pois}(a/m)$. Why is their sum $S_m = \sum_{i=1}^m Y_i$ a $\text{Pois}(a)$? In addition, the individual Poisson terms are essentially just zeros and ones, for growing m : find $q_m = \Pr(\max_{i \leq m} Y_i \leq 1)$ and show $q_m \rightarrow 1$. Generalise: suppose $Y_i \sim \text{Pois}(\theta_i)$, with $\sum_{i=1}^m \theta_i \rightarrow a$ and $\max_{i \leq m} \theta_i \rightarrow 0$. Show $q_m \rightarrow 1$.

Ex. 2.9 *Maximum of uniforms.* Let $X_1, \dots, X_n$ be i.i.d. random variables with the uniform distribution on $(0, \theta)$, and set $M_n = \max_{i \leq n} X_i$.

(a) Show that $M_n \rightarrow_{\text{pr}} \theta$, that is, the maximum of the observations is consistent for the unknown endpoint. In fact, try to show that $M_n \rightarrow \theta$ a.s.. This is an instance of a more general result, see Ex. 2.13.

(b) Find the limit distribution of $V_n = n(\theta - M_n)$, and use this result to find an approximate 90 percent confidence interval for θ , i.e., a random interval $[L_n, U_n]$ such that $\Pr(L_n \leq \theta \leq U_n)$ is approximately 0.90.

see Ch. 3 for
confidence
intervals

Convergence in probability and tail bounds

Ex. 2.10 *Markov, Chebyshev, and the Law of Large Numbers.* In view of the definition of convergence in probability in (2.1), it is certainly useful to find mathematically manageable bounds for so-called tail probabilities, i.e., $\Pr(|X| \geq a)$ for a random variable X . There are several such, as we learn here, with more to come in Ex. 2.16. Their uses include assessing how likely it might be that an estimator is some distance off its target.

(a) Markov's inequality was proven in Ex. 2.1: If X is nonnegative, then $\Pr(X \geq a) \leq E(X)/a$ for each $a > 0$. More generally, if $h(x)$ nonnegative and nondecreasing, with $h(a)$ positive, show that $\Pr(X \geq a) \leq E h(X)/h(a)$. From this, deduce the Chebyshev inequality: if X has finite variance (but is not necessarily nonnegative), then

the Chebyshev
inequality

$$\Pr(|X - EX| \geq \varepsilon) \leq \text{Var}(X)/\varepsilon^2, \quad \text{for } \varepsilon > 0.$$

(b) With $X_1, \dots, X_n$ i.i.d. with mean ξ and standard deviation σ , show that for the empirical mean $\bar{X}_n$ that $\Pr(|\bar{X}_n - \xi| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2)$. Since the right hand side tends to zero as n tends to infinity, you have now proven the version of the *Law of Large Numbers* (LLN) that says that the empirical mean of n i.i.d. random variables with finite second moments converges in probability to its expectation as n tends to infinity.

the LLN via
Chebyshev

(c) Let $Y_1, Y_2, \dots$ be i.i.d. random variables with expectation θ and finite variance σ^2 . Consider the weighted average $\hat{\theta}_n = \sum_{i=1}^n w_i Y_i / \sum_{i=1}^n w_i$, for nonnegative and fixed weights w_i . Give a condition, in terms of these weights, ensuring consistency $\hat{\theta}_n \rightarrow_{\text{pr}} \theta$. What happens for $w_i = 1/i$, and for $w_i = 1/i^{1.5}$?

(d) If X has mean ξ and $E X^4 < \infty$, show that $\Pr(|X - \xi| \geq \varepsilon) \leq E|X - \xi|^4/\varepsilon^4$. For $X \sim N(\xi, \sigma^2)$, deduce that $\Pr(|X - \xi| \geq \varepsilon) \leq 3\sigma^4/\varepsilon^4$.

(e) With $X_1, \dots, X_n$ i.i.d. from a distribution with finite fourth moment, write $\gamma_4 = E\{(X_i - \xi)/\sigma\}^4 - 3$ for its kurtosis. Show that

$$E|\bar{X}_n - \xi|^4 = \frac{\sigma^4}{n^2}\{3 + (1/n)(\gamma_4 - 3)\}.$$

Show hence that $\Pr(|\bar{X}_n - \xi| \geq \varepsilon) \leq 3.01 \sigma^4/(n^2 \varepsilon^2)$, for all large enough n . When is this a sharper result than that of the Chebyshev inequality?

the empirical
c.d.f.

Ex. 2.11 *The binomial and the empirical distribution function.* For i.i.d. observations $Y_1, \dots, Y_n$, we form the empirical cumulative distribution function, the empirical c.d.f., with $F_n(t) = n^{-1} \sum_{i=1}^n I(Y_i \leq t)$. Plotting F_n for a given dataset is informative, also for comparing with given distributions.

(a) Before returning to F_n , consider $X_n \sim \text{binom}(n, p)$ and the familiar ratio $\hat{p}_n = X_n/n$. Show that indeed $\hat{p}_n \rightarrow_{\text{pr}} p$. Historically speaking, this is the earliest clear LLN statement.

(b) Coming back to the empirical c.d.f., explain that $F_n(t)$ for given t is just a binomial proportion, and that $F_n(t) \rightarrow_{\text{pr}} F(t)$ for each t . In particular, the events $F_n(t) \leq F(t) - \varepsilon$ and $F_n(t) \geq F(t) + \varepsilon$ both have probabilities going to zero, for positive ε .

(c) We come back to finer analysis of F_n in Ex. 2.60 and 10.27, but for now find the variance of $F_n(t)$ and the covariance between $F_n(t_1)$ and $F_n(t_2)$.

Ex. 2.12 *Continuous mapping for convergence in probability.* The continuous mapping theorem says that the three modes of convergence introduced above are preserved under a continuous mapping. In this exercise we look at the convergence in probability part of this theorem, coming back to the convergence in distribution part in Ex. 2.20.

(a) Let X_n be a sequence of random vectors such that $X_n \rightarrow_{\text{pr}} a$, with a being a constant. Show that if g is a function continuous at point $x = a$, then $g(X_n) \rightarrow_{\text{pr}} g(a)$.

(b) Suppose that $X_n \rightarrow_{\text{pr}} X$, with the limit being a random vector such that $\Pr(X \in C) = 1$, with $C \subset \mathbb{R}^k$. Show that if g is uniformly continuous on C , then $g(X_n) \rightarrow_{\text{pr}} g(X)$.

(c) Show that in (b) it is sufficient that g is continuous on C . There are several ways of showing this, one involves the subsequences result from Ex. 2.4.

(d) Generalise the above to sequences X_n of more general random elements.

Ex. 2.13 *Convergence of sums of random variables.* A sequence $(x_n)_n$ of real numbers is absolutely summable if $\sum_n |x_n|$ is convergent.

(a) Show that if $(x_n)_n$ is absolutely summable then both $\sum_n x_n$ and $\sum_n x_n^2$ are convergent. Exhibit examples showing that the reverse implications do not hold.

(b) Let $0 \leq Y_1 \leq Y_2 \leq Y_3 \leq \dots$ be random variables such that $Y_n \rightarrow_{\text{pr}} Y$. Show that $Y_n \rightarrow Y$ almost surely.

(c) Let $X_1, X_2, \dots$ be random variable such that $\sum_n E|X_n|$ is convergent. Show that $\sum_n X_n$ is almost surely convergent.

(d) Suppose that $\sum_n \Pr(|X_n| \geq M)$ and $\sum_n E|X_n|I_{|X_n| \leq M}$ are convergent for some $M > 0$. Show that then $\sum_n X_n$ is almost surely convergent.

(e) Show that the summability condition in (c) implies the two conditions in (d).

(f) Consider the sequence $(X_n)_{n \geq 1}$ with distributions given by $\Pr(X_n = n) = 1/n^2$ and $\Pr(X_n = 0) = 1 - 1/n^2$ for $n \geq 1$. Use this to show that the two conditions in (d) are weaker than the summability condition in (c).

Ex. 2.14 Quantiles. Here we learn that the empirical quantiles are consistent for their population counterparts. Further informative analyses are in Ex. 3.21, including convergence in distribution, but here we limit attention to direct convergence in probability.

(a) Suppose first that $U_1, \dots, U_n$ are i.i.d. from the uniform distribution on the unit interval, and let M_n be the empirical median. With the empirical c.d.f. F_n from Ex. 2.11, use the fact that $F_n(t) \rightarrow_{\text{pr}} t$, for each t , to show that $M_n \rightarrow_{\text{pr}} \frac{1}{2}$. Then generalise to the case of observations $Y_1, \dots, Y_n$ from a distribution with continuous and strictly increasing c.d.f. F , showing for the empirical median that $M_n \rightarrow_{\text{pr}} F^{-1}(\frac{1}{2})$.

(b) More generally, for a $q \in (0, 1)$, let $Q_n = Q_n(q)$ be the empirical q -quantile, defined here to be $Y_{(\lfloor nq \rfloor)}$, the $\lfloor nq \rfloor$ order statistic, where $\lfloor x \rfloor = \max\{m \in \mathbb{Z} : m \leq x\}$. Show that Q_n is consistent for the population quantile $F^{-1}(q)$. floor function

(c) Show that the interquartile range, the 0.75 quantile minus the 0.25 quantile, is consistent for the population interquartile range $F^{-1}(0.75) - F^{-1}(0.25)$. Show more generally that if $\theta = g(F^{-1}(q_1), \dots, F^{-1}(q_r))$ is any continuous function of a finite number of quantiles, then $\hat{\theta} = g(Q_n(q_1), \dots, Q_n(q_r))$ is consistent for θ .

Ex. 2.15 Improving on the weak LLN. In this exercise we get rid of the finite variance condition used to prove the (weak) LLN in Ex. 2.10. For the almost sure version of the LLN see Ex. 2.57.

(a) Let $X_1, \dots, X_n$ be i.i.d. with finite mean ξ . Truncate the random variables involved by some constant $M > 0$, and show that for any $\varepsilon > 0$,

$$\Pr(|\bar{X}_n - \xi| \geq \varepsilon) \leq \frac{8M}{n\varepsilon^2} \mathbb{E}|X_1| + \frac{4}{\varepsilon} \mathbb{E}|X_1 - \xi|I_{|X_1 - \xi| > M},$$

or some other workable version thereof.

(b) Let $X_1, X_2, \dots$ be i.i.d. random variables with finite mean ξ , i.e., no assumption is made about its second moment. Show that $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i \rightarrow_p \xi$. law of large numbers

(c) Let $X_1, X_2, \dots$ be i.i.d. random variables with finite variance σ^2 . Show that the empirical variance $n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is consistent for σ^2 .

Ex. 2.16 Further tail bound inequalities. In Ex. 2.10 we learned about the Markov and Chebyshev inequalities; here we work out further tail bounds.

(a) Suppose X has m.g.f. $M(t) = \mathbb{E} \exp(tX)$, as per Ex. 1.32. Show that

$$\Pr(X \geq a) \leq q(a) = \min\{t \geq 0 : \exp(-ta)M(t)\}.$$

Writing $M(t) = \exp\{K(t)\}$, show that this leads to $q(a) = \exp\{K(t_a) - at_a\}$, where t_a is the solution to $K'(t_a) = a$, if positive. Apply this to $X \sim N(0, 1)$ and a positive, and show that $\Pr(X \geq a) \leq \exp(-\frac{1}{2}a^2)$. Show that this is indeed sharper than the tail bound $1/a^2$, from the simpler Chebyshev inequality, for all $a > 0$.

(b) For $X \sim N(0, 1)$ and a positive, show that

$$\Pr(|X| \geq a) \text{ is smaller than each of } \frac{1}{a^2}, \frac{3}{a^4}, \frac{15}{a^6}, 2\exp(-\frac{1}{2}a^2).$$

(c) Let $X \sim \chi_m^2$, which has mean and variance m and $2m$. Consider $p_m(a) = \Pr(X \geq$

$m + am^{1/2}$). Show that $p_m(a) \leq q_m(a)$, where

$$\begin{aligned} q_m(a) &= \min\{t \geq 0: \exp(-ta - t/\sqrt{m})(1 - 2t/\sqrt{m})^{-m/2} \\ &= \exp(-\tfrac{1}{2}a\sqrt{m})(1 + a/\sqrt{m})^{m/2}. \end{aligned}$$

Compare this bound both with bounds from the Markov inequality, and with the exact limit of $p_m(a)$, as m grows.

(d) Let $X_1, X_2, \dots$ be i.i.d. with mean zero and variance one, so that $\sqrt{n}\bar{X}_n \rightarrow_d N(0, 1)$, but you don't need to use that here. Assume the m.g.f. $M(t) = E \exp(tX_1) = \exp\{K(t)\}$ is finite. Show that

$$\Pr(\sqrt{n}\bar{X}_n \geq a) \leq \frac{M(t/\sqrt{n})^n}{\exp(ta)} = \exp\{nK(t/\sqrt{n}) - ta\}, \quad \text{for each } t.$$

Ex. 2.17 Bernshtein and Weierstraß. In c. 1885, Karl Weierstraß proved one of the fundamental and insightful results of approximation theory, that any given continuous function can be approximated uniformly well, on any finite interval, by polynomials. Here we give a constructive and relatively straightforward proof of the Weierstraß theorem, involving so-called Bernshtein polynomials. Let $g: [0, 1] \rightarrow \mathbb{R}$ be continuous, and construct

$$B_n(p) = E_p g(X_n/n) = \sum_{j=0}^n g(j/n) \binom{n}{j} p^j (1-p)^{n-j} \quad \text{for } p \in [0, 1],$$

where $X_n \sim \text{binom}(n, p)$. Note that $B_n(p)$ is a polynomial of degree n .

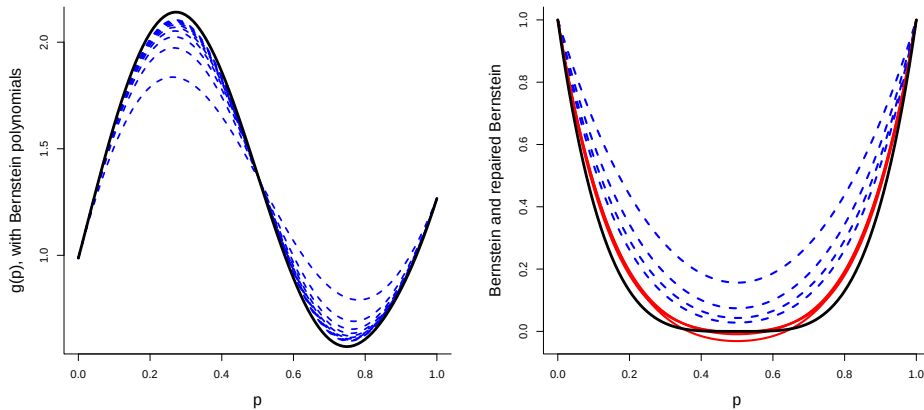


Figure 2.1: Left panel: the given non-polynomial function $g(p)$ (full black curve), along with approximating Bernshtein polynomials of order 10, 20, ..., 100. Right panel: for the function $g_2(p) = (2p - 1)^4$, the Bernshtein polynomials B_n (dashed curves), along with the bias-modified ones B_n^* (full curves), of order 4, 6, 8, 10.

(a) Show that $B_n(p) \rightarrow_{\text{pr}} g(p)$, for each p . Then show that the convergence is actually uniform. In some detail, for $\varepsilon > 0$, find $\delta > 0$ such that $|x - y| < \delta$ implies $|g(x) - g(y)| < \varepsilon$ (which is possible, as a continuous function on a compact interval is always uniformly continuous). Then show

$$|B_n(p) - g(p)| \leq \varepsilon + 2M \Pr(|X_n/n - p| \geq \delta),$$

with M a bound on $|g(p)|$. Show from this that indeed $\max_p |B_n(p) - g(p)| \rightarrow 0$.

(b) Consider the marvellous function

$$g_1(p) = \sin(2\pi p) + \exp(1.234 \sin^3 \sqrt{p}) - \exp(-4.321 \cos^5 p^2)$$

on the unit interval. Compute the Bernshtein polynomials of various orders, and display these in a diagram, alongside the curve of g . Construct a version of Figure 2.1, left panel, which does this for $n = 10, 20, \dots, 90, 100$. How high n is needed for the maximum absolute difference to creep below 0.01?

(c) Assuming smoothness of the g function, and writing $\hat{p} = X_n/n$ for the binomial fraction, use Taylor expansion $g(\hat{p}) = g(p) + g'(p)(\hat{p} - p) + \frac{1}{2}g''(p)(\hat{p} - p)^2 + O_{\text{pr}}(|\hat{p} - p|^3)$ to show that $B_n(p) = g(p) + \frac{1}{2}g''(p)p(1-p)/n + o(1/n)$. This invites a repaired Bernshtein polynomial, of the form $B_n^*(p) = B_n(p) - \frac{1}{2}C_n(p)p(1-p)/n$, with $C_n(p) = E_p g''(X_n/n)$. Explain that this is still a polynomial, now of order $n + 2$, and show in examples that it often succeeds better than B_n in coming close to the target $g(p)$. Make a version of Figure 2.1, right panel, which shows the B_n and the B_n^* , of order 4, 6, 8, 10, for the case of $g_2(p) = (2p - 1)^4$. Check here that $\max_p |B_n^*(p) - g(p)|$ becomes rather smaller than $\max_p |B_n(p) - g(p)|$. notation: as per Ex. 2.22

(d) Assuming g has a derivative, construct a confidence band of the type $B_n(p) \pm 1.96 \hat{\sigma}(p)/\sqrt{n}$, for a certain $\hat{\sigma}(p)$, with the property that it for each given p covers the underlying $g(p)$ with probability tending to 0.95.

(e) Let now $g(x, y)$ be an arbitrary function on the unit simplex, $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 1\}$. Construct a mixed polynomial $B_n(x, y)$ of degree n such that it converges uniformly to g on the simplex.

Convergence in distribution

Ex. 2.18 *The Portmanteau theorem.* Let $(P_n)_n$ and P be probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$; where $\mathcal{X}$ is metrisized by d . We say that P_n converges weakly to P , denoted $P_n \Rightarrow P$, if

$$\int g \, dP_n \rightarrow \int g \, dP, \quad (2.7)$$

weak
convergence
 $P_n \Rightarrow P$

for all bounded and continuous functions $g : \mathcal{X} \rightarrow \mathbb{R}$. When $(P_n)_n$ and P are the distributions of some random elements X_n and X , we might just write $Eg(X_n) \rightarrow Eg(X)$ for the convergence in (2.7). If X_n and X are random vectors, then (2.7) is equivalent to $X_n \rightarrow_d X$ as defined in (2.4). As will become clear as we proceed, however, the definition in (2.7) is vastly more general than (2.4), in particular, it works when $\mathcal{X}$ is a function space, see Ch. 10. It is also often much easier to work with. In this exercise we first prove the classical Portmanteau theorem, valid for all types of metric spaces, then proceed to proving its equivalence with (2.4). In Ex. 2.19 we look into additional equivalent statements. The theorem states that the following four statements are equivalent:

- (i) $P_n \Rightarrow P$;
- (ii) $\limsup_n P_n(F) \leq P(F)$ for every closed set F ;
- (iii) $\liminf_n P_n(B) \geq P(B)$ for every open set B ;
- (iv) $P_n(C) \rightarrow P(C)$ for every set C that is P -continuous, in the sense that $P(\partial C) = 0$, where $\partial C = \bar{C} \setminus C^\circ$ is the boundary of C (the closure minus the interior).

(a) For any set A define $d(x, A) = \inf\{d(x, y) : y \in A\}$, i.e., the distance from x to A , and define the function

$$\eta(t) = \begin{cases} 1, & \text{if } t \leq 0, \\ 1 - t, & \text{if } 0 \leq t \leq 1, \\ 0, & \text{if } t \geq 1. \end{cases}$$

For some $\varepsilon > 0$, the function $\eta(d(x, A)/\varepsilon)$ will be used to approximate the indicator function I_A . Let $F = [a, b]$ be a closed interval on the real line, and make a sketch of $\eta(d(x, F)/\varepsilon)$. Show that for any closed set F , that $x \mapsto \eta(d(x, A)/\varepsilon)$ is bounded and continuous, and that for $F^\varepsilon = \{x : d(x, F) < \varepsilon\}$

$$I_F(x) \leq \eta(d(x, A)/\varepsilon) \leq I_{F^\varepsilon}(x).$$

(b) Use the inequalities from (a) to show that (i) implies (ii) and (iii).

(c) Show that (ii) and (iii) imply (iv).

(d) To show that (iv) implies (i), take f to have its values inside $[0, 1]$, without loss of generality. Write

$$\mathbb{E} f(X_n) = \int \int_0^1 I\{y \leq f(x)\} dy dP_n(x) = \int_0^1 \Pr\{f(X_n) \geq y\} dy,$$

and finish up the argument. You may prove and use that for a continuous function f , we have the inclusions $\overline{f^{-1}(B)} \subset f^{-1}(\bar{B})$ and $f^{-1}(B^\circ) \subset (f^{-1}(B))^\circ$, which entails the inclusion $\partial f^{-1}(B) \subset f^{-1}(\partial B)$.

(e) Finally, consider the setting where we have c.d.f.s $F_n(x) = P_n(-\infty, x]$ and $F(x) = P(-\infty, x]$. Show that for $P_n \Rightarrow P$ it is necessary and sufficient that $F_n \Rightarrow F$. To show the sufficiency part, express the open set $B \subset \mathbb{R}$ as a union of the disjoint open intervals $(a_1, b_1), (a_2, b_2), \dots$, and deduce that $F_n \Rightarrow F$ implies (iii) of the Portmanteau theorem.

Ex. 2.19 More on the Portmanteau. It is often convenient to have smaller classes of functions to work with than the bounded and continuous ones. In this exercise we show that $P_n \Rightarrow P$ is equivalent to $\int f dP_n \rightarrow \int f dP$ for all bounded and continuous functions f with bounded and continuous derivatives of all orders.

(a) As a preliminary that will be useful below, for $k = 1, 2, \dots$ consider the functions $g_k(x, u) = |x - u|^k \exp\{-\frac{1}{2}(x - u)^2\}$. For a given $\varepsilon > 0$ define

$$K_k(u) = \begin{cases} |u + \varepsilon|^k \exp\{-\frac{1}{2}(u + \varepsilon)^2\}, & u < -k^{1/2} - \varepsilon, \\ k^{k/2} \exp(-\frac{1}{2}k), & -k^{1/2} - \varepsilon \leq u < k^{1/2} + \varepsilon, \\ |u - \varepsilon|^k \exp(-\frac{1}{2}(u - \varepsilon)^2), & k^{1/2} + \varepsilon \leq u, \end{cases}$$

and show that $g_k(x, u) \leq K_k(u)$ for all $x \in (-\varepsilon, \varepsilon)$ and that $\int K_k(u) du < \infty$.

(b) Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a bounded and continuous function. With X having any distribution, $Z \sim N(0, 1)$ independent of X , and $\sigma > 0$, show that $\mathbb{E} f(X + \sigma Z) = \mathbb{E} f_\sigma(X)$ where (see Ex. C.25)

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \int f(u) \exp\{-\frac{1}{2}(x - u)^2/\sigma^2\} du.$$

Show that $f_\sigma(x)$ is bounded and continuous, and has bounded and continuous derivatives of all orders.

- (c) Let $Z, Z_1, Z_2, \dots$ be a standard normal random variables with Z_i independent of X_i for each $i \geq 1$. Show that $X_n \rightarrow_d X$ if and only if $X_n + \sigma Z_n \rightarrow_d X + \sigma Z$ for each $\sigma > 0$.
- (d) From (b) and (c), conclude that $X_n \rightarrow_d X$ if and only if $E f(X_n) \rightarrow E f(X)$ for all bounded and continuous functions, having bounded and continuous derivatives of all orders. This version of the Portmanteau theorem will be of use in Ex. 2.32.

Ex. 2.20 *Continuous mapping for convergence in distribution.* Let $X_1, X_2, \dots$ and X be random elements with values in some metric space $\mathcal{X}$, and let g be a function that is continuous on a set $C \subset \mathcal{X}$ such that $\Pr(X \in C) = 1$. Suppose that $X_n \rightarrow_d X$.

continuous
mapping

- (a) Assume that g is continuous on all of $\mathcal{X}$, and show that $g(X_n) \rightarrow_d g(X)$.
- (b) Now suppose that C is a proper subset of $\mathcal{X}$, and let F be a closed set. The challenge here is that while the sets $\{g(X_n) \in F\}$ and $\{X_n \in g^{-1}(F)\}$ are identical, the set $g^{-1}(F)$ need not be closed, as g is only assumed continuous on C . Use the Portmanteau theorem to establish that

$$\limsup_n \Pr(X_n \in g^{-1}(F)) \leq \Pr(X \in \overline{g^{-1}(F)}).$$

It then remains to prove that $\Pr(X \in \overline{g^{-1}(F)}) \leq \Pr(X \in g^{-1}(F))$ as the reverse inequality is immediate from the inclusion of a set in its closure. To show this inequality, start by combining the continuity of g on C and the fact that a closed set contains all its limits points to show that $\overline{g^{-1}(F)} \cap C \subset g^{-1}(F)$, and take it from there.

Ex. 2.21 *The Slutsky–Cramér rules.* The usefulness of the three results in (c) below, together known as the Slutsky–Cramér rules, will become abundantly clear as we progress.

Evgeniĭ Slutskiĭ
1925, Harald
Cramér 1946

- (a) Show that if $X_n \rightarrow_d X$, and $Y_n - X_n \rightarrow_{\text{pr}} 0$, then also $Y_n \rightarrow_d X$. This says that variables which are essentially close, for growing n , have identical limit distributions.
- (b) Show that if X_n and Y_n are sequences of random vectors such that $X_n \rightarrow_d X$ and $Y_n \rightarrow_{\text{pr}} a$, for a random variable X and a constant a , then $(X_n, Y_n) \rightarrow_d (X, a)$.
- (c) Show that if $X_n \rightarrow_d X$ and $Y_n \rightarrow_{\text{pr}} a$, as above, then (i) $X_n + Y_n \rightarrow_d X + a$; (ii) $X_n Y_n \rightarrow_d X a$; (iii) $X_n / Y_n \rightarrow_d X / a$, provided $a \neq 0$. Explain why rules (i) and (ii) also hold when Y_n and a are matrices.

Slutsky–Cramér

- (d) Let $X_1, \dots, X_n$ be i.i.d. Bernoulli(p). Look back at Ex. 2.7 and Ex. 2.15 to show that $\sqrt{n}(\bar{X}_n - p)/\hat{\sigma}^2 \rightarrow_d N(0, 1)$, where $\hat{\sigma}^2 = \hat{p}(1 - \hat{p})$.

- (e) Suppose $X_n \rightarrow_d X$, in dimension p , and that $A_n \rightarrow_{\text{pr}} A$, a symmetric matrix, in the sense that $A_{n,i,j} \rightarrow_{\text{pr}} A_{i,j}$ for each component (i, j) . Show that $X_n^t A_n X_n \rightarrow_d X^t A X$.

- (f) Assume $X_n \rightarrow_d N_p(0, \Sigma)$, with a positive definite Σ , and that there is a consistent estimator $\hat{\Sigma}$ for this matrix. Show that $X_n^t \hat{\Sigma}^{-1} X_n \rightarrow_d \chi_p^2$.

- (g) The claim in (b) is in general not true if the constant limit in probability a is replaced with a (nondegenerate) random variable. Construct an example where $X_n \rightarrow_d X$ and $Y_n \rightarrow_{\text{pr}} Y$, for a random variable Y , but (X_n, Y_n) does not converge in distribution.

Ex. 2.22 *Stochastic O_{pr} and o_{pr} symbols.* The symbols $O_{\text{pr}}(\cdot)$ and $o_{\text{pr}}(\cdot)$ are the stochastic analogues of $O(\cdot)$ and $o(\cdot)$. Let X_n be a sequence of random variables. Then $X_n = o_{\text{pr}}(1)$ if $X_n \rightarrow_{\text{pr}} 0$; and $X_n = O_{\text{pr}}(1)$ if X_n is bounded in probability, i.e., for any $\varepsilon > 0$ there is an M such that $\Pr(|X_n| > M) < \varepsilon$ for all n . If r_n is another sequence of random variables, then $X_n = o_p(r_n)$ if $X_n = r_n o_{\text{pr}}(1)$, and $X_n = O_{\text{pr}}(r_n)$ if $X_n = r_n O_{\text{pr}}(1)$. Typically, r_n will be a sequence of constants, $r_n = n^{-1}$ and $r_n = n^{-1/2}$ for example, appear frequently in parametric statistics.

- (a) Show that $X_n = o_{\text{pr}}(r_n)$ if and only if $\Pr(|X_n| \geq \varepsilon r_n) \rightarrow 0$ for each $\varepsilon > 0$.

- (b) Show that $X_n = O_{\text{pr}}(r_n)$ if and only if for any $\varepsilon > 0$ there are M and n_0 such that $\Pr(|X_n| \geq Mr_n) < \varepsilon$ for all $n \geq n_0$.
- (c) Show that if $\sqrt{n}X_n \rightarrow_d N(0, \sigma^2)$, then $X_n = O_{\text{pr}}(1/\sqrt{n})$. More generally, show that if $n^\alpha X_n \rightarrow_d Z$, then $X_n = O_{\text{pr}}(n^{-\alpha})$.
- (d) Show that $o_{\text{pr}}(1) + o_{\text{pr}}(1) = o_{\text{pr}}(1)$; $o_{\text{pr}}(1) + O_{\text{pr}}(1) = O_{\text{pr}}(1)$; $o_{\text{pr}}(1)O_{\text{pr}}(1) = o_{\text{pr}}(1)$; $1/(1 + o_{\text{pr}}(1)) = O_{\text{pr}}(1)$; $o_{\text{pr}}(O_{\text{pr}}(1)) = o_{\text{pr}}(1)$; and $o_{\text{pr}}(1)/(1 + o_{\text{pr}}(1)) = o_{\text{pr}}(1)$.
- (e) Let $r: \mathbb{R}^k \rightarrow \mathbb{R}$ be a function such that $g(0) = 0$, and X_n a sequence of random vectors. Show that if $r(x) = o(\|x\|^p)$ as $x \rightarrow 0$, then $R(X_n) = o_{\text{pr}}(\|X_n\|^p)$; and that if $r(x) = O(\|x\|^p)$ as $x \rightarrow 0$, then $r(X_n) = O_{\text{pr}}(\|X_n\|^p)$.
- (f) Let $g: \mathbb{R} \rightarrow \mathbb{R}$ be a function such that $g(\theta) - g(\theta_0) = g'(\theta_0)(\theta - \theta_0) + o(|\theta - \theta_0|)$ as $\theta \rightarrow \theta_0$, and let $\hat{\theta}_n$ be an estimator sequence such that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution. Show that $\sqrt{n}\{g(\hat{\theta}_n) - g(\theta_0)\} = g'(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_{\text{pr}}(1)$, using the rules about $o_{\text{pr}}(\cdot)$ and $O_{\text{pr}}(\cdot)$ that you have derived above in the above exercises.
- (g) Suppose that $X_n = o_{\text{pr}}(n^{-\alpha})$ for some $\alpha > 0$, and that $n^\alpha X_n$ is dominated by an integrable random variable Y , or is uniformly integrable (which is weaker, as we recall from Ex. 2.5). Show that $E X_n = o(n^{-\alpha})$.

Tightness, Skorokhod, Helly, Prokhorov, Lévy

Ex. 2.23 *Skorokhod's theorem.* If $X_n \rightarrow_d X$, we can construct random variables Y_n and Y on a common probability space such that $Y_n \sim X_n$ and $Y \sim X$ and $Y_n(\omega) \rightarrow Y(\omega)$ for all ω . This result is interesting in its own right, but also leads to simplifications of certain proofs.

(a) Let F be a c.d.f. in $\mathbb{R}$ and suppose that the inverse function $F^{-1}: [0, 1] \rightarrow \mathbb{R}$ exists, and for $U \sim \text{unif}(0, 1)$ define $Y = F^{-1}(U)$. Show that $Y \sim F$. A more formal way of presenting this result is to consider the probability space $([0, 1], \mathcal{B}, \text{Leb})$, i.e., the unit interval equipped with the Borel- σ -algebra and Lebesgue measure, and define $Y(\omega) = F^{-1}(\omega)$ for each $\omega \in (0, 1)$. Then that $\text{Leb}(Y \leq y) = \text{Leb}(\{\omega \in (0, 1): Y(\omega) \leq y\}) = F(y)$, which is to say that Y has distribution F .

(b) If F^{-1} does not exist, the results from (a) still holds. Define

$$Y(\omega) = \inf\{x: F(x) \geq \omega\}, \quad \text{for } \omega \in (0, 1).$$

Argue that $\{x: F(x) \geq \omega\} = [Y(\omega), \infty)$ since F is a c.d.f., in particular nondecreasing and right-continuous; and that consequently $F(x) \geq \omega$ if and only if $Y(\omega) \leq x$. Show that $\text{Leb}(Y \leq y) = F(y)$ yet again.

(c) Let F_n and F be c.d.f.s on $\mathbb{R}$ and suppose that $F_n \Rightarrow F$. Construct $Y_n(\omega) = \inf\{x: F_n(x) \geq \omega\}$ and $Y(\omega) = \inf\{x: F(x) \geq \omega\}$ as above. Given $\varepsilon > 0$, choose a continuity point x of F such that $Y(\omega) - \varepsilon < x < Y(\omega)$, and for $\omega' > \omega$ choose a continuity point y of F such that $Y(\omega) \leq Y(\omega') < y < Y(\omega') + \varepsilon$. Show that $Y(\omega) \leq \liminf_n Y_n(\omega) \leq \limsup_n Y_n(\omega) \leq Y(\omega')$, and use that Y only has countably many points of discontinuity to conclude that $Y_n(\omega) \rightarrow Y(\omega)$ for Lebesgue almost all $\omega \in (0, 1)$. Finally, modify Y_n and Y on the null set to get pointwise convergence.

(d) Let $(X_n)_{n \geq 1}$ be an uniformly integrable sequence of random variables such that $X_n \rightarrow_d X$. Combine Skorokhod's theorem with Ex. 2.5(e) to show that X is integrable and $E X_n \rightarrow E X$.

(e) Prove the continuous mapping theorem (see Ex. 2.20(b)) for random variables, this time using Skorokhod's theorem.

(f) For random variables, use Skorokhod's theorem to prove that (2.4) implies (i) and (iv) of the Portmanteau theorem.

Ex. 2.24 Tightness. Consider the sequence $(X_n)_{n \geq 1}$ of random variables with distribution $\Pr(X_n = x) = 1/3$ for $x = 0, 1/2, n$. You may verify that the sequence $F_n(x) = \Pr(X_n \leq x)$ of distribution functions converges pointwise the limiting function,

$$G(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{3}, & 0 \leq x < \frac{1}{2}, \\ \frac{2}{3}, & x \geq \frac{1}{2}. \end{cases}$$

This function, albeit a limit of distributions functions, is *not* itself a distribution function. The problem is that $G(\infty) = 2/3$, meaning that one third of the probability mass of $F_n(x)$ has escaped to infinity. Tightness is a condition ensuring that a limit of distribution functions is itself a distribution function. A sequence $X_1, X_2, \dots$ of random variables is tight if for any $\varepsilon > 0$ there exists a constant K so that

tightness

$$\Pr(|X_n| > K) < \varepsilon, \quad \text{for all } n.$$

A tight sequence of random variables is also said to be bounded in probability.

(a) Show that $(X_n)_{n \geq 1}$ is tight if and only if for any $\varepsilon > 0$ there are a and b such that

$$\Pr\{X_n \in (a, b]\} > 1 - \varepsilon, \quad \text{for all } n,$$

and that this is the same as requiring that for any $\varepsilon > 0$ we can find a and b such that $F_n(a) < \varepsilon$ and $F_n(b) > 1 - \varepsilon$ for all n , in terms of the c.d.f.s F_n of X_n .

(b) Show that $(X_n)_{n \geq 1}$ is tight if and only if for any $\varepsilon > 0$ there is a $K > 0$ so that $\limsup_n \Pr(|X_n| > K) < \varepsilon$.

(c) Show that $(X_n)_{n \geq 1}$ is tight if and only if for any sequence of constants $a_1, a_2, \dots \geq 0$ tending to zero, $a_n Y_n \rightarrow_p 0$.

(d) Suppose that for some $\delta > 0$ there is an $M > 0$ and an $n_0 \geq 1$ such that $E|X_n|^\delta \leq M$ for all $n \geq n_0$. Show that $(X_n)_{n \geq 1}$ is tight.

(e) Show that if $(X_n)_{n \geq 1}$ is uniformly integrable, then it is tight.

(f) Let $(X_n)_{n \geq 1}$ be a tight sequence of random variables with c.d.f.s $(F_n)_{n \geq 1}$. Suppose that there is a non-decreasing and right-continuous function F such that $F_n(x) \rightarrow F(x)$ at each continuity point of F . Show that F must be a c.d.f., and hence $X_n \rightarrow_d X$, where the random variable X has distribution F .

Ex. 2.25 Helly's theorem. The theorem we are concerned with here says that for any sequence $(F_n)_{n \geq 1}$ of c.d.f.s there is a non-decreasing and right-continuous function F and a subsequence $(F_{n_k})_{k \geq 1}$ such that $F_{n_k}(x) \rightarrow F(x)$ at each x where F is continuous. The limit function F need not be a c.d.f., as it may lack one of the three defining properties of a c.d.f. We know from Ex. 2.24(f), however, that if $(F_n)_{n \geq 1}$ is a tight sequence, then F is a bona fide c.d.f. In this exercise we prove Helly's theorem for c.d.f.s, first on the real line, then on $\mathbb{R}^k$.

(a) Let $\mathbb{Q} = \{q_1, q_2, \dots\}$ be the rational numbers, and consider the infinite array

$$\begin{array}{cccc} F_1(q_1) & F_2(q_1) & F_3(q_1) & \dots \\ F_1(q_2) & F_2(q_2) & F_3(q_2) & \dots \\ F_1(q_3) & F_2(q_3) & F_3(q_3) & \dots \\ \vdots & \vdots & \vdots & \end{array}$$

Since $F_n(q_1)$ lies between zero and one for all n , the Bolzano–Weierstrass theorem says that there is a convergent subsequence $(F_{n_{1,k}}(q_1))_{k \geq 1}$, say $F_{n_{1,k}}(q_1) \rightarrow G(q_1)$ as $k \rightarrow \infty$. The sequence $(F_{n_{1,k}}(q_2))_{k \geq 1}$ is also bounded between zero and one, so we can extract a further subsequence $(n_{2,k})_{k \geq 1} \subset (n_{1,k})_{k \geq 1}$ such that $F_{n_{2,k}}(q_2)$ converges to a limit $G(q_2)$ as $k \rightarrow \infty$. Continuing in this manner we get the array

the diagonal
method

$$\begin{array}{cccc} n_{1,1} & n_{1,2} & n_{1,3} & \dots \\ n_{2,1} & n_{2,2} & n_{2,3} & \dots \\ n_{3,1} & n_{3,2} & n_{3,3} & \dots \\ \vdots & \vdots & \vdots & \end{array}$$

where each row is included in the row above, i.e., $(n_{j,k})_{k \geq 1} \subset (n_{j-1,k})_{k \geq 1}$ for $j = 2, 3, \dots$. Argue that the diagonal sequence $n_k = n_{k,k}$ is such that $F_{n_k}(q) \rightarrow G(q)$ for each $q \in \mathbb{Q}$.

(b) Define the function

$$F(x) = \inf\{G(q) : q > x, q \in \mathbb{Q}\}.$$

Since $F(x)$ is an infimum over an set that shrinks with growing x , $F(x)$ is non-decreasing. Also, as $F(x)$ is an infimum we can for any x and any $\varepsilon > 0$ find an $s \in \mathbb{Q}$ with $s > x$ such that $G(s) < F(x) + \varepsilon$. Use this to prove that $F(x)$ must be right-continuous.

(c) Let x be a continuity point of F (see Ex. C.2(c) for the why such points exist). Argue that for any $\varepsilon > 0$ we can find $y \in \mathbb{R}$ and $s, r \in \mathbb{Q}$ such that $y < r < x < s$ and

$$F(x) - \varepsilon \leq G(r) \leq G(s) \leq F(x) + \varepsilon.$$

Use these inequalities to prove that $\lim_k F_{n_k}(x) = F(x)$, proving Helly's theorem.

(d) Let $\mathbb{Q}^k = \{q_1, q_2, \dots\}$ be an enumeration of the k -dimensional vectors with rational components. Mimic the one-dimensional proof and generalise Helly's theorem to sequences of c.d.f.'s on $\mathbb{R}^k$. The changes you need to make are essentially notational.

Ex. 2.26 Prokhorov's theorem. The theorem worked with here comprises (a) and (b) of this exercise. With Helly's theorem at hand, part (a) was essentially proved in Ex. 2.24(f), but we include it here for completeness.

(a) If $(X_n)_{n \geq 1}$ is tight, then there is a subsequence $(n_k)_{k \geq 1}$ such that $X_{n_k} \rightarrow_d X$ for some random variable X .

(b) If $X_n \rightarrow_d X$ for some random variable X , then $(X_n)_{n \geq 1}$ is tight.

(c) The importance of Prokhorov's theorem stems from the following lemma, which you are to show: The sequence $(X_n)_{n \geq 1}$ converges in distribution to X if and only if each subsequence $(X_{n_k})_{k \geq 1}$ contains a further subsequence $(X_{n_{k_j}})_{j \geq 1}$ that converges in distribution to X . The proof is similar to the analogous result from real analysis.

(d) Suppose that $(X_n)_{n \geq 1}$ is tight, and that every subsequence of X_{n_k} that converges weakly at all, converges to the same random variable X . Show that then $X_n \rightarrow_d X$.

Ex. 2.27 *Characteristic functions converging to a characteristic function.* Let $(X_n)_{n \geq 1}$ be a sequence of random variables with characteristic functions $\varphi_1(t), \varphi_2(t), \dots$. Suppose that $\varphi_n(t) \rightarrow \varphi(t)$, and that we recognise $\varphi(t)$ as the characteristic function of some random variable X . In such cases we can conclude that $X_n \rightarrow_d X$ without further ado.

(a) Suppose that X_n and X have characteristic functions $\varphi_n(t)$ and $\varphi(t)$, respectively, and that $\varphi_n(t) \rightarrow \varphi(t)$, where $\varphi(t)$ is a characteristic function. As an auxiliary assumption, suppose that $\varphi_n(t)$ is dominated by a function g with finite integral. By Ex. C.25(h) the X_n then have densities

$$f_n(x) = \frac{1}{2\pi} \int \exp(-itx) \varphi_n(t) dt.$$

Show that under these assumptions $X_n \rightarrow_d X$.

(b) Next, we reduce the general case to the special case studied in (a), by using the result from Ex. 2.19(c). Suppose that X_n and X have characteristic functions $\varphi_n(t)$ and $\varphi(t)$, respectively, and that $\varphi_n(t) \rightarrow \varphi(t)$. Let Z be a standard normal random variable, and σ some positive constant. Show that $X_n + \sigma Z \rightarrow_d X + \sigma Z$, and conclude that $\varphi_n(t) \rightarrow \varphi(t)$ implies $X_n \rightarrow_d X$.

Ex. 2.28 *Lévy's continuity theorem and some more.* In Ex. 2.27 we made two crucial assumptions: that $\varphi_n(t) \rightarrow \varphi(t)$, and that the limiting function φ is indeed a characteristic function. What if we drop the second assumption and only assume that the characteristic functions φ_n converge to *some* function φ ? Lévy's continuity theorem says that if this limiting function is continuous at zero, then it must be a characteristic function. In more terse prose, if $\varphi_n(t) = E \exp(itX_n)$ and $\varphi_n(t) \rightarrow \varphi(t)$ with φ some function continuous at zero, then φ is the characteristic function of some random variable X , and $X_n \rightarrow_d X$. The key to this theorem is that the continuity of φ at zero entails that $(X_n)_{n \geq 1}$ is tight. In fact, given that $\varphi_n(t) \rightarrow \varphi(t)$ for some function $\varphi(t)$, there is equivalence between the following three statements:

- (i) $\varphi(t)$ is a characteristic function,
- (ii) the X_n sequence is tight;
- (iii) $\varphi(t)$ is continuous at zero;

and, as we know from Ex. 2.27, the convergence $\varphi_n(t) \rightarrow \varphi(t)$ combined with (iii) implies that convergence in distribution takes place. The implication (i) $\Rightarrow$ (iii) is immediate from the uniform continuity of characteristic functions, see Ex. C.24; and (i) $\Rightarrow$ (ii) follows from combining Ex. 2.26 and Ex. 2.27. In (a) of this exercise we prove (ii) $\Rightarrow$ (i); while (iii) $\Rightarrow$ (i), which is Lévy's continuity theorem, is proven in (b)–(c).

(a) Suppose that X_n has characteristic function φ_n for $n = 1, 2, \dots$, and that $(X_n)_{n \geq 1}$ is tight. Use the lemma in Ex. 2.26(c) to show that if $\varphi_n(t) \rightarrow \varphi(t)$ for *some* function φ , then $\varphi(t)$ must be a characteristic function.

(b) Start by applying Fubini's theorem (see Ex. B.19) to show that if X has characteristic function φ and cumulative c.d.f. F , then

$$\int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\} dt = 2\varepsilon \int \left\{1 - \frac{\sin(x\varepsilon)}{x\varepsilon}\right\} dF(x).$$

In particular, the integral of $\varphi(t)$ on a symmetric interval around zero is really a real number, that is, the complex component disappears. Deduce that

$$\frac{1}{\varepsilon} \int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\} dt \geq 2 \int_{|x\varepsilon| \geq c} \left\{1 - \frac{\sin(x\varepsilon)}{x\varepsilon}\right\} dF(x) \geq 2(1 - 1/c) \Pr\{|X| \geq c/\varepsilon\},$$

with the value $c = 2$ yielding the tail inequality

$$\Pr\{|X| \geq 2/\varepsilon\} \leq \frac{1}{\varepsilon} \int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\} dt. \quad (2.8)$$

(c) If we now have a collection of random variables, where their characteristic functions have approximately the same level of smoothness around zero, then we should get tightness from (2.8). Assume that $X_1, X_2, \dots$ have characteristic functions $\varphi_1, \varphi_2, \dots$, and that $\varphi_n(t)$ converges to *some* function $\varphi(t)$ that is continuous at zero. For a given $\varepsilon' > 0$, find $\varepsilon > 0$ such that $|1 - \varphi(t)| \leq \varepsilon'$ for $|t| \leq \varepsilon$. Show that

$$\limsup_{n \rightarrow \infty} \Pr\{|X_n| \geq 2/\varepsilon\} \leq \frac{1}{\varepsilon} \int_{-\varepsilon}^{\varepsilon} \{1 - \varphi(t)\} dt \leq 2\varepsilon'.$$

We've hence found a broad interval, namely $[-2/\varepsilon, 2/\varepsilon]$, inside which each X_n lies, with high enough probability, which means that $(X_n)_{n \geq 1}$ is tight, by Ex. 2.24(b).

(d) As a bonus, show that if $X_1, X_2, \dots$ is a tight sequence of random variables with characteristic functions $\varphi_1, \varphi_2, \dots$, then $(\varphi_n)_{n \geq 1}$ is uniformly equicontinuous. That is, for each $\varepsilon > 0$ there is a $\delta > 0$ such that $|t - s| < \delta$ implies $|\varphi_n(t) - \varphi_n(s)| < \varepsilon$ for all n .

Ex. 2.29 *Showing convergence in distribution via moment generating functions.* Here we spell out an m.g.f. analogue of the Lévy continuity theorem, see Ex. 2.28. Let $(X_n)_{n \geq 1}$ have m.g.f.s $\{M_n(t)\}_{n \geq 1}$ and suppose that all these exist on the same interval $[-a, a]$.

(a) Expand on Ex. 2.16 to show that

$$\Pr(|X_n| \geq K) \leq \exp(-Ka)\{M_n(a) + M_n(-a)\}, \quad \text{for all } n,$$

and use this to argue that if $M_n(t)$ converges to *some* function $M(t)$ for all $t \in [-a, a]$, then $(X_n)_{n \geq 1}$ is tight.

(b) Deduce from (a) that $X_n \rightarrow_d X$, where X is a random variable with m.g.f. $M(t)$. Note that since $x \mapsto \exp(tx)$ is not bounded, this last part is more involved than in the characteristic function case.

(c) For illustration, consider $X_1, \dots, X_n$ being independent 0-1 Bernoulli variables with $p_i = \Pr(X_i = 1)$. Show that if $\max_{i \leq n} p_i \rightarrow 0$ and $\sum_{i=1}^n p_i \rightarrow \lambda$, then $\sum_{i=1}^n X_i \rightarrow_d \text{Pois}(\lambda)$. Note that this generalises the classical binomial to Poisson theorem, see Ex. 2.8, and also that proving this result directly from the point probabilities would be difficult.

(d) Suppose $X_1, X_2, \dots$ are independent Bernoulli with $p_i = i/n$, and consider $Y_n(t) = \sum_{i \leq t\sqrt{n}} X_i$. Show that $Y_n(t) \rightarrow_d \text{Pois}(\frac{1}{2}t^2)$. The limit is actually a full Poisson process in t , with independent increments (see Ch. 10).

Ex. 2.30 *Showing convergence in two steps.* Suppose one wishes to prove that $X_n \rightarrow_d X$, but that technical issues make it easier to first prove that an approximation to X_n converges to an approximation to X . With a suitable extra condition this might suffice. Here we suppose that the random elements involved take values in a common (separable) metric space $(\mathcal{X}, d)$.

(a) For the approximations $A_{n,k}$ to X_n and A_k to X , suppose (i) that $A_{n,k} \rightarrow_d A_k$, for each k ; (ii) that $A_k \rightarrow_d X$ as $k \rightarrow \infty$; and also (iii) that

$$\lim_{k \rightarrow \infty} a_k(\varepsilon) \rightarrow 0 \text{ for each } \varepsilon > 0, \quad \text{where } a_k(\varepsilon) = \limsup_{n \rightarrow \infty} \Pr(d(X_n, A_{n,k}) \geq \varepsilon) = 0.$$

Consider the closed sets F and $F^\varepsilon = \{x: d(x, F) \leq \varepsilon\}$, and use the Portmanteau theorem to show that $X_n \rightarrow_d X$.

- (b) Suppose somebody clever has managed to prove the CLT for bounded i.i.d. variables (for example using m.g.f.s): If $A_1, A_2, \dots$ are i.i.d. with mean zero and variance one, and $|A_i| \leq k$, then $Z_n = \sqrt{n} \bar{A}_n \rightarrow_d N(0, 1)$. How can you then prove that the CLT is valid also for unbounded i.i.d. random variables, as long as only their variance is finite?
- (c) (xx we invent one more example. xx)

Central limit theorems

Ex. 2.31 *Applying the CLT: binomial, Poisson, chi-squares, etc.* We have formulated the basic CLT already in (2.5). In exercises below we develop tools used to prove this, along with generalisations and related statements, but we can start out by seeing how to apply it.

- (a) For $X \sim \text{binom}(n, p)$, show that $(X - np)/\{np(1-p)\}^{1/2} \rightarrow_d N(0, 1)$. de Moivre 1738,
Laplace 1812
- (b) Let $Y \sim \text{Pois}(\lambda)$. Write Y as a sum of i.i.d. variables, and show that Y is approximately a $N(\lambda, \lambda)$, for growing λ .
- (c) Similarly, for $Z \sim \chi_n^2$: show that $(Z - n)/(2n)^{1/2} \rightarrow_d N(0, 1)$. For $n = 100$, compute $\Pr(Z_n \in [90, 110])$ via normal approximation, and check with the exact probability.
- (d) Let $X_1, \dots, X_n$ be i.i.d. from the normal mixture $f = 0.90 N(0, 1) + 0.10 N(0, 10)$. Find a normal approximation formula for the distribution of $\bar{X}_n$.
- (e) Suppose $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are independent samples from two distributions with common mean and finite variances. With sample means $\bar{X}, \bar{Y}$ and sample variances $\hat{\sigma}_1^2, \hat{\sigma}_2^2$, show that $(\bar{Y} - \bar{X})/(\hat{\sigma}_1^2/n + \hat{\sigma}_2^2/n)^{1/2}$ tends to the standard normal, perhaps via Slutsky–Cramér.

Ex. 2.32 *The central limit theorem, Lindeberg's proof.* Let $X_1, X_2, \dots$ be i.i.d. random variables with mean zero and unit variance, define $X_{n,i} = X_i/\sqrt{n}$, and let $Z \sim N(0, 1)$. We will show that $E f(\sum_{i=1}^n X_{n,i}) \rightarrow E f(Z)$ for all bounded and continuous functions with bounded and continuous derivatives of all orders, which, by Ex. 2.19 is equivalent to $\sum_{i=1}^n X_{n,i} \rightarrow_d N(0, 1)$. Introduce $Z_{n,i} = Z_i/\sqrt{n}$ for $i = 1, 2, \dots, n$ and $n \geq 1$, where the $Z_1, Z_2, \dots$ are i.i.d. standard normal random variables independent of the $X_1, X_2, \dots$. In particular, $\sum_{i=1}^n Z_{n,i} \sim N(0, 1)$ by Ex. C.25, for example.

- (a) The essence of Lindeberg's idea is to show that $E f(\sum_{i=1}^n X_{n,i}) - E f(\sum_{i=1}^n Z_{n,i})$ tends to zero by replacing the summands $X_{n,i}$ by the Gaussian summands $Z_{n,i}$, one by one. Show that by so doing the difference $f(\sum_{i=1}^n X_{n,i}) - f(\sum_{i=1}^n Z_{n,i})$ can be expressed back in 1922 as the (telescoping) sum

$$f\left(\sum_{i=1}^n X_{n,i}\right) - f\left(\sum_{i=1}^n Z_{n,i}\right) = \sum_{k=1}^n \{f(A_k) - f(A_{k-1})\} = \sum_{k=1}^n \{f(R_k + X_{n,k}) - f(R_k + Z_{n,k})\},$$

where $A_k = \sum_{i=1}^k X_{n,i} + \sum_{i=k+1}^n Z_{n,i}$ and $R_k = \sum_{i=1}^{k-1} X_{n,i} + \sum_{i=k+1}^n Z_{n,i}$, so that $A_n = \sum_{i=1}^n X_{n,i}$ and $A_0 = \sum_{i=1}^n Z_{n,i}$.

- (b) Let f be a bounded and continuous functions with bounded and continuous derivatives up to and including the third order. Show that for any $\varepsilon > 0$ there are $\delta > 0$ and K such that

$$|f(x+y) - f(x) - f'(x)y - \frac{1}{2}f''(x)y^2| \leq \varepsilon y^2 I(|y| < \delta) + Ky^2 I(|y| \geq \delta),$$

uniformly in x and y .

(c) Verify that for each $k = 1, \dots, n$ and $n \geq 1$, $X_{n,k}$ is independent of R_k and that $Z_{n,k}$ is independent of R_k . Use this with the above to show that given $\varepsilon > 0$ we can choose $\delta > 0$ and find K such that

$$\left| \mathbb{E} f\left(\sum_{i=1}^n X_{n,i}\right) - \mathbb{E} f\left(\sum_{i=1}^n Z_{n,i}\right) \right| \leq 2\varepsilon + K \sum_{k=1}^n \mathbb{E} X_{n,k}^2 I_{|X_{n,k}| \geq \delta} + K \sum_{k=1}^n \mathbb{E} Z_{n,k}^2 I_{|Z_{n,k}| \geq \delta}.$$

Please conclude, and you will have shown the CLT for i.i.d. random variables.

(d) Suppose $Y_1, Y_2, \dots$ are i.i.d. with finite mean ξ and standard deviation σ . Show that the random sum, normalised to have mean zero and variance one, becomes normal:

$$\left(\sum_{i=1}^n Y_i - n\xi\right)/\sqrt{n\sigma^2} = \sqrt{n}(\bar{Y}_n - \xi)/\sigma \rightarrow_d N(0, 1).$$

Ex. 2.33 *The Lindeberg theorem.* We now extend the central limit theorem to independent random variables that do not necessarily have the same distribution.

Lindeberg CLT,
1922

(a) Let $X_1, X_2, \dots$ be independent mean zero random variables with variances $\sigma_1^2, \sigma_2^2, \dots$, and form $X_{n,j} = X_j/B_n$, where $B_n^2 = \sigma_1^2 + \dots + \sigma_n^2$. In view of the remainder term in Ex. 2.32(c), the Lindeberg condition is natural: Assume that

$$\sum_{j=1}^n \mathbb{E} X_{n,j}^2 I(|X_{n,j}| \geq \varepsilon) \rightarrow 0, \quad \text{as } n \rightarrow \infty, \text{ for each } \varepsilon > 0. \quad (2.9)$$

Show that the Lindeberg condition implies that $\max_{i \leq n} \sigma_i^2/B_n^2 \rightarrow 0$ as $n \rightarrow \infty$; and proceed to show that $\sum_{j=1}^n X_{n,j} \rightarrow_d N(0, 1)$ by making the appropriate modifications to the proof for the i.i.d. case studied in Ex. 2.32.

(b) Suppose $X_1, X_2, \dots$ are i.i.d. with mean zero and variance σ^2 , then form from these $Z_n = \sum_{i=1}^n iX_i/(\sum_{i=1}^n i^2)^{1/2}$. Show that $Z_n \rightarrow_d N(0, \sigma^2)$. Generalise.

Ex. 2.34 *Convergence in distribution for bounded variables.* Suppose variables X_n all live on a bounded interval, which we for simplicity and without loss of generality below take to be $[0, 1]$.

(a) With X a candidate limit variable, also on $[0, 1]$, we know from the Portmanteau theorem of Ex. 2.18 that $X_n \rightarrow_d X$ if and only if $\mathbb{E} g(X_n) \rightarrow \mathbb{E} g(X)$ for each bounded continuous g . Show via the Weierstraß polynomial approximation theorem of Ex. 2.17 that this is also equivalent to $\mathbb{E} X_n^k \rightarrow \mathbb{E} X^k$ for each k .

(b) Then show that this again is equivalent to convergence of the m.g.f.s, i.e. $M_n(t) \rightarrow M(t)$ for each t . For an illustration with an infinite sum of Bernoullis, check Story #85.

(c) We saw in Ex. 2.7 that if X_n is uniform on $1/n, 2/n, \dots, n/n$, then $X_n \rightarrow_d U$, the uniform on the unit interval. Show from this that $s_n(k) = (1/n^{k+1})(1^k + 2^k + \dots + n^k) \rightarrow 1/(k+1)$ for each k .

Ex. 2.35 *Proving the CLT with moment generating functions.* Here we prove a central limit theorem for i.i.d. random variables $X_1, X_2, \dots$ whose m.g.f. $M(t) = \mathbb{E} \exp(tX_1)$ exist on an interval around zero. See Ex. C.21 for what this assumption entails about the random variables. Throughout, we assume that $\mathbb{E} X_1 = 0$ and $\text{Var } X_1 = \sigma^2$. The CLT proof via characteristic functions is clearly more general, but in various cases it is simply more convenient to give proofs via m.g.f.s.

(a) Show that $M(t) = 1 + \frac{1}{2}\sigma^2 t^2 + o(t^2)$ as $t \rightarrow 0$.

(b) Show that $\sqrt{n}\bar{X}_n = n^{-1/2} \sum_{i=1}^n X_i$ has moment generating function of the form

$$M_n(t) = M(t/\sqrt{n})^n = \{1 + \frac{1}{2}\sigma^2 t^2/n + o(t^2/n)\}^n,$$

and conclude that $M_n(t) \rightarrow \exp(\frac{1}{2}t^2\sigma^2)$ as $n \rightarrow \infty$. Explain, in view of Ex. 2.29, that this proves the CLT, under these m.g.f. conditions.

Ex. 2.36 *Two useful lemmas.* In proving the central limit theorems below, the following two lemmas will be useful.

(a) Demonstrate that with $z_{n,1}, z_{n,2}, \dots$ a sequence of real numbers coming closer to zero, we have $\prod_{i=1}^n (1 + z_{n,i}) \rightarrow \exp(z)$, provided (i) $\sum_{i=1}^n z_{n,i} \rightarrow z$; (ii) $\max_{i \leq n} |z_{n,i}| \rightarrow 0$; and (iii) $\sum_{i=1}^n |z_{n,i}|$ stays bounded. It may be helpful to show first that

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots = x + x^2 K(x),$$

with $K(x)$ a continuous function such that $|K(x)| \leq 1$ for $|x| \leq \frac{1}{2}$, and $K(x) \rightarrow -\frac{1}{2}$ when $x \rightarrow 0$. These statements are valid also when the $z_{n,i}$ and the x are complex numbers inside the unit ball, in which case the logarithm is the natural complex extension of the real logarithm.

(b) Let $x_1, \dots, x_n$ and $y_1, \dots, y_n$ be complex numbers such that $|x_j|, |y_j| \leq 1$ for all j . Show that $|x_1 x_2 - y_1 y_2| \leq |x_1 - y_1| + |x_2 - y_2|$, and proceed by induction to show that

$$\left| \prod_{j=1}^n x_j - \prod_{j=1}^n y_j \right| \leq \sum_{j=1}^n |x_j - y_j|, \quad \text{for all } n.$$

Ex. 2.37 *The Lindeberg CLT via truncations.* Here we present a proof of the CLT for independent random variables using m.g.f.s., but without assuming that the m.g.f.s. of the random variables involved necessarily exist. Since the m.g.f. of a bounded random variable always exists (prove it), the trick is to truncate the random variables involved: Let $X_1, X_2, \dots$ be independent random variables with mean zero and variances $\sigma_1^2, \sigma_2^2, \dots$. Write $B_n^2 = \sigma_1^2 + \dots + \sigma_n^2$ and $X_{n,j} = X_j/B_n$ for $j = 1, \dots, n$. We are to show that $\sum_{j=1}^n X_{n,j}$ converges in distribution to a standard normal random variable, provided the Lindeberg condition (2.9) is satisfied, which we assume throughout.

(a) For some $\varepsilon > 0$, define $\mu_{n,j} = E X_{n,j} I(|X_{n,j}| < \varepsilon)$, as well as the random variables $Y_{n,j} = X_{n,j} I(|X_{n,j}| < \varepsilon) - \mu_{n,j}$ and $v_{n,j} = X_{n,j} I(|X_{n,j}| \geq \varepsilon) + \mu_{n,j}$. Verify that

$$X_{n,j} = Y_{n,j} + v_{n,j},$$

and also $|Y_{n,j}| \leq 2\varepsilon$, the latter implying that the m.g.f. of $Y_{n,j}$ exist for all j and n . Show that

$$E \exp(tY_{n,j}) = 1 + \frac{1}{2}t^2 E Y_{n,j}^2 + r_{n,j}(t),$$

where the remainders are such that $|\sum_{j=1}^n r_{n,j}(t)| \leq \varepsilon \exp(2\varepsilon)$ for $|t| \leq 1$.

(b) By Ex. 2.36(a), showing that $\prod_{j=1}^n E \exp(tY_{n,j}) \rightarrow \exp(\frac{1}{2}t^2)$ now comes down to verifying that (i) $\sum_{j=1}^n E Y_{n,j}^2 \rightarrow 1$; (ii) $\max_{j \leq n} E Y_{n,j}^2 \rightarrow 0$ and $\max_{j \leq n} |r_{n,j}| \rightarrow 0$; and (iii) $\limsup_{n \rightarrow \infty} |E \exp(tY_{n,j}) - 1|$ is finite. Please verify these conditions, remembering that it is sufficient to consider $|t| \leq 1$.

(c) Finally, show that $\sum_{j=1}^n v_{n,j} \rightarrow_{\text{pr}} 0$, and conclude that $\sum_{j=1}^n X_{n,j} \rightarrow_d N(0, 1)$.

Ex. 2.38 *Proving the CLT with characteristic functions.* It can be argued that the most elegant, unified, and general proof of the CLT is obtained by the use of characteristic functions. This is because, contrary to the m.g.f., the characteristic function of a random variable always exists.

(a) Show that if X has finite mean ξ , then its characteristic function satisfies $\varphi(t) = 1 + i\xi t + o(t)$ as $t \rightarrow 0$. Also, its derivative exists, with $\varphi'(t) = E iX \exp(itX)$, and in particular $\varphi'(0) = i\xi$ (see Ex. C.24 for more details).

(b) Show similarly that if X has finite variance σ^2 , then

$$\varphi(t) = 1 + i\xi t - \frac{1}{2}(\xi^2 + \sigma^2)t^2 + o(t^2) \quad \text{as } t \rightarrow 0.$$

(c) If $X_1, X_2, \dots$ are i.i.d. with mean zero and finite variance σ^2 , then show that $Z_n = \sqrt{n}\bar{X}_n = n^{-1/2} \sum_{i=1}^n X_i$ has characteristic function of the form

$$\varphi_n(t) = \{1 - \frac{1}{2}\sigma^2 t^2/n + o(1/n)\}^n.$$

Prove the CLT from this.

Ex. 2.39 *The Lindeberg CLT, yet again.* Let $(X_{n,i})_{i \leq n, n \geq 1}$ be a triangular array of mean zero random variables with variances $(\sigma_{n,i}^2)_{i \leq n, n \geq 1}$ such that $X_{n,1}, \dots, X_{n,n}$ are independent for each n . We say that the triangular array is row wise independent.

(a) Let $(Y_{n,i})_{i \leq n, n \geq 1}$ be a triangular array of row-wise independent mean zero normally distributed random variables with variances $\text{Var } Y_{n,i} = \sigma_{n,i}^2$ for all i and n . Using the second lemma of Ex. 2.36, show that for any $\varepsilon > 0$

$$\left| E \exp\left(it \sum_{i=1}^n X_{n,i}\right) - E \exp\left(it \sum_{i=1}^n Y_{n,i}\right) \right| \leq \frac{1}{3}|t|^3 \varepsilon \sum_{i=1}^n \sigma_{n,i}^2 + L_n(\varepsilon) + L'_n(\varepsilon),$$

where $L_n(\varepsilon) = \sum_{i=1}^n E X_{n,i}^2 I(|X_{n,i}| \geq \varepsilon)$ and $L'_n(\varepsilon) = \sum_{i=1}^n E Y_{n,i}^2 I(|Y_{n,i}| \geq \varepsilon)$.

(b) Deduce that $\sum_{i=1}^n X_{n,i} \rightarrow_d N(0, 1)$ provided $\sum_{i=1}^n \sigma_{n,i}^2 \rightarrow 1$ and the Lindeberg condition holds.

Ex. 2.40 *Lindeberg and Liapunov and some more.* In this exercise we summarise various implications pertaining to the Lindeberg condition. Let $X_1, X_2, \dots$ be independent mean zero random variables with variances $\sigma_1^2, \sigma_2^2, \dots$, set $B_n^2 = \sigma_1^2 + \dots + \sigma_n^2$, and define $X_{n,i} = X_i/B_n$. Recall that the Lindeberg condition is satisfied if for each $\varepsilon > 0$, we have $\sum_{i=1}^n E X_{n,i}^2 I(|X_{n,i}| \geq \varepsilon) \rightarrow 0$ as n tends to infinity.

(a) Show that if the Lindeberg condition holds, then $\max_{i \leq n} \sigma_i^2/B_n^2 \rightarrow 0$.

(b) Show that if B_n^2 is convergent, then the Lindeberg condition fails.

(c) In view of (b), we assume that B_n diverges. Show that the Lindeberg condition holds for uniformly bounded random variables, i.e., $|X_i| \leq M$ for all i .

(d) Show that if $E|X_i|^{2+\delta} < M$ for all i and some $\delta > 0$, still assuming that B_n diverges, of course, then the Lindeberg condition holds.

(e) Suppose that there is a sequence of constants $K_1, K_2, \dots$ such that $|X_i| < K_i$ for each i , and that $K_n/B_n \rightarrow 0$. Show that the Lindeberg condition holds.

(f) Show that the Liapunov condition, that is $\sum_{i=1}^n E|X_{n,i}|^{2+\delta} \rightarrow 0$ for some $\delta > 0$, implies the Lindeberg condition.

(g) Suppose that $n^{-1}B_n^2 \rightarrow \sigma^2 > 0$, and show that then

$$\frac{1}{n} \sum_{i=1}^n E X_i^2 I(|X_i| \geq \sqrt{n}\varepsilon) \rightarrow 0,$$

for each $\varepsilon > 0$ if and only if the Lindeberg condition holds; and that under these conditions $\sqrt{n}\bar{X}_n \rightarrow_d N(0, \sigma^2)$, where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$.

(h) Let $Y_1, Y_2, \dots$ be i.i.d. random variables with finite second moments. Show that $\max(Y_1, \dots, Y_n)/\sqrt{n}$ tends to zero in probability.

(i) Show that if $Y_1, Y_2, \dots$ are i.i.d. with mean μ and variance σ^2 , then the Lindeberg condition holds. Thus, the Lindeberg CLT contains the CLT for i.i.d. random variables.

Ex. 2.41 *Limiting normality of linear combinations of i.i.d. variables.* Let $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d. from some distribution with mean zero and finite variance σ^2 . For a sequence of multiplicative constants $a_1, a_2, \dots$, consider

$$Z_n = \frac{\sum_{i=1}^n a_i \varepsilon_i}{B_n} = \sum_{i=1}^n (a_i/B_n) \varepsilon_i, \quad \text{with } B_n^2 = \sigma^2 \sum_{i=1}^n a_i^2,$$

which has mean zero and variance 1. The question is what should be demanded of the a_i sequence, to ensure that $Z_n \rightarrow_d N(0, 1)$ (even if the ε_i distribution might be looking say skewed and multimodal and strange).

(a) Let $D_n = \max_{i \leq n} |a_i|/B_n$. Writing G for the distribution of ε_i , show that

$$\sum_{i=1}^n \mathbb{E} \left| \frac{a_i \varepsilon_i}{B_n} \right|^2 I \left(\left| \frac{a_i \varepsilon_i}{B_n} \right| \geq \delta \right) \leq \sum_{i=1}^n \frac{a_i^2}{B_n^2} \mathbb{E} \varepsilon_i^2 I(D_n |\varepsilon_i| \geq \delta) \leq \int_{|u| \geq \delta/D_n} u^2 dG(u),$$

and conclude that $Z_n \rightarrow_d N(0, 1)$ provided $D_n \rightarrow 0$.

(b) Under a variety of setups, one actually has $D_n \rightarrow 0$, which is hence not at all a strict condition. Verify that the condition holds, and hence limiting normality, in the following cases: (i) $a_i = 1$ (which corresponds to the plain CLT); (ii) all $|a_i|$ inside some positive $[b, c]$ interval; (iii) $a_i = i$; (iv) $a_i = i^2$ (and generalise); (v) $a_i = 1/\sqrt{i}$. Show however that the condition does not hold for $a_i = 1/i$.

(c) A grasshopper sits at zero and then starts jumping, to the right and left with equal probability, and with jump sizes $1, 2, 3, \dots$. With S_n her position after n jumps, show that $S_n/n^{3/2}$ has a normal limit. Though she keeps on passing zero, she will not be in that vicinity; show that $\Pr(|S_n| \leq cn^{1.49}) \rightarrow 0$, for each c .

(d) [xx fikse xx] Another important case to understand well is when the a_i can be considered an i.i.d. sequence, drawn from their own distribution. Show that $D_n \rightarrow_{\text{pr}} 0$ if the a_i distribution has finite variance. (xx nils thinks this is if and only if, actually. what happens with Z_n if the a_i are drawn from say the $1/|x|^2$ distribution, for $|x| \geq 1$? xx)

Ex. 2.42 *Toeplitz lemma for random variables.* The Toeplitz lemma says that if $a_1, a_2, \dots$ is a nonnegative sequence such that $\sum_{j=1}^{\infty} a_j = \infty$ and $x_1, x_2, \dots$ is a sequence such that $x_n \rightarrow x$, then $\sum_{j=1}^n a_j x_j / \sum_{j=1}^n a_j \rightarrow x$. If we set $a_j = 1$ for all j , the lemma gives $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i \rightarrow x$ provided $x_n \rightarrow x$. An interesting question is when this lemma carries over to convergence in probability.

Toeplitz lemma

(a) Prove the Toeplitz lemma.

(b) Let X_n be a sequence of independent random variables with distributions $\Pr(X_n = 0) = 1 - n^{-1}$ and $\Pr(X_n = n) = n^{-1}$. Show that $X_n \rightarrow_p 0$, and that the empirical mean $\bar{X}_n$ is such that its characteristic function $\mathbb{E} \exp(it\bar{X}_n) \rightarrow \exp\{\int_0^1 (e^{itx} - 1)/x dx\}$. Explain that the limit here is the characteristic function of a bona fide random variable, say Z . We are thus in the situation that $X_n \rightarrow_p 0$, while $\bar{X}_n \rightarrow_d Z$, which shows that Toeplitz lemma does not translate directly to convergence in probability.

(c) Suppose that X_n is a uniformly integrable sequence of random variables such that $X_n \rightarrow_{\text{pr}} 0$, and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Show that $\bar{X}_n \rightarrow_p X$.

Ex. 2.43 *Higher-order expansions of m.g.f.s.* (xx to be polished. the aim is to give more info for CLT, of the type $|F_n(t) - \Phi(t)| \leq c/\sqrt{n}$ for some c . xx)

(a) Consider a variable Y , with m.g.f. $M(t) = E \exp(tY)$, assumed to be finite in at least a neighbourhood around zero. We have seen in Ex. C.21 that $E Y^r = M^{(r)}(0)$. Write ξ and σ^2 for the mean and variance of Y . Show that $M(t) = 1 + \xi t + o(t)$, for $|t|$ small. Taking a Taylor expansion to the next step, show that $M(t) = 1 + \xi t + \frac{1}{2}(\xi^2 + \sigma^2)t^2 + o(t^2)$. Deduce also that $\log M(t) = \xi t + \frac{1}{2}\sigma^2 t^2 + o(t^2)$.

(b) We may also take the expansion to the third order, but it is simpler and more insightful to proceed from $Y = \xi + Y_0$, with Y_0 having mean zero. Show that

$$M(t) = \exp(t\xi) E \exp(tY_0) = \exp(t\xi) \{1 + \frac{1}{2}\sigma^2 t^2 + \frac{1}{6}\gamma_3 t^3 + o(|t|^3)\},$$

where $\gamma_3 = E(Y - \xi)^3$.

(c) Consider $Y_1, \dots, Y_n$ i.i.d. from a distribution with mean zero and m.g.f. $M(t)$ being finite around zero. Show that $Z_n = \sqrt{n} \bar{Y}$ has

$$\begin{aligned} M_n(t) &= E \exp(tZ_n) = M(t/\sqrt{n})^n \\ &= \{1 + \frac{1}{2}\sigma^2 t^2/n + \frac{1}{6}\gamma_3 t^3/n^{3/2} + o(|t|^3/n^{3/2})\}^n. \end{aligned}$$

Show from this that under the assumptions given, $\log M_n(t) = \frac{1}{2}\sigma^2 t^2 + \frac{1}{6}\gamma_3 t^3/\sqrt{n} + o(1/\sqrt{n})$. Explain why this is a proof of the CLT.

Ex. 2.44 *Characterisations of the normal and the Cauchy.* There are many characterisation theorems in probability theory, results saying that appropriate assumptions or properties fully characterise certain distributions- Here we give a few such, using characteristic functions.

(a) Suppose F is a distribution, symmetric around zero, such that if X and Y are independent from this distribution, then $(X + Y)/\sqrt{2}$ has the same distribution. Prove that the distribution is normal.

(b) Suppose F is a distribution, symmetric around zero, such that if X and Y are independent from this distribution, then $(X + Y)/2$ has the same distribution. Prove that the distribution is a Cauchy.

Joint convergence in distribution

Ex. 2.45 *Characteristic functions of random vectors and Cramér–Wold.* With $X = (X_1, \dots, X_k)^t$ a random vector, in dimension p , we define its characteristic functions as

$$\varphi(t_1, \dots, t_k) = E \exp(it^t X) = E \exp\{i(t_1 X_1 + \dots + t_k X_k)\}$$

for $t = (t_1, \dots, t_k)^t$. Let $F(x) = F(x_1, \dots, x_k)$ be the c.d.f. of X , and $\varphi(t) = \varphi(t_1, \dots, t_k)$ its characteristic function. From Ex. C.26 we have the inversion formula

$$F(b) - F(a) = \lim_{\sigma \rightarrow 0} \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \left\{ \prod_{j=1}^k \frac{\exp(-it_j b_j) - \exp(-it_j a_j)}{-it_j} \right\} \varphi(t) \exp(-\frac{1}{2}t^t \sigma^2) dt,$$

valid for all continuity points $a = (a_1, \dots, a_k)$ and $b = (b_1, \dots, b_k)$ of F , using the shorthand notation $dt = dt_1 \cdots dt_k$. Ex. C.26 also contains other results pertaining to the c.f. $\varphi(t_1, \dots, t_p)$ that you may need in this exercise.

(a) Let $X_n = (X_{n,1}, \dots, X_{n,p})^t$ be a sequence of random vectors with characteristic functions $\varphi_n(t_1, \dots, t_p)$. Suppose that $\varphi_n(t_1, \dots, t_p) \rightarrow \varphi(t_1, \dots, t_p)$, where φ is the characteristic function of some random vector $X = (X_1, \dots, X_p)^t$. Mimic the steps taken in Ex. 2.27 to show that then $X_n \rightarrow_d X$.

(b) Show that $X_n \rightarrow_d X$ if and only if $a^t X_n \rightarrow_d a^t X$ for each $a \in \mathbb{R}^k$, i.e., if and only if all linear combinations converge. This is the Cramér–Wold theorem, also known as the Cramér–Wold device. Cramér–Wold device

(c) Show that a random pair (X_n, Y_n) converges in distribution to the binormal $N_2(0, \Sigma)$, with Σ having diagonal elements 1, 1, and correlation ρ , if and only if $aX_n + bY_n \rightarrow_d N(0, a^2 + b^2 + 2\rho ab)$ for each (a, b) .

Ex. 2.46 *Tightness, Prokhorov, and the continuity theorem in $\mathbb{R}^k$.* A sequence of random vectors $X_n = (X_{n,1}, \dots, X_{n,k})^t$ is tight if for any $\varepsilon > 0$ there is a K such that

$$\Pr(\|X_n\| > K) < \varepsilon \quad \text{for all } n.$$

(a) Let F_n be the c.d.f. of X_n , and show that X_n being tight is equivalent to that for every $\varepsilon > 0$ there are $a = (a_1, \dots, a_k)$ and $b = (b_1, \dots, b_k)$ such that $F_n(b) - F_n(a) > 1 - \varepsilon$ for all n , where .

(b) Show that all the $X_{n,1}, \dots, X_{n,k}$ are tight if and only if X_n is tight. Deduce that X_n is tight if and only if any linear combination $a^t X_n = a_1 X_{n,1} + \dots + a_p X_{n,p}$ is tight.

(c) Show that if the sequence $(X_n)_{n \geq 1}$ of random vectors is tight, then the corresponding sequence of characteristic functions $\{\varphi_n(t_1, \dots, t_p)\}_{n \geq 1}$ is uniformly equicontinuous.

(d) Prokhorov’s theorem, proven for the one-dimensional case in Ex. 2.26, extends to the multi-dimensional case (and further, see Ex. 10.12): Demonstrate that (i) if $X_n \rightarrow_d X$, then X_n is tight; and (ii) that if $(X_n)_{n \geq 1}$ is tight, then there is a subsequence $(n_k)_{k \geq 1}$ such that $(X_{n_k})_{k \geq 1}$ converges in distribution. Recall that Helly’s theorem for random vectors was proven in Ex. 2.25(d) Prokhorov’s theorem in $\mathbb{R}^k$

(e) Let φ_n and F_n be, respectively, the characteristic function and the c.d.f. of the random vector $X_n = (X_{n,1}, \dots, X_{n,k})$ for $n = 1, 2, \dots$. Amend the derivations of Ex. 2.28 to show that for any $\varepsilon > 0$, and with $A_\varepsilon = (-\varepsilon, \varepsilon) \times \dots \times (-\varepsilon, \varepsilon)$,

$$\begin{aligned} \int_{A_\varepsilon} \{1 - \varphi_n(t)\} dt &\geq 2^k \varepsilon^k \int_{\cap_{j=1}^k \{|\varepsilon x_j| \geq c_k\}} \left\{1 - \prod_{j=1}^k \frac{\sin(x_j \varepsilon)}{x_j \varepsilon}\right\} dF_n(x_1, \dots, x_k) \\ &\geq 2^k \varepsilon^k (1 - 1/c_k^k) \Pr(\cap_{j=1}^k \{|X_{n,j}| \geq c_k/\varepsilon\}), \end{aligned}$$

so that with $c_k = 2/(2^k - 1)^{1/k}$ we get the inequality

$$\Pr(\cap_{j=1}^k \{|X_{n,j}| \geq c_k/\varepsilon\}) \leq \frac{1}{\varepsilon^k} \int_{-\varepsilon}^{\varepsilon} \dots \int_{-\varepsilon}^{\varepsilon} \{1 - \varphi_n(t_1, \dots, t_k)\} dt_1 \dots dt_k.$$

Lévy’s
continuity
theorem in
dim. k

Combine this inequality with (b) to show that if $\phi_n(t_1, \dots, t_k)$ converges to some limit function $\varphi(t_1, \dots, t_p)$ that is continuous at $(t_1, \dots, t_p) = (0, \dots, 0)$, then $(X_n)_{n \geq 1}$ is tight, and that consequently the limit function φ must be a c.f., which by Ex. 2.45(a) entails that $X_n \rightarrow_d X$, where X is a random variable with c.f. φ .

Ex. 2.47 *The multi-dimensional CLT.* Here we prove two canonical central limit theorems for random vectors taking values in $\mathbb{R}^k$.

(a) Suppose $X_1, X_2, \dots$ are i.i.d. random vectors with mean ξ and variance matrix Σ . Invoke the Cramér–Wold device to show that $\sqrt{n}(\bar{X}_n - \xi) \rightarrow_d N_k(0, \Sigma)$. multivariate CLT

multivariate
Lindeberg
theorem

(b) Let $X_1, X_2, \dots$ be independent random vectors of dimension k , with means $\xi_1, \xi_2, \dots$, and with finite positive definite variance matrices $\Sigma_1, \Sigma_2, \dots$. Let $B_n = (\Sigma_1 + \dots + \Sigma_n)^{1/2}$, and write $Z_{n,i} = B_n^{-1}(X_i - \xi_i)$. Show that

$$Z_n = (\Sigma_1 + \dots + \Sigma_n)^{-1/2} \sum_{i=1}^n (X_i - \xi_i) = \sum_{i=1}^n Z_{n,i} \rightarrow_d N_k(0, I_k),$$

provided the multivariate Lindeberg condition holds, i.e.,

$$\sum_{i=1}^n E \|Z_{n,i}\|^2 I(\|Z_{n,i}\| \geq \varepsilon) \rightarrow 0, \quad \text{for each } \varepsilon > 0. \quad (2.10)$$

(c) Show that the multi-dimensional Liapunov condition, that $\sum_{i=1}^n E \|Z_{n,i}\|^{2+\delta} \rightarrow 0$ for some $\delta > 0$, implies the Lindeberg condition in (2.10).

(d) Suppose that B_n^2/n converges to some positive definite Σ , and that the Lindeberg condition $n^{-1} \sum_{i=1}^n E \|X_i - \xi_i\|^2 I(\|X_i - \xi_i\| \geq \sqrt{n}\varepsilon) \rightarrow 0$ for each $\varepsilon > 0$, holds. Show that $n^{-1/2} \sum_{i=1}^n (X_i - \xi_i) \rightarrow_d N_p(0, \Sigma)$.

(e) Suppose that B_n^2/n converges to some positive definite Σ , and that there is a K and $\delta > 0$ such that $E |X_{i,j} - \xi_{i,j}|^{2+\delta} \leq K$ for $j = 1, \dots, k$ and for all $i \geq 1$. Show that $n^{-1/2} \sum_{i=1}^n (X_i - \xi_i) \rightarrow_d N_p(0, \Sigma)$.

Ex. 2.48 *Joint limiting normality for empirical mean and variance.* Suppose $X_1, X_2, \dots$ are i.i.d. with mean ξ and standard deviation σ . We assume that also the skewness and kurtosis are finite, $\gamma_3 = E(X_i - \xi)^3/\sigma^3$ and $\gamma_4 = E(X_i - \xi)^4/\sigma^4 - 3$.

(a) Show from the two-dimensional CLT that

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \xi) \\ \sqrt{n}(\hat{\sigma}_0^2 - \sigma^2) \end{pmatrix} \rightarrow_d N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \gamma_3\sigma^3 \\ \gamma_3\sigma^3 & \sigma^4(2 + \gamma_4) \end{pmatrix}\right),$$

where $\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^n (X_i - \xi)^2$.

(b) Then show that with $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, which is a ‘real estimator’, as opposed to $\hat{\sigma}_0^2$, which uses ξ , then we have $\sqrt{n}(\hat{\sigma}^2 - \hat{\sigma}_0^2) \rightarrow_{\text{pr}} 0$; see Ex. 2.21. Conclude that the two-dimensional limit distribution result above continues to hold with $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$ replacing $\sqrt{n}(\hat{\sigma}_0^2 - \sigma^2)$.

(c) Let in particular the distribution of the X_i be normal, so $X_i \sim N(\xi, \sigma^2)$. Show that γ_3 and γ_4 are equal to zero, so the general result above simplifies to

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \xi) \\ \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \end{pmatrix} \rightarrow_d N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right).$$

For another instructive application, involving the Gamma distribution, see Ex. 3.26.

(d) Let $X_1, \dots, X_n$ be independent Bernoulli variables with $\Pr(X_i = 1) = p$. With $\hat{p} = \bar{X} = Y/n$ the familiar binomial ratio, explain first that the empirical variance here is $\hat{\sigma}^2 = \hat{p}(1 - \hat{p})$. Then use the results above to find the joint limiting distribution of $\sqrt{n}(\hat{p} - p, \hat{\sigma}^2 - p(1 - p))$.

The delta method

Ex. 2.49 *The delta method: the basics.* There are important and often reasonably simple to use approximation methods, in probability theory and statistics, going by the name of

the delta method. It is related to functions of variables often being approximately linear, if the variables in question are not too spread out, with consequences for approximate normality.

(a) We start out in dimension one. Suppose in generic terms that $\sqrt{n}(A_n - a) \rightarrow_d Z$, with random variables A_n and a constant a . The limits are taken with respect to the index n tending to infinity, where n typically but not always is the size of an underlying sample. Show first that $A_n \rightarrow_{\text{pr}} a$. Now consider a function g , defined in at least a neighbourhood around a , and assumed to have a continuous derivative there. Show via the mean value theorem that $g(A_n) - g(a) = g'(B_n)(A_n - a)$, for some B_n between A_n and a , and use this to show that $\sqrt{n}\{g(A_n) - g(a)\} \rightarrow_d g'(a)Z$.

(b) For the typical case where the limit is a zero-mean normal, say $Z \sim N(0, \tau^2)$, explain that $\sqrt{n}\{g(A_n) - g(a)\} \rightarrow_d N(0, g'(a)^2 \tau^2)$.

(c) For the vector case, assume $\sqrt{n}(A_n - a) \rightarrow_d Z$ in dimension p , and consider a smooth function $g(a) = g(a_1, \dots, a_p)$, assumed to have continuous first order derivatives in a neighbourhood around a . Explain that $A_n \rightarrow_{\text{pr}} a$, so A_n will be inside that neighbourhood with probability tending to one. With $c = g^*(a)$ the vector of derivatives $\partial g(a)/\partial a_1, \dots, \partial g(a)/\partial a_p$, evaluated at a , show that $\sqrt{n}\{g(A_n) - g(a)\} \rightarrow_d c^t Z = c_1 Z_1 + \dots + c_p Z_p$. In particular, if $Z \sim N_p(0, \Sigma)$, explain that the limit $c^t Z$ is a $N(0, c^t \Sigma c)$. The main point is that the convergence in distribution, from the previous points, holds also jointly.

(d) Sometimes one also needs the vector-to-vector extension of results above, leading to the joint limit distribution of several such g functions. Let $g = (g_1, \dots, g_q)^t: \mathbb{R}^p \rightarrow \mathbb{R}^q$ be such a function, with component functions having first order derivatives at position a , yielding a Jacobi matrix $J = \partial g(a)/\partial a$ of dimension $q \times p$. Show that $\sqrt{n}\{g(A_n) - g(a)\} \rightarrow_d JZ$. In particular, if $Z \sim N_p(0, \Sigma)$, then the limit is $N_q(0, J\Sigma J^t)$.

(e) The delta method has been formulated here in terms of limit distributions, implying approximate normality for function of approximately normal variables. Explain that in the framework above, with $\sqrt{n}(A_n - a) \rightarrow_d N_p(0, \Sigma)$, then $g(A_n)$ is approximately normal, with mean $g(a)$ and variance $c^t \Sigma c/n$. The delta method hence gives approximations also to means, variances, and covariances of smooth functions of other variables.

(f) The $\sqrt{n}$ factor above comes from the most typical uses of these methods, where variances of estimators go to zero with speed $1/n$, in terms of sample size. Show however that the mathematics goes through for any increasing sequence; if $d_n(A_n - a) \rightarrow_d Z$, with d_n tending to infinity, then $d_n\{g(A_n) - g(a)\} \rightarrow_d g'(a)Z$. There are indeed situations where the rate might be $n^{2/3}$ or $n^{1/3}$.

Ex. 2.50 *Applying the delta method.* Here we exercise our delta method muscles, to see how the general recipes of Ex. 2.49 may be applied in a few situations. As is clear, once we have established limiting normality for certain quantities, e.g. via the CLT, there is a host of easy-to-harvest applications for functions of these start quantities.

(a) If $\sqrt{n}(A_n - a) \rightarrow_d N(0, 1)$, find out what happens to $\sqrt{n}(A_n^3 - a^3)$, and to $\sqrt{n}\{\exp(kA_n) - \exp(ka)\}$, where k is a constant.

(b) For Y a binomial (n, p) , we have of course $\text{Var } \hat{p} = p(1-p)/n$ for the classic estimator $\hat{p} = Y/n$. Use the delta method to find approximations to the means, variances, and distributions of (i) the estimated odds ratio $\hat{p}/(1-\hat{p})$; (ii) the estimated log-odds-ratio $\log \hat{p} - \log(1-\hat{p})$; (iii) the transformed estimator $\hat{\gamma} = 2 \arcsin(\hat{p}^{1/2})$. In particular, for the latter, show $\sqrt{n}(\hat{\gamma} - \gamma) \rightarrow_d N(0, 1)$, with $\gamma = 2 \arcsin(p^{1/2})$.

you may check
Story #1

(c) Suppose $\hat{p}_1 = Y_1/n$ and $\hat{p}_2 = Y_2/n$ are two binomial estimates, with the same sample size n . Then $\sqrt{n}(\hat{p}_1 - p_1) \rightarrow_d Z_1$ and $\sqrt{n}(\hat{p}_2 - p_2) \rightarrow_d Z_2$, where $Z_j \sim N(0, p_j(1 - p_j))$. Find the approximate normal distribution of $\hat{p}_1/\hat{p}_2$, viewed as an estimator of p_1/p_2 . Modify arguments appropriately to find a good approximation to the variance of $\hat{p}_1/\hat{p}_2$, and its approximate normal distribution, also in the case of unequal sample sizes, say n_1 and n_2 .

(d) Suppose $Y_1, \dots, Y_n$ are independent from the geometric distribution with $\Pr(Y_i = y) = (1 - p)^{y-1}p$ for $y = 1, 2, \dots$. We learned in Ex. 1.25 that the mean and variance are $1/p$ and $(1 - p)/p^2$. Find first the limiting distribution of $\sqrt{n}(\bar{Y} - 1/p)$ and then that of $\sqrt{n}(\hat{p} - p)$, where $\hat{p} = 1/\bar{Y}$.

(e) Suppose (a, b) is a certain position on the map, where one only has estimates, say A_n and B_n , for its x- and y-coordinates. Assume these are independent, approximately unbiased, and approximate normal, after n measurements. We formalise a version of this as $\sqrt{n}(A_n - a) \rightarrow_d N_1$ and $\sqrt{n}(B_n - b) \rightarrow_d N_2$, the limit variables N_1, N_2 being independent and standard normal. Having observed A_n and B_n , explain how you can put up 90 percent confidence intervals for a and b separately. Construct also a 90 percent confidence circle for (a, b) .

(f) Let us pass from Cartesian to polar coordinates, letting

$$R_n = \|(A_n, B_n)\| = (A_n^2 + B_n^2)^{1/2} \quad \text{and} \quad \hat{\alpha}_n = \arctan(B_n/A_n),$$

seen as estimators of the length $r = \|(a, b)\|$ and angle $\alpha = \arctan(b/a)$. Find the limit distributions for $\sqrt{n}(R_n - r)$ and $\sqrt{n}(\hat{\alpha}_n - \alpha)$, and show that these are independent in the limit.

(g) Suppose one observes $(A_n, B_n) = (4.44, 2.22)$, with $n = 100$. Construct and display an approximate 90 percent confidence circle for (a, b) , and then approximate 90 percent confidence intervals for the length r and angle α . How can you construct confidence intervals for r and α jointly, say $I_{r,n}$ and $I_{\alpha,n}$, such that the probability that $(r \in I_{r,n}) \cap (\alpha \in I_{\alpha,n})$ converges to 0.90?

Ex. 2.51 *Limiting normality for multinomials.* Consider the multinomial setup of Ex. 1.5, with $(Y_1, \dots, Y_k)$ counting the number of events of type $1, \dots, k$ in n independent experiments, each time with probabilities $p = (p_1, \dots, p_k)^t$. Here we sort out the basic large-sample behaviour of the relative frequencies $\hat{p}_j = Y_j/n$.

(a) Show that these $\hat{p}_j = Y_j/n$ are consistent, with $\sqrt{n}(\hat{p}_j - p_j) \rightarrow_d N(0, p_j(1 - p_j))$. Show more generally that there is full joint convergence in distribution here; $X_n = \sqrt{n}(\hat{p} - p) \rightarrow_d Z \sim N_k(0, \Sigma)$, where Σ is the matrix with elements $\sigma_{j,\ell} = p_j\delta_{j,\ell} - p_jp_\ell$. It may be written $\Sigma = D - pp^t$ with D diagonal with elements p_j . Verify that this is consistent with $\sum_{j=1}^k Z_j = 0$.

(b) For $\gamma = g(p_1, \dots, p_k)$ any smooth function of the relative frequencies, with natural estimator $\hat{\gamma} = g(\hat{p}_1, \dots, \hat{p}_k)$, show that $\sqrt{n}(\hat{\gamma} - \gamma) \rightarrow_d N(0, \tau^2)$, with $\tau^2 = c^t \Sigma c = c^t D c - (c^t p)^2$, where $c = \partial g(p)/\partial p$. Check what this says, for the case of $\gamma = p_1 + \dots + p_k$.

(c) Consider (X, Y, Z) being trinomial (n, p, q, r) . With $\hat{p} = X/n$, $\hat{q} = Y/n$, $\hat{r} = Z/n$, find the limit distribution for $\hat{\gamma} = \hat{p}/(\hat{q}\hat{r})^{1/2}$, as well as for $\hat{\delta} = 2 \arcsin(\hat{p}^{1/2}) - 2 \arcsin(\hat{q}^{1/2})$.

Ex. 2.52 *Smooth functions of means.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from some distribution. For the below, we also point to Ex. 3.21, with joint normality for empirical

quantiles, after which methods of this exercise apply also to these; smooth functions of means and quantiles are approximately normal.

(a) Consider sample averages $\bar{g}_j = n^{-1} \sum_{i=1}^n g_j(Y_i)$, for functions $g_1, \dots, g_p$, with finite variances. With $\xi_j = E g_j(Y_i)$ the means, show from the CLT that $\sqrt{n}(\bar{g}_1 - \xi_1, \dots, \bar{g}_p - \xi_p)^t$ tends to a zero-mean multinormal $(X_1, \dots, X_p)$, with covariance matrix Σ , having elements $\sigma_{j,k} = \text{cov}\{g_j(Y_i), g_k(Y_i)\}$. Let then $A(x_1, \dots, x_p)$ be any smooth function of these means. With $\bar{A}_n = A(\bar{g}_1, \dots, \bar{g}_p)$, show that

$$\sqrt{n}\{\bar{A}_n - A(\xi_1, \dots, \xi_p)\} \rightarrow_d a_1 X_1 + \dots + a_p X_p \sim N(0, \tau^2),$$

with $\tau^2 = a^t \Sigma a$, in terms of $a_j = \partial A(\xi_1, \dots, \xi_p) / \partial \xi_j$.

(b) For an illustration, let $Y_1, \dots, Y_n$ be i.i.d. from the unit exponential distribution. Show that the skewness and kurtosis are skew = 2 and kurt = 6. Identify the multinormal limit of $\sqrt{n}(\bar{g}_1 - 1, \bar{g}_2 - 2, \bar{g}_3 - 6, \bar{g}_4 - 24)$, for $\bar{g}_j = n^{-1} \sum_{i=1}^n Y_i^j$. Find the limit distribution for $\hat{\sigma} = (\bar{g}_2 - \bar{g}_1^2)^{1/2}$. Then find the limit distribution of $\sqrt{n}(\text{skew}^* - 2)$ and $\sqrt{n}(\text{kurt}^* - 6)$, with the empirical skew* = $n^{-1} \sum_{i=1}^n W_i^3$ and kurt* = $n^{-1} \sum_{i=1}^n W_i^4 - 3$, with $W_i = (Y_i - \bar{g}_1) / \hat{\sigma}$. Find also the joint binormal limit for skew*, kurt*.

Ex. 2.53 Delta method calculus for the normal case. Let $Y_1, \dots, Y_n$ be i.i.d. from the normal $N(\xi, \sigma^2)$, with standard estimators $\hat{\xi} = \bar{Y}$ and $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. In various exercises in Ch. 1 we have worked with *exact finite-sample* calculus, for certain basic parameters, like the mean, variance, the quantile $\gamma_q = \xi + z_p \sigma$. Here we show how the delta method, starting with the basic limit distributions for the two parameters, can be used to put up large-sample normal approximations for *any* functions of the parameters, in cases where it might be too hard to carry out exact finite-sample calculus.

(a) Since the skewness and the kurtosis for the normal are zero, show that the general result from Ex. 2.47 implies that $(\sqrt{n}(\bar{Y} - \xi), \sqrt{n}(\hat{\sigma} - \sigma))^t$ tends to say $(A, B)^t$, with these being independent and zero-mean normals with variances σ^2 and $\frac{1}{2}\sigma^2$. Show this directly, from the normality assumptions, as opposed to deriving it as a special case of the general statement. Note also that the $\sqrt{n}(\bar{Y} - \xi) \sim N(0, \sigma^2)$ holds exactly, for each finite n .

(b) With $\alpha = g(\xi, \sigma)$, for any smooth function of the two parameters, the natural estimator is $\hat{\alpha} = g(\bar{Y}, \hat{\sigma})$. Show that

$$\sqrt{n}(\hat{\alpha} - \alpha) \rightarrow_d cA + dB \sim N(0, (c^2 + \frac{1}{2}d^2)\sigma^2),$$

where c and d are the partial derivatives of g , evaluated at the position (ξ, σ) . Show how this leads to construction of confidence intervals for the α parameter.

(c) Consider the probability $p = \Pr(Y \geq y_0) = 1 - \Phi((y_0 - \mu)/\sigma)$, for some given threshold y_0 , and the associated estimator $\hat{p} = 1 - \Phi((y_0 - \bar{Y})/\hat{\sigma})$. Find the limit distribution of $\sqrt{n}(\hat{p} - p)$, and use this to put up a confidence interval for p , with coverage level converging to 0.90. Compare with the simpler estimator $p^* = n^{-1} \sum_{i=1}^n I(Y_i \geq y_0)$, the binomial proportion, which bypasses the normal assumption.

(d) Then consider the parameter $\kappa = \xi/\sigma$, the normalised mean (so its value is unchanged when one passes from say millimetres to metres). For $\hat{\kappa} = \bar{Y}/\hat{\sigma}$, show that $\sqrt{n}(\hat{\kappa} - \kappa) \rightarrow_d N(0, 1 + \frac{1}{2}\kappa^2)$, and construct an approximate 90 percent confidence interval. For this particular parameter, check also Ex. 3.38 for an exact method. Then work with the particular function $g(\kappa) = \sqrt{2} \log(\kappa + (2 + \kappa^2)^{1/2})$. Show that it stabilises the variance: $\sqrt{n}\{g(\hat{\kappa}) - g(\kappa)\}$ tends to $N(0, 1)$. xx

(e) Assume ξ is positive, so that the so-called coefficient of variation $\gamma = \sigma/\xi$ has a natural interpretation. With $\hat{\gamma} = \hat{\sigma}/\hat{\xi}$ the plug-in estimator, show that $\sqrt{n}(\hat{\gamma} - \gamma)$ tends to a normal with variance $\tau^2 = \frac{1}{2}\gamma^2(1 + 2\gamma^2)$. Find also a variance stabilising transformation, from γ to $\gamma^* = h(\gamma)$, with the property that $\sqrt{n}(\hat{\gamma}^* - \gamma^*) \rightarrow_d N(0, 1)$: one such is $h(\gamma) = \log \gamma - \log(1 + (1 + 2\gamma^2)^{1/2})$. Explain how confidence intervals can be set for γ via this.

Ex. 2.54 *Estimating mean and standard deviation outside normality.* Let $Y_1, \dots, Y_n$ be i.i.d. from a distribution with finite fourth moment, and consider the usual mean $\bar{Y}_n$ and empirical standard deviation S_n . Under normality we have precise finite-sample results regarding their distributions, see Ex. 1.49, but here we investigate behaviour outside normality.

(a) Let as on previous occasions γ_3 and γ_4 be the skewness and kurtosis of the distribution. Use the delta method, with previous results from Ex. 2.47, to show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y}_n - \xi) \\ \sqrt{n}(\hat{\sigma}_n - \sigma) \end{pmatrix} \rightarrow_d N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3 & \frac{1}{2} + \frac{1}{4}\gamma_4 \end{pmatrix}\right).$$

Explain that this implies $\Pr(|\bar{Y}_n - \xi| \leq 1.96\hat{\sigma}) \rightarrow 0.95$, even outside normality; the normality-based standard recipes regarding inference for the mean parameter is hence not much hampered by that modelling assumption.

(b) The situation is different when it comes to inference for the standard deviation, however. Show that $\sqrt{n}(\hat{\sigma} - \sigma)/\hat{\kappa} \rightarrow_d N(0, 1)$, if $\hat{\kappa}$ is a consistent estimator for $\kappa = (\frac{1}{2} + \frac{1}{4}\gamma_4)^{1/2}$. Construct indeed such an estimator; see also Ex. 3.12. For an application of this, with sample size up to a million, see Story #93.

Ex. 2.55 *The empirical correlation coefficient under binormality.* For i.i.d. data pairs (X_i, Y_i) , the classical empirical correlation coefficient is

$$R_n = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\{\sum_{i=1}^n (X_i - \bar{X})^2\}^{1/2} \{\sum_{i=1}^n (Y_i - \bar{Y})^2\}^{1/2}} = n^{-1} \sum_{i=1}^n \frac{X_i - \bar{X}}{\hat{\sigma}_1} \frac{Y_i - \bar{Y}}{\hat{\sigma}_2}, \quad (2.11)$$

with $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ the empirical variances $n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and $n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Here we find the limit distribution of R_n under binormality.

(a) Assume first that the (X_i, Y_i) pairs are from a zero-mean binormal with variances 1 and correlation $\rho \in (-1, 1)$; see Ex. 1.43. Use results from Ex. 1.44, including $Y_i | X_i \sim N(\rho X_i, 1 - \rho^2)$, to derive expressions for $E X_i^2 Y_i^2$, $E X_i^3 Y_i$, $E X_i Y_i^3$. Use these to show that

$$\Sigma = \text{Var} \begin{pmatrix} X_i^2 \\ Y_i^2 \\ X_i Y_i \end{pmatrix} = \begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1 + \rho^2 \end{pmatrix}.$$

Use the CLT to argue that

$$\begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} = \sqrt{n} \begin{pmatrix} n^{-1} \sum_{i=1}^n X_i^2 - 1 \\ n^{-1} \sum_{i=1}^n Y_i^2 - 1 \\ n^{-1} \sum_{i=1}^n X_i Y_i - \rho \end{pmatrix} \rightarrow_d \begin{pmatrix} A \\ B \\ C \end{pmatrix} \sim N_3(0, \Sigma).$$

With $R_{n,0} = C_n/(A_n B_n)^{1/2}$, use the delta method to show that $\sqrt{n}(R_{n,0} - \rho) \rightarrow_d Z = -\frac{1}{2}\rho A - \frac{1}{2}\rho B + C$, and that in fact $Z \sim N(0, (1 - \rho^2)^2)$.

(b) Then generalise to the situation where the (X_i, Y_i) pairs are i.i.d. from a zero-mean binormal, with standard deviations σ_1, σ_2 and correlation ρ . Show that we still have $\sqrt{n}(R_{n,0} - \rho) \rightarrow_d N(0, (1 - \rho^2)^2)$.

(c) Then go one step further, to the full five-parameter binormal situation, with unknown means ξ_1, ξ_2 , standard deviations σ_1, σ_2 , and correlation ρ . Argue first that we must have $\sqrt{n}(R_{n,1} - \rho) \rightarrow_d N(0, (1 - \rho^2)^2)$, where $R_{n,1}$ is as in (2.11) but using ξ_1, ξ_2 instead of $(\bar{X}, \bar{Y})$. Then, finally, show what we really are after, that $\sqrt{n}(R_n - \rho)$ must have the same limit distribution.

(d) Explain that if $h(\rho)$ is smooth, then $\sqrt{n}\{h(R_n) - h(\rho)\} \rightarrow_d h'(\rho)(1 - \rho^2)N(0, 1)$. Show that with the particular choice $\zeta = \frac{1}{2} \log\{(1 + \rho)/(1 - \rho)\}$ the variance is being stabilised, and $\sqrt{n}(\hat{\zeta} - \zeta) \rightarrow_d N(0, 1)$, where $\hat{\zeta} = \frac{1}{2} \log\{(1 + R_n)/(1 - R_n)\}$. This is called Fisher's zeta. Show that $\hat{\zeta} \pm 1.645/\sqrt{n}$ becomes an approximate 90 percent confidence interval for ζ , and transform this to an approximate 90 percent confidence interval for ρ . Show in fact that this inverse transformation is $\rho = \{\exp(2\zeta) - 1\}/\{\exp(2\zeta) + 1\}$.

variance
stabilising
transformation

Fisher's zeta,
with inverse

Ex. 2.56 *The empirical correlation coefficient, general case.* Here we use some of the arguments of Ex. 2.55 to find the limit distribution of the empirical correlation R_n of (2.11) also outside binormality. Assume $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. pairs from a distribution with means ξ_1, ξ_2 , standard deviations σ_1, σ_2 , and correlation ρ . Write $a_{j,k} = E U_i^j V_i^k$ for cross moments of the standardised $U_i = (X_i - \xi_1)/\sigma_1$ and $V_i = (Y_i - \xi_2)/\sigma_2$, where it is assumed that fourth order moments $a_{4,0}$ and $a_{0,4}$ are finite.

(a) Show that $\sqrt{n}(R_n - R_{n,0}) \rightarrow_{pr} 0$, where $R_{n,0}$ is as R_n , but using the real ξ_1, ξ_2 instead of their estimators $\bar{X}, \bar{Y}$. Show also that the distribution of R_n and $R_{n,0}$ must depend on ρ but not on $\xi_1, \xi_2, \sigma_1, \sigma_2$. We may hence carry out our large-sample investigation with the standardised (U_i, V_i) rather than the (X_i, Y_i) . Work with

$$\begin{pmatrix} A_n \\ B_n \\ C_n \end{pmatrix} = \begin{pmatrix} \sqrt{n}(n^{-1} \sum_{i=1}^n U_i^2 - 1) \\ \sqrt{n}(n^{-1} \sum_{i=1}^n V_i^2 - 1) \\ \sqrt{n}(n^{-1} \sum_{i=1}^n U_i V_i - \rho) \end{pmatrix},$$

and show that $(A_n, B_n, C_n)^t \rightarrow_d (A, B, C)^t \sim N_3(0, \Sigma)$, for the variance matrix Σ of $(U_i^2, V_i^2, U_i V_i)^t$. Spell out the elements of this matrix, using the $a_{j,k}$. Check that this agrees with the Σ of Ex. 2.55 under binormality. Explain that under independence, we have $\sqrt{n}R_n \rightarrow_d N(0, 1)$, and that cases where $|R_n| > 1.96/\sqrt{n}$ should make us suspect that there is non-zero correlation (without necessarily going for binormality).

(b) Then show that $\sqrt{n}(R_n - \rho) \rightarrow_d Z = -\frac{1}{2}\rho A - \frac{1}{2}\rho B + C$, and give an expression for the limit distribution variance τ^2 . Explain how τ may be estimated from the data, and how this leads to confidence intervals of the type $R_n \pm 1.96 \hat{\tau}/\sqrt{n}$ for ρ .

(c) For a concrete illustration, consider the joint density $f(x, y) = 1 + a(x - \frac{1}{2})(y - \frac{1}{2})$ for (x, y) in the unit square. Find the allowed parameter range for a , and a formula for the correlation coefficient ρ in terms of a . Then apply the above to find the limit distribution of $\sqrt{n}(R_n - \rho)$.

The strong law of large numbers

Ex. 2.57 *The Strong Law of Large Numbers: the basics.* Suppose $X_1, X_2, \dots$ are i.i.d. from a distribution with finite $E|X_i|$. Then the mean $\xi = E X_i$ exists, and we are aiming to

prove the strong LLN of (2.3), that the event

$$A = \{\bar{X}_n \rightarrow \xi\} = \cap_{\varepsilon > 0} \cup_{n_0 \geq 1} \cap_{n \geq n_0} \{|\bar{X}_n| \leq \varepsilon\}$$

has probability equal to one hundred percent. We may for simplicity and without loss of generality take $\xi = 0$ below.

(a) Show that A is the same as $\cap_{N \geq 1} \cup_{n_0 \geq 1} \cap_{n \geq n_0} \{|\bar{X}_n| \leq 1/N\}$, and deduce in particular from this that A is actually measurable – so it does make well-defined sense to work with its probability.

(b) Show that if $\Pr(A_N) = 1$ for all N , then $\Pr(\cap_{N \geq 1} A_N) = 1$ – if you’re fully certain about a countable number of events, then you’re also fully certain about all of them, jointly. This is actually not true with a bigger index set: if $X \sim N(0, 1)$, then you’re 100 percent sure that $B_x = \{X \text{ is not } x\}$ takes place, for each single x , but from this does it *not* follow that you should be sure about $\cap_{\text{all } x} B_x$. Explain why.

(c) Show that $\Pr(A) = 1$ if and only if $\Pr(B_{n_0}) \rightarrow 0$, for each $\varepsilon > 0$, where $B_{n_0} = \cup_{n \geq n_0} \{|\bar{X}_n| \geq \varepsilon\}$. In words: for a given ε , the probability should be very low that there is *any* $n \geq n_0$ with $|\bar{X}_n| \geq \varepsilon$.

(d) A simple bound is of course $\Pr(B_{n_0}) \leq \sum_{n \geq n_0} \Pr\{|\bar{X}_n| \geq \varepsilon\}$, so it suffices to show, if possible, under appropriate conditions, that $\sum_{n \geq 1} \Pr\{|\bar{X}_n| \geq \varepsilon\}$ is a convergent series. With finite variance σ^2 , explain that the classic simple Chebyshov bound, see Ex. 2.10, does *not* solve any problem here.

(e) Show, however, that if the fourth moment is finite, then

$$\Pr\{|\bar{X}_n| \geq \varepsilon\} \leq \frac{1}{\varepsilon^4} \mathbb{E}|\bar{X}_n|^4 \leq \frac{c}{\varepsilon^4} \frac{1}{n^2},$$

for a suitable c ; see also Ex. 2.10. So under this condition, which is moderately hard, we’ve proven the strong LLN.

(f) One may squeeze more out of the chain of arguments here, which we now indicate, without full details. Assume $\mathbb{E}|X_i|^r$ is finite, for some $r > 2$, like $r = 2.02$. Then one may show, via arguments in von Bahr (1965), that the sequence $\mathbb{E}|\sqrt{n}\bar{X}_n|^r$ is bounded. This leads to the bound

$$\Pr\{|\bar{X}_n| \geq \varepsilon\} \leq \frac{\mathbb{E}|\sqrt{n}\bar{X}_n|^r}{(\sqrt{n}\varepsilon)^r},$$

and these form a convergent series. We have hence proven (modulo the von Bahr thing) that the strong LLN holds for finite $\mathbb{E}|X_i|^{2+\varepsilon}$, an improvement over the finite $\mathbb{E}|X_i|^4$ condition. – To get further, trimming away on the conditions until we are at the Kolmogorovian position of only requiring finite mean, we need more technicalities; see the following Ex. 2.58.

Ex. 2.58 *The Strong Law of Large Numbers: nitty-gritty details.* This exercise goes through the required extra technical details, along with a few intermediate lemmas, to secure a full proof of the full LLN theorem: as long as $\mathbb{E}|X_i|$ is finite, the infinite sequence of sample means $\bar{X}_n$ will with probability equal to one hundred percent converge to $\xi = \mathbb{E}X_i$.

(a) We start with Kolmogorov’s inequality: Consider independent zero-mean variables $X_1, \dots, X_n$ with variances $\sigma_1^2, \dots, \sigma_n^2$, and with partial sums $S_i = X_1 + \dots + X_i$. Then

$$\Pr\{\max_{i \leq n} |S_i| \geq \varepsilon\} \leq \frac{\text{Var } S_n}{\varepsilon^2} = \frac{1}{\varepsilon^2} \sum_{i=1}^n \sigma_i^2.$$

Kolmogorov’s
inequality

Note that this is a much stronger result than the special case of caring only about $|S_n|$, with $\Pr\{|S_n| \geq \varepsilon\} \leq \text{Var } S_n / \varepsilon^2$, which is the Chebyshev inequality. To prove it, work with the disjoint decomposition

$$A_i = \{|S_1| < \varepsilon, \dots, |S_{i-1}| < \varepsilon, |S_i| \geq \varepsilon\} \quad \text{and} \quad A = \cup_{i=1}^n A_i = \{\max_{i \leq n} |S_i| \geq \varepsilon\}.$$

Show that $E S_n^2 \geq E S_n^2 I(A) = \sum_{i=1}^n E S_n^2 I(A_i)$, that

$$E S_n^2 I(A_i) = E (S_i + S_n - S_i)^2 I(A_i) \geq \varepsilon^2 \Pr(A_i),$$

and that this leads to the inequality asked for.

(b) Consider a sequence of independent $X_1, X_2, \dots$ with means zero and variances $\sigma_1^2, \sigma_2^2, \dots$. Show that if $\sum_{i=1}^{\infty} \sigma_i^2$ is convergent, then $\sum_{i=1}^{\infty} X_i$ is convergent with probability 1. – It suffices to show that the sequence of partial sums $S_n = X_1 + \dots + X_n$ is Cauchy with probability 1. Show that this is the same as

$$\lim_{n \rightarrow \infty} \Pr[\cup_{i,j \geq n} \{|S_i - S_j| \geq \varepsilon\}] = 0 \quad \text{for each } \varepsilon > 0.$$

Use the Kolmogorov inequality to show this.

(c) A quick example to illustrate this result is as follows. Consider $X = X_1/10 + X_2/100 + X_3/1000 + \dots$, a random number in the unit interval, with the X_i independent, and with no further assumptions. Show that X exists with probability 1.

(d) Prove that if $\sum_{i=1}^{\infty} a_i/i$ converges, then $\bar{a}_n = (1/n) \sum_{i=1}^n a_i \rightarrow 0$. To show this, consider $b_n = \sum_{i=1}^n a_i/i$, so that $b_n \rightarrow b$ for some b . Show $a_n = n(b_n - b_{n-1})$, valid also for $n = 1$ if we set $b_0 = 0$, and which leads to $\sum_{i=1}^n a_i = nb_n - b_0 - b_1 - \dots - b_{n-1}$.

(e) From the above, deduce that if $X_1, X_2, \dots$ are independent with means $\xi_1, \xi_2, \dots$ and variances $\sigma_1^2, \sigma_2^2, \dots$, and $\sum_{i=1}^{\infty} \sigma_i^2/i^2$ converges, then $\bar{X}_n - \bar{\xi}_n \rightarrow_{\text{a.s.}} 0$. Here $\bar{\xi}_n = (1/n) \sum_{i=1}^n \xi_i$.

(f) Use the above to show that if $X_1, X_2, \dots$ are independent with zero means, and all variances are bounded, then indeed $\bar{X}_n \rightarrow_{\text{a.s.}} 0$. Note that this is a solid generalisation of what we managed to show in (xx calibrate xx) – first, the distributions are allowed to be different (not identical); second, we have landed at a.s. convergence with the mild assumption of finite and bounded variances, whereas we there needed the harsher conditions of finite fourth moments.

(g) We're close to the Pole. For i.i.d. zero mean variables $X_1, X_2, \dots$, split them up with the little trick

$$X_i = Y_i + Z_i, \quad \text{with} \quad Y_i = X_i I(|X_i| < i), \quad Z_i = X_i I(|X_i| \geq i).$$

We have $\bar{X}_n = \bar{Y}_n + \bar{Z}_n$, so it suffices to demonstrate that $\bar{Y}_n \rightarrow_{\text{a.s.}} 0$ and $\bar{Z}_n \rightarrow_{\text{a.s.}} 0$ (since an intersection of two sure events is sure). Use the Borel–Cantelli lemma, in concert with $E|X_i| = \int_0^{\infty} \Pr(|X_i| \geq x) dx$, to show that only finitely many Z_i are non-zero. Then use previous results to demonstrate $\bar{Y}_n - \bar{\xi}_n \rightarrow_{\text{a.s.}} 0$ and $\bar{\xi}_n \rightarrow 0$, where $\bar{\xi}_n$ is the average of $\xi_i = E Y_i$.

(h) So we've managed to prove the Strong LLN; good. Attempt also to prove the interesting converse that if $E|X_i| = \infty$, then the sequence of sample means is pretty erratic indeed:

$$\Pr\{\limsup_{n \rightarrow \infty} \bar{X}_n = \infty\} = 1, \quad \Pr\{\liminf_{n \rightarrow \infty} \bar{X}_n = -\infty\} = 1.$$

Simulate a million realisations from the density $f(x) = 1/x^2$, for $x \geq 1$, in your nearest computer, display the sequence of $\bar{X}_n$ on your screen, and comment.

Ex. 2.59 *Yes, we converge with probability 1.* We've proven that the sequence of empirical means converges almost surely to the population mean, under the sole condition that this mean is finite. This half-automatically secures almost sure convergence of various other natural quantities, almost without further efforts.

(a) Suppose $X_1, X_2, \dots$ are i.i.d. with finite variance σ^2 . Show that the classical empirical standard deviation $\hat{\sigma} = \{\sum_{i=1}^n (X_i - \bar{X}_n)^2 / (n-1)\}^{1/2}$ converges a.s. to σ . Note again that nothing more is required than a finite second moment.

(b) Suppose the third moment is finite, such that the skewness $\gamma_3 = E\{(X - \xi)/\sigma\}^3$ is finite. Show that $\hat{\gamma}_{3,n} = (1/n) \sum_{i=1}^n (X_i - \bar{X}_n)^3 / \hat{\sigma}^3$ is strongly consistent for γ_3 .

(c) Then suppose the fourth moment is finite, such that the kurtosis $\gamma_4 = E\{(X - \xi)/\sigma\}^4 - 3$ is finite. Construct a strongly consistent estimator for this kurtosis.

(d) Assume that $(X_1, Y_1), (X_2, Y_2), \dots$ is an i.i.d. sequence of random pairs, with finite variances, and define the population correlation coefficient in the usual fashion, as $\rho = \text{cov}(X, Y) / (\sigma_1 \sigma_2)$. Show that the usual empirical correlation coefficient R_n of (2.11) converges with probability one to ρ .

(e) Formulate and prove a suitable statement regarding almost sure convergence of smooth functions of means.

(f) Let $X_1, X_2, \dots$ be an i.i.d. sequence of nonnegative variables such that $E \log X_i = \xi$ is finite. Show that the harmonic means $H_n = (X_1 \cdots X_n)^{1/n}$ converge with probability 1 to $\exp(\xi)$.

Ex. 2.60 *Glivenko–Cantelli theorem.* For i.i.d. observations $Y_1, \dots, Y_n$, we form the empirical c.d.f., as in Ex. 2.11, with $F_n(t) = n^{-1} \sum_{i=1}^n I(Y_i \leq t)$. We have seen that since $F_n(t)$ is just a binomial ratio, $F_n(t) \rightarrow_{\text{a.s.}} F(t)$, for each t . It is a remarkable fact that this convergence also takes place uniformly, with probability 1. This is the Glivenko–Cantelli theorem: with $D_n = \max_t |F_n(t) - F(t)|$, the max taken over all t in the domain in question, we have $\Pr(D_n \rightarrow 0) = 1$. This means that regardless of any strange or complicated aspects of the distribution F , with enough data one will be able to learn these. See also Ex. 3.9 and 10.27 for more information, regarding the speed with which $D_n \rightarrow 0$.

Glivenko–
Cantelli

(a) Choose $t_1 < \dots < t_m$, creating a finite number of cells $[t_j, t_{j+1})$, where we take $t_0 = -\infty$ and $t_{m+1} = \infty$. With $A_{m,j}$ the event that $F_n(t_j) \rightarrow F(t_j)$, argue that $\Pr(\cap_{j=1}^m A_{m,j}) = 1$.

(b) Consider any t in the cell $[t_j, t_{j+1})$. Writing $D_n(t) = F_n(t) - F(t)$, use monotonicity of F_n and F to show that

$$D_n(t_j) - \{F(t_{j+1}) - F(t_j)\} \leq F_n(t) - F(t) \leq D_n(t_{j+1}) + F(t_{j+1}) - F(t_j).$$

Deduce that

$$\max_{t_j \leq t < t_{j+1}} |D_n(t)| \leq B_m + C_m,$$

where $B_m = \max_{1 \leq j \leq m} |D_n(t_j)|$ and $C_m = \max_{1 \leq j \leq m} \{F(t_{j+1}) - F(t_j)\}$.

(c) Show that $\Pr(\limsup D_n \leq C_m) = 1$.

(d) For each $\varepsilon > 0$, show that a partition into cells can be arranged, with high m if required, so that $C_m \leq \varepsilon$. Conclude that $\Pr(D_n \rightarrow 0) = 1$.

(e) Choose some moderately complicated normal mixture, of the type $f = \sum_{j=1}^k p_j N(\mu_j, \sigma_j^2)$. Then simulate a high number n of data from this distribution, and read off $D_n = \max_t |F_n(t) - F(t)|$. Check out how high n needs to be to have $D_n \leq 0.01$, say, with high probability, in a few situations.

Leftovers, more material, more limits (and some are nonnormal)

Ex. 2.61 *Approximate variances and the delta method.* We have built a well-working apparatus around the delta method and seen various applications. Statements reached are in terms of precise limit distributions, though without going into the quality of the resulting approximations. The present exercise goes into the details of variances and covariances for functions of sample averages.

(a) Suppose $X_1, \dots, X_n$ are i.i.d. with mean zero, variance σ^2 , and finite skewness $\gamma_3 = E(X_i/\sigma)^3$ and kurtosis $\gamma_4 = E(X_i/\sigma)^4 - 3$. With $\bar{X}_n$ as usual being the average, show then that

$$\begin{aligned} E \bar{X}_n^2 &= \sigma^2/n, \\ E \bar{X}_n^3 &= (\sigma^3/n^2)\gamma_3, \\ E \bar{X}_n^4 &= (\sigma^4/n^2)(3 + \gamma_4/n), \\ \text{Var } \bar{X}_n^2 &= (\sigma^4/n^2)(2 + \gamma_4/n). \end{aligned}$$

(b) Let then $Y_1, \dots, Y_n$ be i.i.d., with finite mean ξ , standard deviation σ , skewness γ_3 , kurtosis γ_4 . With $\bar{Y}_n$ the sample average, consider then the variable

$$Z_n = a_0 + a_1(\bar{Y}_n - \xi) + \frac{1}{2}a_2(\bar{Y}_n - \xi)^2.$$

Show that $E Z_n = a_0 + \frac{1}{2}a_2\sigma^2/n$, and that

$$\text{Var } Z_n = a_1^2\sigma^2/n + (1/n^2)\{\frac{1}{4}a_2^2\sigma^4(2 + \gamma_4/n) + a_1a_2\sigma^3\gamma_3\}.$$

(c) Consider any smooth function $Z_n = g(\bar{Y}_n)$. Since $\bar{Y}_n$ is close to ξ with high probability, it makes sense to carry out a Taylor expansion,

$$Z_n = g(\xi) + g'(\xi)(\bar{Y}_n - \xi) + \frac{1}{2}g''(\xi)(\bar{Y}_n - \xi)^2 + \delta_n,$$

where δ_n is a smaller-sized remainder term; you may prove that $n^{3/2}\delta_n$ is bounded in probability, provided g has three derivatives in a neighbourhood around ξ . Show from the above that

$$\text{Var } Z_n = g'(\xi)^2\sigma^2/n + (1/n^2)\{\frac{1}{4}g''(\xi)^2\sigma^4(2 + \gamma_4/n) + g'(\xi)g''(\xi)\sigma^3\gamma_3\} + o(1/n^2).$$

There is hence a clear leading $O(1/n)$ term for the variance, with other terms being of size $O(1/n^2)$. Explain how this relates to the basics of the delta method from Ex. 2.49.

(d) Suppose A is a random variable with mean a and finite variance, and that $g(y)$ is smooth in a neighbourhood around a . Use the Taylor approximation

$$g(y) = g(a) + g'(a)(y - a) + \frac{1}{2}g''(a)(y - a)^2 + O(|y - a|^3),$$

valid for y close to a , to show that

$$E g(A) \doteq g(a) + \frac{1}{2}g''(a)\text{Var } Y, \quad \text{Var } g(Y) \doteq g'(a)^2 \text{Var } Y,$$

and indicate the sizes of the error terms involved.

Ex. 2.62 *How many records are set?* Consider i.i.d. observations $X_1, X_2, \dots$ from some continuous distribution on the line. Define $R_n = 1$ if X_n is bigger than all previous datapoints, i.e. a new record has been set.

(a) Explain first that the number of records set, in the course of the first n occasions, is $Z_n = \sum_{i=1}^n R_i$. Show that $\Pr(R_n = 1) = 1/n$, so that $E Z_n = H_n = 1 + 1/2 + \dots + 1/n$, the partial sum of the harmonic series. As we have noted earlier, $H_n - \log n \rightarrow \gamma_e = 0.5772\dots$, the Euler constant, so the number of records set is approximately $\log n + \gamma_e$.

(b) Show that $R_1, \dots, R_n$ are independent, and deduce from this that $\text{Var } Z_n = H_n - \sum_{i=1}^n 1/i^2 \doteq \log n + \gamma_e - \pi^2/6$. Use the Lindeberg theorem to show that $(Z_n - H_n)/H_n^{1/2} \rightarrow_d N(0, 1)$. Show also that $(Z_n - \log n)/(\log n)^{1/2} \rightarrow_d N(0, 1)$, which gives a simpler but somewhat more crude approximation to Z_n probabilities. Use these normal limits to give a prediction band, with $\Pr(a_n \leq Z_n \leq b_n) \rightarrow 0.95$.

(c) Now consider $A_n = Z_{2n} - Z_n$, the number of records set during $n, n+1, \dots, 2n$. Show that A_n tends to the Poisson with mean $\log 2$. Generalise.

Ex. 2.63 *Limiting normality of rank sums statistics.* (xx with a brief pointer to Story #75. need to round off the conditioning thing. xx) In a population of n individuals, followed on some continuous scale, a subgroup of interest, of size m , has ranks $X_1, \dots, X_m$. These form a randomly selected subset of size m from $\{1, \dots, n\}$, with all such $\binom{n}{m}$ subsets equally likely. The rank sum $Z_n = X_1 + \dots + X_m$ is the Wilcoxon statistic.

(a) Explain that one may write $Z_n = \sum_{i=1}^n iJ_i$, where the 0-1 variables J_i are such that precisely m of them are 1, and with all $\binom{n}{m}$ subsets of such 1s being equally likely. Find $E J_i$, $\text{Var } J_i$, $\text{cov}(J_i, J_j)$ for $j \neq i$. Writing $p = m/n$ for the sample ratio, show using either of the representations $\sum_{i=1}^m X_i$ or $\sum_{i=1}^n iJ_i$ that

$$E Z_n = \frac{1}{2}m(n+1) \doteq \frac{1}{2}n^2p, \quad \text{Var } Z_n = (1/12)(n+1)m(n-m) \doteq (1/12)n^3p(1-p).$$

(b) We aim indeed at showing limiting normality of Z_n here, with both n and m becoming larger, with $m/n \rightarrow p$. Explain that this must mean

$$(Z_n - \frac{1}{2}n^2p)/n^{3/2} \rightarrow_d N(0, (1/12)p(1-p)).$$

We cannot use CLT or Lindeberg for studying Z_n , since the X_i are dependent, as are the J_i . Consider however a different parallel setup, involving independent Bernoulli variables $J_1^*, \dots, J_n^*$ with $\Pr(J_i^* = 1) = p = m/n$. Explain that the distribution of Z_n is the same as the distribution of $Z_n^* = \sum_{i=1}^n iJ_i^*$ given $\sum_{i=1}^n J_i^* = m$. Show now that

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = \begin{pmatrix} (1/\sqrt{n}) \sum_{i=1}^n (i/n)(J_i^* - p) \\ (1/\sqrt{n}) \sum_{i=1}^n (J_i^* - p) \end{pmatrix} \rightarrow_d \begin{pmatrix} A \\ B \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/3 & 1/2 \\ 1/2 & 1 \end{pmatrix}\right).$$

(c) Find the distribution of $A|(B=b)$, and show in particular that $A|(B=0) \sim N(0, 1/12)$. This gives a clear limiting normality statement for Z_n^* , conditional on $\sum_{i=1}^n J_i^* = m$, and nicely solves our Wilcoxon problem; show that it corresponds precisely to the $(Z_n - \frac{1}{2}n^2p)/n^{3/2}$ limiting statement above.

Ex. 2.64 *Nonnormal limits.* (xx polish, and calibrate with Story #100. xx) Normally limits are normal, but not always. Here we shall indeed work with variables with mean zero and variance one, where the sample averages have nonnormal limits. The basic construction is as follows. Let $U_1, U_2, \dots$ be i.i.d., with mean zero and variance one, and with m.g.f. $M_0(s) = E \exp(sU_i)$ finite in a neighbourhood around zero; in particular,

all moments for the U_i are finite. Let independently of these $J_1, J_2, \dots$ be independent Bernoulli variables with $\Pr(J_i = 1) = 1/i$, $\Pr(J_i = 0) = 1 - 1/i$. Then form

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n J_i \sqrt{i} U_i = \sum_{i=1}^n J_i \sqrt{i/n} U_i.$$

A picture to have in mind is that most of the terms will be zero, with non-zero contributions becoming both more rare and more big as time proceeds.

(a) Show that there will with probability one be infinitely many $J_i = 1$, i.e. non-zero terms in the Z_n sum as n grows.

(b) Show that the terms $J_i \sqrt{i} U_i$ have mean zero and variance one; hence also the normalised sample average Z_n has mean zero and variance one. Find also an expression for the kurtosis $\kappa_n = E Z_n^4 - 3$ of Z_n , and show that $\kappa_n \rightarrow \frac{1}{2} a_4$, where $a_4 = E U_i^4$. Compare this to what we are ‘used to’ from the Lindeberg theorem.

(c) We already know that if Z_n has a limit distribution, it can’t be normal. Working with the m.g.f., show that

$$M_n(t) = E \exp(t Z_n) = \prod_{i=1}^n \left[1 + \frac{1}{i} \{ M_0(t \sqrt{i/n}) - 1 \} \right],$$

for all t around zero for which $M_0(t)$ is finite.

(d) Show that

$$\prod_{i=1}^n \left[1 + \frac{1}{i} \{ M_0(t \sqrt{i/n}) - 1 \} \right] \rightarrow \exp \left\{ \int_0^1 \frac{M_0(t \sqrt{x}) - 1}{x} dx \right\}.$$

Work first with Special Case One, where we let U_i have the simple symmetric two-point distribution $\Pr(U_i = 1) = \Pr(U_i = -1) = \frac{1}{2}$. Find the limiting kurtosis for Z_n in this case. Show that $M_0(s) = \frac{1}{2} e^s + \frac{1}{2} e^{-s} = 1 + (1/2!)s^2 + (1/4!)s^4 + \dots$, and use this to find an infinite-sum expression for the limit of $M_n(t)$. Have you now proved that Z_n has a limit distribution?

(e) Then work with Special Case Two, where the U_i have a double exponential distribution, of the form $f(u) = \frac{1}{2} \sqrt{2} \exp(-\sqrt{2}|u|)$ on the real line (the $\sqrt{2}$ factor is there to ensure variance one). Find the m.g.f. $M_0(s)$ for the U_i , and then the m.g.f. $M(t)$ for the limit distribution of Z_n .

(f) For most cases, regarding the distribution for the U_i , it is hard to learn the explicit distribution for Z_n (even in cases where there might be a clear distribution for its limit). For Special Case Two, however, find the explicit distribution for Z_n , for any given n .

CLTs for dependent random variables

Ex. 2.65 A CLT for 1-dependent variables. (xx decide later if these few should be pushed to Ch. 13. xx) Consider a stationary sequence $Y_1, Y_2, \dots$, with mean zero and variance one, being 1-dependent. Stationarity means $(Y_1, \dots, Y_r)$ having the same distribution as $(Y_{i+1}, \dots, Y_{i+r})$, for any i and block lengths r , and 1-dependence means that Y_i, Y_{i+1} may be dependent, but $Y_1, \dots, Y_i$ is independent of $Y_{i+2}, Y_{i+3}, \dots$. This exercise reaches a CLT for $\sum_{i=1}^n Y_i$, representing a genuine extension of the usual CLT and Lindeberg theorems from independence.

(a) Writing $\rho = \text{corr}(Y_i, Y_{i+1})$, show that $(1/k)\text{Var}(Y_1 + \cdots + Y_k) = 1 + 2(1 - 1/k)\rho$ which then goes to $1 + 2\rho$ for increasing k .

(b) For a given block length k , split $Y_1 + \cdots + Y_n$ into $[n/k]$ blocks, and write block j of these as $U_j + V_j$, with U_j as sum of $k - 1$ consecutive observations and V_j the last one of that block. Write then

$$Z_n = (1/\sqrt{n}) \sum_{i=1}^n Y_i = (1/\sqrt{n}) \left(\sum_{j=1}^{[n/k]} U_j + \sum_{j=1}^{[n/k]} V_j + E_n \right),$$

with E_n any extra left after the $k[n/k]$ variables captured in these first $[n/k]$ blocks.

(c) Explain why $U_1, U_2, \dots$ are independent, so that the usual CLT applies to these. Show that $(1/\sqrt{n}) \sum_{j=1}^{[n/k]} U_j \rightarrow_d N(0, \tau_k^2)$, with $\tau_k^2 = (1/k)\text{Var}(Y_1 + \cdots + Y_{k-1})$.

(d) Then use Ex. 2.30 to prove that $(1/\sqrt{n}) \sum_{i=1}^n Y_i \rightarrow_d N(0, 1 + 2\rho)$, i.e. a CLT for 1-dependent variables.

(e) Assume $X_1, X_2, \dots$ are i.i.d. with mean zero and variance one. Consider $Z_n = (1/\sqrt{n})(X_1 X_2 + X_2 X_3 + \cdots + X_{n-1} X_n)$. Show that $Z_n \rightarrow_d N(0, 1)$. Show also that

$$Z'_n = (1/\sqrt{n}) \sum_{i=1}^{n-1} (X_i - \bar{X}_n)(X_{i+1} - \bar{X}_n)$$

has the same limit distribution, where $\bar{X}_n$ as usual is the sample mean.

Ex. 2.66 A CLT for m -dependent variables. In natural generalisation of Ex. 2.65, consider a stationary m -dependent sequence $Y_1, Y_2, \dots$, with mean zero and variance σ^2 . There is accordingly potential dependence among $Y_1, \dots, Y_m$, but for any i , $(Y_1, \dots, Y_i)$ is independent of $(Y_{i+m+1}, \dots, Y_n)$.

(a) Writing $\text{cov}(Y_i, Y_j) = \sigma^2 \rho(|j - i|)$, with the autocorrelation function $\rho(\cdot)$, show first that in general terms,

$$(1/n) \text{Var}\left(\sum_{i=1}^n Y_i\right) = \sigma^2 \left\{ 1 + 2 \sum_{j=1}^n (1 - j/n) \rho(j) \right\}.$$

Then show that for the case of m -dependence, for any $k \geq m$, we have $(1/k) \text{Var}(Y_1 + \cdots + Y_k) \rightarrow \sigma^2 \{1 + 2 \sum_{j=1}^m \rho(j)\}$,

(b) Extend arguments and techniques from Ex. 2.65 to show that $(1/\sqrt{n}) \sum_{i=1}^n Y_i$ tends to a zero-mean normal with variance $\sigma^2 \{1 + 2\rho(1) + \cdots + 2\rho(m)\}$.

(c) (xx acf for an i.i.d. sequence: $\sqrt{n}\hat{\rho}(j) \rightarrow_d N(0, 1)$, for each j . xx)

Ex. 2.67 Summing geometrically many terms. Suppose $Y_1, Y_2, \dots$ are i.i.d. with mean zero, variance σ^2 , and m.g.f. $M_0(t)$. With $\text{Pr}(N = n) = (1 - p)^{n-1}p$ for $n \geq 1$, i.e. a geometric, consider $Z_p = p^{1/2}(Y_1 + \cdots + Y_N)$, with the Y_i being independent of N .

(a) Show first that the generating function for N is $E s^N = ps / \{1 - (1 - p)s\}$ for $|s| < 1/(1 - p)$; see Ex. 1.38. Show that Z_p has variance σ^2 , and that its m.g.f. may be written

$$K_p(t) = E \exp(tZ_p) = pM_0(p^{1/2}t) / \{1 - (1 - p)M_0(p^{1/2}t)\}.$$

(b) Then use $M_0(t) = 1 + \frac{1}{2}\sigma^2 t^2 + o(|t|^2)$ to demonstrate that as $p \rightarrow 0$, with increasing number of terms $EN = 1/p$, we have $K_p(t) \rightarrow 1/(1 - \frac{1}{2}\sigma^2 t^2)$ for $|t| < \sqrt{2}/\sigma$. This shows that $Z_p \rightarrow_d L_\sigma$, the Laplace distribution with standard deviation σ , see Ex. 1.34.

(c) For the particular case of a sum of randomly many normal terms, let the Y_i be i.i.d. standard normal. Show what $Z_p | N \sim N(0, pN)$, and that $pN \rightarrow_d \text{Expo}(1)$ as $p \rightarrow 0$. Explain how this matches Ex. 1.34.(c).

(d) (xx one or two further illustrations, with Laplace limit of $p^{1/2}(Y_1 + \dots + Y_N)$. with $Y_i = Q_i - 1$, Poisson, we learn $p^{1/2}(V_N - N) \rightarrow_d L$, where $V_N \sim \text{Pois}(N)$. similarly with $p^{1/2}(W_N - N)$, where $W_N | N \sim \chi_N^2$, randomly many degrees of freedom. xx)

(e) (xx something to simulate, to illustrate the cusp behaviour at the centre. with the Y_i having mean ξ and variance 1, we have

$$p^{1/2}(Y_1 - \xi + \dots + Y_N - \xi) = p^{1/2}N(\bar{Y} - \xi) = (pN)^{1/2}N^{1/2}(\bar{Y} - \xi) \rightarrow_d L_1.$$

this gives different inference for ξ , and different predictions for $\bar{Y}$, than what we're used to from normal terrain. we're also in scale mixtures of normal terrain, with random variance tending to a unit exponential. we do like the $pN \rightarrow_d \text{Expo}(1)$, since it gives the cool cusp in the limit for Z_p , but other variations for $pN \rightarrow_d V$ are ok too. xx)

(f) (xx just ranting away a bit until things settle. xx) More generally, with $Y_1, Y_2, \dots$ being i.i.d. with zero mean and unit variance, consider

$$Z_p = p^{1/2}(Y_1 + \dots + Y_N) = (pN)^{1/2}N^{1/2}\bar{Y}_N,$$

with N having a distribution such that $pN \rightarrow_d V$, say, as $p \rightarrow 0$. From the CLT, $N^{1/2}\bar{Y}_N \rightarrow_d N(0, 1)$, so this amounts to a situation with a normal limit, but a random variance, as in $X | V \sim N(0, V)$ but V random. Here $E \exp(tX) = E \exp(\frac{1}{2}t^2V) = M_0(\frac{1}{2}t^2)$, where $M_0(u) = E \exp(uV)$ is the mgf for V . Also, X has density

$$\bar{f}(x) = \int \phi(x/v^{1/2})(1/v^{1/2}) dH(v),$$

with H the distribution of V . – The special case above amounts to $N \sim \text{geom}(1/p)$, where pN tends to the unit exponential, and where X gets the Laplace distribution. For another case, consider $N | \lambda \sim \text{Pois}(\lambda/p)$, and with λ having its own distribution with mean and variance λ_0 and τ_0^2 , say. Show that $E pN = \lambda_0$ and that $\text{Var } pN \rightarrow \tau_0^2$. Also consider the special case for this setup where $\lambda \sim \text{Gam}(a, b)$. From

$$E \{ \exp(-spN) | \lambda \} = \exp[-(\lambda/p)\{1 - \exp(-sp)\}]$$

deduce

$$E \exp(-spN) = \frac{1}{[1 + (1/b)(1/p)\{1 - \exp(-sp)\}]^a} \rightarrow \frac{1}{(1 + s/b)^a},$$

and that $pN \rightarrow_d V_{a,b}$, another $\text{Gam}(a, b)$. With the construction above, the normal scale mixture variable X has mgf $1/(1 - \frac{1}{2}t^2/b)^a$, and density

$$\bar{f}(x) = \int \phi(x/v^{1/2})(1/v^{1/2}) \frac{b^a}{\Gamma(a)} v^{a-1} \exp(-bv) dv.$$

This is a Laplace, for $a = 1$. (xx but different, and interesting, for other a . round this off. xx)

Notes and pointers

(xx to come. we point to certain famous things from the past: [Kolmogorov \(1933b\)](#), [Lindeberg \(1922\)](#), Borel and Cantelli. tail bounds. emil’s extension of the [Inlow \(2010\)](#) paper, from CLT to Lindeberg. more on Lindeberg and the history of CLT developments in [Cramér \(1976\)](#), also see [Schweder \(1980, 1999\)](#). mention the classical text [Billingsley \(1968\)](#). xx)

(xx mention: Billingsley invented Mr Portmanteau. xx)

(xx point to [von Bortkiewicz \(1898\)](#), and to Poisson himself. The classical, 1899, horse kicks, plus $\sum_{i=1}^n X_i \rightarrow_d \text{Pois}(\lambda)$ when $\sum_{i=1}^n p_i \rightarrow \lambda$ and $\max_{i \leq n} p_i \rightarrow 0$. xx)

The example in and lemma in Ex. 2.42 are from [Linero and Rosalsky \(2013\)](#).

The equivalence of Ex. 2.19 can be shown in several ways, the trick we use here is from [Pollard \(1984\)](#).

(xx Weierstraß, [Hveberg \(2019\)](#), A generation or so later, such results had been generalised to so-called Stone–Weierstraß theorems, stating, in various forms, that certain classes of functions are rich enough to deliver uniform approximations to bigger classes of functions. This is useful also in branches of probability theory. xx)

(xx the first of the two lemmas: Nils beta 1990. Weierstrass: see also [Hveberg \(2019\)](#). our result xxxx if an extension of one by [Inlow \(2010\)](#). xx)

(xx include also something on Fra Preface i *Life and Times of CLT*: “For those who teach a course in probability whose objective is to prove the Central limit theorem of interest [...] is commentary on the characteristic function approach employed by Lyapunov versus the ‘very simple’ proof, as Le Cam describes it, given by Lindeberg”. also, [LeCam \(1986\)](#). xx)

[xx for Scheffé: see [Scheffé \(1947\)](#), but also [Kusolitsch \(2010\)](#), who explains that the result is a special case of results published by F. Riesz in 1928. see also what Scheffe says in his paper about comments he got from Morse. xx]

(xx push this to Notes. xx) [Inlow \(2010\)](#) has shown how one can prove the usual CLT without the technical use of characteristic and hence complex functions. Essentially, he writes the X_i in question as $Y_i + Z_i$ with $Y_i = X_i I\{|X_i| \leq \varepsilon\sqrt{n}\}$ and $Z_i = X_i I\{|X_i| > \varepsilon\sqrt{n}\}$, after which ‘ordinary’ m.g.f.s may be used for the part involving the Y_i , yielding the normal limit, supplemented with analysis to show that the part involving the Z_i tends to zero in probability. – It is a non-trivial matter to extend Inlow’s arguments, from the CLT to the Lindeberg theorem, but this is precisely what is done in [Stoltenberg \(2019\)](#). Check that note, on the book website, and make sure you understand its main tricks and steps.

(xx When Jarl Waldemar Lindeberg was reproached for not being sufficiently active in his scientific work, he said, ‘Well, I am really a farmer’. And if somebody happened to say that his farm was not properly cultivated, his answer was, ‘Of course my real job is to be a mathematics professor’. Hundred years ago!, i.e. in 1920, he published his first paper on the CLT, and in 1922 he generalised his findings to the classical Lindeberg Theorem, with the famous Lindeberg Condition, securing limiting normality of a sum of independent but not identically distributed random variables. He did not know about Ляпунов’s earlier work, and therefore not about условие Ляпунова, the Lyapunov condition, which we treat below as a simpler-to-reach condition than the more general one of Lindeberg. Other luminaries whose work touch on these themes around the 1920ies and beyond include Paul Lévy, Harald Cramér, William Feller, and, intriguingly, Alan

Turing who (allegedly) won the war and invented computers etc. xx)

(xx point to a couple of characterisation theorems books, kagan linnik rao, one more. xx)

(xx for Notes: The little $\log(1+x)$ lemma is stated, proven, and used in [Hjort \(1990b, Appendix\)](#). xx)

(xx the material is from [Hjort and Pollard \(1993\)](#) and [Hjort \(1986a\)](#). xx)
 [xx For notes and pointers: Ex. 2.19(b) is from [Jacod and Protter \(2004, p. 166\)](#), they have it from [Pollard \(1984\)](#) xx].

ToDo notes, of 16 August 2026.

Clean and calibrate. Include a couple of classic nonparametric test procedures, like Wilcoxon, the sign test, more, to showcase the use of Lindeberg things to show limiting normality of such statistics too. point to [Hjort and Pollard \(1993\)](#) and [Hjort \(1986a\)](#).

Include some non-normal limits. Can take $\hat{\mu}^* = \{1 - c(D_n)\}\hat{\mu}_{\text{narr}} + c(D_n)\hat{\mu}_{\text{wide}}$ from model selection. And point to $n^{2/5}$ rates for f estimation.

I.3

Parameters, estimators, precision, confidence

With data observed from a statistical model, the theme of this chapter is that of constructing estimators for unknown statistical parameters, along with assessing their precision. Basic concepts include the bias, the variance, the mean squared error of estimators. The development also naturally leads to the important notion and basic machinery of confidence intervals. General estimation methods covered here include the method of moments and the method of quantiles; these can also be combined. For regression setups, with response variables influenced by covariates, we go through the method of least squares. To understand and utilise the properties of classes of estimators in general models, we utilise the machinery of large-sample normal approximations, from Ch. 2. This also enables one to assess precision and to compare different competing estimators. Similar remarks apply also for the more versatile method of maximum likelihood, treated in Ch. 5.

Key words: approximations, confidence, estimators, linear regression, model parameters, moment matching, quantile matching, risk

Most statistical models have *parameters*, as we learn from the generous variety of models in Ch. 1. Parameters may then be fine-tuned, or estimated, from data, which is the grand theme of the present chapter. In generic terms, if a model has density $f(y, \theta)$, with $\theta = (\theta_1, \dots, \theta_p)^t$ its parameter vector, we use data $\mathcal{D}$ to construct *an estimator* $\hat{\theta} = \hat{\theta}(\mathcal{D})$. Thus $f(y, \hat{\theta})$ is the fitted model, which we use for interpretation and inference, themes we return to in more detail in later chapters. The data $\mathcal{D}$ can often be in the form of direct independent observations $y_1, \dots, y_n$ from the model, but can be in vector form, and also be different in character, involving censoring mechanisms, or measurement error.

A sensible minimum demand for an estimator is that it should tend to the right value, with increasing data volume. Formally, if $\hat{\theta} = \hat{\theta}_n$ is the estimator for some parameter whose true value is θ_0 , based on a sample of size n , then we say that the estimator (or, more pedantically, the sequence of estimators) is *consistent* provided $\hat{\theta}_n$ converges in probability to θ_0 , i.e. $\hat{\theta}_n \rightarrow_{\text{pr}} \theta_0$ in the terminology of Ch. 2.

One often needs estimates and inference methods for *focus parameters*, those of particular and context-driven interest, which are one-dimensional functions $\mu = \mu(\theta_1, \dots, \theta_p)$ of the underlying model parameter vector. If $\hat{\mu}$ is an estimator for this parameter, we often care about its mean, represented here as

$$E_{\theta} \hat{\mu} = \mu + b(\theta). \tag{3.1}$$

consistency of
an estimator
focus parameter

The footnote signals that the expectation operator is at work at the parameter position θ . The $b(\theta)$ is termed *the bias*; in various cases it is a function of μ only, but in general it depends on the full parameter vector θ . If $E_\theta \hat{\mu} = \mu$, at all positions θ , we say the estimator is *unbiased*. In addition to wishing for estimators with small bias, we care about its variability, and often about its *mean squared error*

unbiased
estimator

mean squared
error

$$\text{mse}(\hat{\mu}, \theta) = E_\theta (\hat{\mu} - \mu)^2 = \text{Var}_\theta \hat{\mu} + b(\theta)^2, \quad (3.2)$$

the classic variance plus squared bias. This is a function of the unknown parameter, and gives a way of understanding and comparing performance for competing estimation schemes. When we can sort out the mathematics properly, depending on the situation at hand, we then choose estimators with smaller mse than those of competitors.

The $\text{mse}(\hat{\mu}, \theta)$ of (3.2) is sometimes called the risk, or risk function, and relates to having chosen squared error $(\hat{\mu} - \mu)^2$ as the underlying measure of quality. Other ways in which to compare and rank performance, involving also different quality functions and risk functions, are dealt with in Ch. 9.

Fundamental and conspicuous instruments in the statistical toolkits, when summarising and reporting findings of investigations, are *confidence intervals* and *testing of null hypotheses*. The development in the present chapter deals with the former, regarding interpretation, construction, properties, performance, whereas the following chapter handles the latter, along with further connections.

Consider in general terms data y , perhaps a long vector or a data matrix, from a model with parameter $\theta = (\theta_1, \dots, \theta_p)$, with $\mu = \mu(\theta_1, \dots, \theta_p)$ a parameter of interest. Then $[L(y), U(y)]$ is a *confidence interval*, with confidence level α , like 0.90 or 0.95 or an even higher 0.99, provided

confidence
interval

$$\Pr_\theta \{L(Y) \leq \mu \leq U(Y)\} = \alpha \quad \text{for all } \theta. \quad (3.3)$$

Thus $[L(Y), U(Y)]$ is a random interval, and with a high number of repeated situations it will capture the underlying μ a fraction α of the times. The reported confidence interval is $[L_{\text{obs}}, U_{\text{obs}}] = [L(y_{\text{obs}}), U(y_{\text{obs}})]$, computed based on the actually observed data y_{obs} . Occasionally one also wants to construct *confidence regions* for a parameter vector, as opposed to confidence intervals for each of its components. Thus if (a, b) is a parameter pair in a model, a random region R , based on data, is a 95 percent confidence region provided $\Pr_\theta((a, b) \in R) = 0.95$ for all model parameters θ .

In various setups one can study distributions, biases, variances, confidence etc. quite accurately, as will be seen in many exercises below. Often enough this might be too complicated, however, and one relies instead on good approximations. There is indeed a host of normal approximations, sometimes with additional tools for finetuning these, as studied in Ch. 2, with yet more to come in Ch. 5. These methods may be understood, appreciated, seen in action, and used for new situations, without necessarily having been through each δ and ε of Ch. 2.

In this chapter we learn certain estimation principles, including those associated with the method of moments and the method of quantiles. There is also room for combining such methods, or for coming up with new estimators in unfamiliar waters. We go on to more advanced models and hence estimation methods in later chapters, but included below is the basics of linear regression and the least sum of squares methods. The more versatile and often well-performing method of *maximum likelihood* will be studied with care in Ch. 5.

and in several
of our stories

Precise estimation in a few classical models

Ex. 3.1 *Mean squared error.* Suppose data lead to an estimator $\hat{\mu}$ for a focus parameter $\mu = \mu(\theta)$, in a model with parameter θ .

(a) Verify the mse formula (3.2), and note its Pythagorean character. Make a little right-angle triangle figure with absolute bias and standard deviation for the two short sides and root-mse, $\text{rmse} = \text{mse}^{1/2}$, on the long side. This root operation bring the risk components down to the original scale of the measurements.

(b) For a simple situation, let $Y \sim N(\theta, 1)$, with θ to be estimated. Find formulae for the mean squared errors of the three estimators $0.9Y$, Y , $1.1Y$. Note the interplay between bias and variance.

(c) Generalise to the case of $Y_1, \dots, Y_n$ being i.i.d. from $N(\theta, 1)$. Find $\text{mse}(\hat{\theta}, \theta)$ for the three estimators $0.99\bar{Y}$, $\bar{Y}$, $1.01\bar{Y}$, with $\bar{Y}$ the sample average. Comment on what you find.

(d) In somewhat more general terms, consider an i.i.d. sample $Y_1, \dots, Y_n$ from a distribution with unknown mean μ and variance σ^2 . Show that $\bar{Y}$ is unbiased with variance σ^2/n . If your estimator for μ is $c_n\bar{Y}$, what is required of c_n , in order for the mean squared error to go to zero with growing n ?

Ex. 3.2 *Binomial estimation.* Consider Y being binomial (n, p) , as in Ex. 1.3.

(a) To estimate p the canonical choice is $\hat{p} = Y/n$. Find its mean and variance, and a formula for $\text{mse}(\hat{p}, p) = E_p(\hat{p} - p)^2$.

(b) Then compare the simple binomial unbiased proportion with the estimator $\hat{p}_B = (Y+1)/(n+2)$; this is arguably the very first Bayes estimator. Find its bias and variance, and a formula for $\text{mse}(\hat{p}_B, p)$. Draw the two mse functions in a diagram, for say $n = 10$. When is the Bayes estimator better than the Y/n , according to this criterion?

Ex. 3.3 *Estimating the normal mean.* Consider independent observations $Y_1, \dots, Y_n$ from the normal distribution $N(\xi, \sigma^2)$.

(a) Prove that the sample average $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ has the $N(\xi, \sigma^2/n)$ distribution. This $\bar{Y}$ is the canonical estimator for μ . Find also a clear formula for its risk, or mean squared error, namely $\text{mse}(\bar{Y}, \mu, \sigma) = E_{\mu, \sigma}(\bar{Y} - \mu)^2$. The subscript indicates that the mean operator is with respect to the probability mechanism dictated by (μ, σ) .

(b) Then generalise the above somewhat, by finding the mean and variance also for the estimator $\hat{\mu} = b\bar{Y}$, with b a constant (which might be close to 1). Use this to put up a clear expression for

$$\text{mse}(b\bar{Y}, \mu, \sigma) = E_{\mu, \sigma}(b\bar{Y} - \mu)^2.$$

Illustrate this, for values $b = 0.98, 1.00, 1.02$, and comment. For what values of the parameters (μ, σ) will the estimator $0.98\bar{Y}$ be better than the classic $\bar{Y}$? Are there values of the parameters where $1.02\bar{Y}$ is better than the plain $1.00\bar{Y}$?

(c) Suppose the starting assumptions about the data at hand are changed to merely saying that the Y_i are i.i.d. with mean μ and standard deviation σ , i.e. we avoid saying that the distribution of the error terms $\varepsilon_i = (Y_i - \mu)/\sigma$ needs to be exactly standard normal. How does this affect your findings and claims for the previous points?

Ex. 3.4 *Estimating the normal variance.* As in Ex. 3.3, suppose there are i.i.d. data $Y_1, \dots, Y_n$ from the $N(\xi, \sigma^2)$. Here we care about the standard deviation parameter σ .

Bayes estimator
from 1763; see
Ch. 7

As we saw in Ex. 1.49, $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi_m^2$, where $m = n - 1$. Also, the Z is stochastically independent of the sample mean $\bar{Y}$.

(a) Use the statement above to find the mean and variance of $\hat{\sigma}^2 = cZ$ (where c ought to be about $1/n$). Find the mean squared error $\text{mse}(cZ, \sigma) = E_\sigma (cZ - \sigma^2)^2$. Check in particular the result for $c = 1/(n-1)$, the classical factor to make the estimator unbiased; for $c = 1/n$, which comes out of the maximum likelihood paradigm (see Ch. 5); and for $c = 1/(n+1)$. Find the best possible constant c for estimators of this type cZ , using the mean squared error on the σ^2 scale as criterion.

(b) Find also the mean and variance of $dZ^{1/2}$, seen as an estimator of σ , i.e. on the standard deviation scale, not that of the variance. Find an expression for $\text{mse}(dZ^{1/2}, \sigma) = E_\sigma (dZ^{1/2} - \sigma)^2$. Find the best d , according to this criterion.

(c) We may also finetune σ estimation on the log-scale. Examine the risk function $\text{mse}(kZ^{1/2}, \sigma) = E_\sigma \{\log(kZ^{1/2}) - \log \sigma\}^2$, and find the best value of k .

(d) A different solution to the issue of determining ‘the best constant’ when estimating σ , disregarding tradition and mathematical convenience, might be as follows. With $\hat{\sigma}^2 = Z/(n-1)$ being the traditional sample variance, with $1/(n-1)$ selected to achieve unbiasedness on the σ^2 scale, consider $\sigma^* = c_n \hat{\sigma}$, with c_n to be fine-tweaked perhaps a little bit away from 1. Find the c_n that makes $\text{risk}(c_n \hat{\sigma}, \sigma) = E_\sigma |c_n \hat{\sigma} - \sigma|$ smallest. This means relying on absolute error as loss function, and the solution needs numerical minimisation of a function which needs numerical integration. Give a table with these optimal c_n for say $n = 10, \dots, 30$. Show that $c_n \rightarrow 1$ as n grows.

Ex. 3.5 *Confidence interval for an exponential rate.* Choose a sample size n , and simulate i.i.d. variables $Y_1, \dots, Y_n$ from the exponential distribution, see Ex. 1.8, with parameter θ equal to say $\theta_0 = 3.33$.

(a) Construct a 90 percent confidence interval $[L, U]$ for θ . Check if θ_0 is contained in this interval, for the data you generated. Repeat the experiment say 100 times, and record how often the intervals contain θ_0 . What is in fact the distribution of N , the number of the 100 intervals which cover the truth?

(b) In addition to checking whether the intervals cover the truth, compute the length $D = U - L$, and give a histogram of its distribution. Find ED . Repeat also these experiments with a couple of other sample sizes, and comment.

Ex. 3.6 *Confidence interval for a normal variance.* Let $Y_1, \dots, Y_n$ be i.i.d. from a normal distribution. How can we set up confidence intervals for the standard deviation σ (or, equivalently, for the variance σ^2)? Writing $m = n - 1$, the sample variance is $\hat{\sigma}^2 = Z/m$, with $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi_m^2$, from Ex. 1.49.

(a) Start with $[a, b] = [\Gamma_m^{-1}(0.05), \Gamma_m^{-1}(0.95)]$, an interval covering the χ_m^2 with probability 0.90. Transform $a \leq m\hat{\sigma}^2/\sigma^2 \leq b$ to the confidence interval $\text{ci} = [(m/b)^{1/2}\hat{\sigma}, (m/a)^{1/2}\hat{\sigma}]$. Show in detail that indeed $\Pr_\sigma(\sigma \in \text{ci}) = 0.90$.

(b) Most often one wishes to estimate and assess the σ parameter directly, being on the same scale as the measurements, but once in a while it would be more natural to communicate and interpret results on the variance scale. Show in suitable detail that $\Pr_\sigma((m/b)\hat{\sigma}^2 \leq \sigma^2 \leq (m/a)\hat{\sigma}^2) = 0.90$; confidence intervals can in this fashion be easily transformed, say from θ to $g(\theta)$, as here, from σ^2 to σ , or the other way around.

(c) The construction above is ‘equitailed’, starting with 0.05 probability to the left of a and 0.05 probability to the right of b . One might somewhat more generally use any $[a, b]$

with 0.90 probability for the χ_m^2 , needing $\Gamma_m(b) - \Gamma_m(a) = 0.90$. The length of the 90 percent σ interval above is proportional to $1/a^{1/2} - 1/b^{1/2}$. Minimise this function, say for $m = 10, 20, 30, 40$. Compare these length-minimising 0.90 intervals with the simpler ones, and comment.

(d) Go through the analogue exercise, this time working with confidence intervals on the σ^2 scale, where the length is proportional to $1/a - 1/b$; minimise this, under the constraint $\Gamma_m(b) - \Gamma_m(a) = 0.90$, and compare with the above, for say $m = 10, 20, 30, 40$. The moral is that such delicate stretching doesn't matter so much, and in statistical practice we are typically content with the equitailed construction. See however the more principled construction of confidence distributions in Ch. 8.

Ex. 3.7 *Confidence interval for a normal mean.* Here we go through the basics for constructing confidence intervals for normal means. Since approximations to normality abound in statistical theory and practice, what we learn here quickly finds use also outside the strict normality assumptions.

(a) Start with the simplest prototype setup, a single Y from the $N(\xi, 1)$ model. Show that the random interval $[Y - 1.96, Y + 1.96]$ captures ξ with probability 0.95.

(b) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the $N(\xi, \sigma^2)$ model, at the moment with standard deviation parameter σ taken known. With $\bar{Y}$ as usual being the data average, show that $\sqrt{n}(\bar{Y} - \xi)/\sigma$ is standard normal, and deduce from this that $\bar{Y} \pm 1.96 \sigma/\sqrt{n}$ is a 95 percent confidence interval for ξ .

(c) In most cases also the σ is unknown, however. Let $\hat{\sigma}$ be the usual empirical standard deviation, see e.g. Ex. 3.4. Show that the natural construction

$$t = \frac{\bar{Y} - \xi}{\hat{\sigma}/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \xi)}{\hat{\sigma}} = \frac{\sqrt{n}(\bar{Y} - \xi)/\sigma}{\hat{\sigma}/\sigma} \quad (3.4)$$

has a distribution not depending on the two parameters, and that this distribution, call it G_n , is symmetric around zero. Deduce also that with $t_{0,n} = G_n^{-1}(0.975)$, the random interval $\bar{Y} \pm t_{0,n} \hat{\sigma}/\sqrt{n}$ covers ξ with probability 0.95.

(d) It is then a matter of finding and perhaps tabulating the distribution G_n of t . It is in fact the celebrated t distribution, with $df = n - 1$ degrees of freedom, see Ex. 1.50. But even without that specific knowledge detail, we could easily have simulated a high number for t , from (3.4), and read off the required quantile.

(e) In more generality, suppose β is a model parameter for which there is an estimator $\hat{\beta}$ with distribution $N(\beta, c_n^2 \sigma^2)$, say, with a known factor c_n . Suppose also that there is a statistically independent estimator $\hat{\sigma}$ for σ , with the property that $\hat{\sigma}^2/\sigma^2 \sim \chi_m^2/m$, for a known m . Then show that $t = (\hat{\beta} - \beta)/(c_n \hat{\sigma}) \sim t_m$. Put up a 0.99 confidence interval for β based on this.

(f) Suppose $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ are random samples from two normal groups with equal variance, say $N(\xi_1, \sigma^2)$ and $N(\xi_2, \sigma^2)$. Find the distribution of the difference of sample means $D_n = \bar{X} - \bar{Y}$, and construct a confidence interval for $\delta = \xi_1 - \xi_2$. This is related to the classic two-sample t-test, see Ex. 4.4.

Ex. 3.8 *Normal quantiles: estimation and confidence.* Consider again the setup of Ex. 3.3, with a sample of Y_i from the normal $N(\xi, \sigma^2)$ model. In the present exercise we care about quantiles, as opposed to ‘only’ the mean or the standard deviation.

(a) Show that the c.d.f. of Y_i may be written $F(y) = \Pr(Y_i \leq y) = \Phi((y - \xi)/\sigma)$, with $\Phi(x)$ the c.d.f. for the standard normal. Show that the q quantile $F^{-1}(q)$ is equal to

$\gamma_q = \xi + z_q \sigma$, with $z_q = \Phi^{-1}(q)$. Thus the 0.95 quantile is $\gamma_{0.95} = \xi + 1.645 \sigma$, etc.

(b) Find the mean and variance of the natural estimator $\hat{\gamma}_q = \bar{Y} + z_q \hat{\sigma}$, where $\hat{\sigma} = (Z/m)^{1/2}$, with Z as in Ex. 3.4 and $m = n - 1$.

(c) Show that we may write

$$\begin{aligned}\hat{\gamma}_q - \gamma_q &\sim \xi + (\sigma/\sqrt{n})N + z_q \sigma (K_m/m)^{1/2} - \xi - z_q \sigma \\ &= \sigma[(1/\sqrt{n})N + z_q\{(K_m/m)^{1/2} - 1\}],\end{aligned}$$

in terms of $N \sim N(0, 1)$ and $K_m \sim \chi_m^2$, with these being independent. Verify that the pivot

$$W_{n,q} = \frac{\sqrt{n}(\hat{\gamma}_q - \gamma_q)}{\hat{\sigma}} \sim \frac{N + z_q \sqrt{n}\{(K_m/m)^{1/2} - 1\}}{(K_m/m)^{1/2}}$$

pivot: formally
defined in
Ex. 8.11

has a distribution not depending on ξ, σ . It can be simulated, for any given sample size n and quantile level q . With a_n such that $\Pr(-a_n \leq W_{n,q} \leq a_n) = 0.95$, convert this to a 95 percent confidence interval for γ_q . Note that for $q = 0.50$, the median case, the distribution of the $W_{n,q}$ is the t_m , the t with degrees of freedom $m = n - 1$.

(d) Use the delta method of Ex. 2.49 to show that $\sqrt{n}\{(K_m/m)^{1/2} - 1\} \rightarrow_d N(0, \frac{1}{2})$, and explain that $W_{n,q} \rightarrow_d N(0, 1 + \frac{1}{2}z_q^2)$. Construct an approximate 95 percent confidence interval for γ_q based on this. Simulate a dataset of size e.g. $n = 50$, use both methods from this exercise to find 95 percent confidence intervals for the $\gamma_{0.75}$ quantile. Note that the delta method of Ex. 2.53 leads to approximate confidence intervals for *any* smooth $\gamma = g(\xi, \sigma)$; the point here is that a more exact analysis works for the quantile.

Ex. 3.9 *The empirical distribution function.* Assume there is an i.i.d. dataset $Y_1, \dots, Y_n$ from an unknown distribution, with c.d.f. $F(t) = \Pr(Y_i \leq t)$. We have learned about the empirical distribution function in Ex. 2.11 and 2.60, with $F_n(t) = n^{-1} \sum_{i=1}^n I(Y_i \leq t)$ converging uniformly to the underlying F , with probability one; here we learn about how fast this happens, via convergence in distribution.

more on the
empirical c.d.f.

(a) Explain that the empirical c.d.f. is the cumulative of the probability measure that puts probability mass $1/n$ at each data point. This is the natural nonparametric estimator of the unknown F .

(b) Construct a version of Figure 3.1, left panel, where $n = 100$ datapoints are simulated from the distribution $f = 0.50 \text{Expo}(r_1) + 0.50 \text{Expo}(r_2)$, with rates $r_1 = 2.00$ and $r_2 = 4.00$. Check how well the nonparametric estimator F_n succeeds in coming close to F .

(c) Since we know so much about the binomial, we quickly learn a few basic properties of the F_n . Show that $F_n(t)$ is unbiased for $F(t)$, and that its variance is $F(t)\{1 - F(t)\}/n$.

(d) Consider the process $Z_n(t) = \sqrt{n}\{F_n(t) - F(t)\}$. Show that it has mean zero, and that $Z_n(t) \rightarrow_d Z(t)$, say, where $Z(t)$ is a zero-mean normal with variance $F(t)\{1 - F(t)\}$. Show also that

$$\text{cov}\{Z_n(t), Z_n(t')\} = F(t)\{1 - F(t')\} \quad \text{for } t \leq t'.$$

Compute and display the Z_n plot, using the same data values as for the previous figure; in other words, construct a version of Figure 3.1, right panel. Such plots may e.g. be used to check model adequacy – if the data come from a distribution not close to the F used to construct the plot, then the Z_n plot will deviate significantly from the zero line. To understand what might qualify as ‘significantly different from the zero line’ means

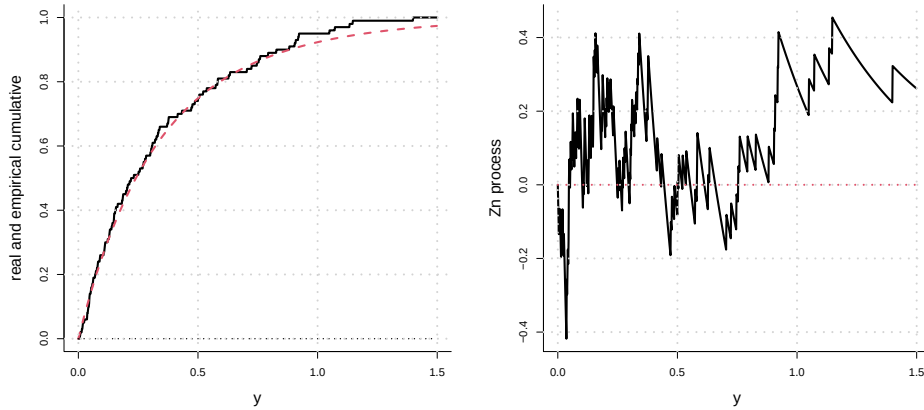


Figure 3.1: Left panel: the real underlying data-generating $F(t)$ (dashed, red), with the empirical distribution function $F_n(t)$ (full line, black), computed from a sample of $n = 100$ data points from F . Right panel: the process $Z_n(t)$, computed for the data used for the same data. In 95 percent of such cases, the maximum absolute value of the Z_n process will be below 1.358.

we need theory for the behaviour of the full Z_n process, not merely the pointwise result that $Z_n(t) \rightarrow_d N(0, F(t)\{1 - F(t)\})$. We return to such questions in Ch. 10.

Ex. 3.10 *Estimating the normal density.* Most often the statistical interest lies in estimating some parameter related to, or expressed through, the normal distribution, like the mean, spread, or quantile, as illustrated above. In some situations one wishes to estimate the density itself. Consider once again a sample $Y_1, \dots, Y_n$ from the normal $N(\xi, \sigma^2)$.

(a) For the parameter σ , we shall again use $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi_m^2$, with $m = n - 1$, as in several previous exercises. With the traditional default estimator $\hat{\sigma}^2 = Z/(n - 1)$ of (1.6), find formulae for the mean of $1/\hat{\sigma}^2$ and $\log \hat{\sigma}$.

(b) In Ex. 9.42 we construct an unbiased estimator of the normal density $f(y, \xi, \sigma)$. For the present exercise, construct unbiased estimators for $1/\sigma^2$ and for $\log \sigma$, and then for the log-density function $\log f(y, \xi, \sigma) = -\log \sigma - \frac{1}{2}(y - \xi)^2/\sigma^2 - \frac{1}{2} \log(2\pi)$. For this task, show in fact that $E \log \hat{\sigma} = \log \sigma + a_m$, with $a_m = E \log(\chi_m^2/m) = \{\log 2 + \psi(\frac{1}{2}m)\}/(2m)$, and that

$$E \frac{m}{m-2} \frac{(y - \hat{\xi})^2}{\hat{\sigma}^2} = \frac{(y - \xi)^2 + \sigma^2/n}{\sigma^2}.$$

(c) For independent samples from normal populations $N(\xi_1, \sigma_1^2)$ and $N(\xi_2, \sigma_2^2)$, construct an unbiased estimator for $d(y) = \log\{f_2(y)/f_1(y)\}$.

Confidence via normal approximations

Ex. 3.11 *Confidence intervals via normal approximations.* As we've already seen in various situations of Ch. 2, there are often estimators for interest parameters for which there is approximate normality. Then various recipes under strict normality can still be used, but as approximations.

(a) Suppose μ is such a parameter of interest, for which there is an estimator $\hat{\mu}$, being approximately normal, in the mathematical sense of $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_d N(0, \tau^2)$, for some

check also Wald tests, Ex. 4.5

appropriate limiting variance τ^2 . Suppose also that there is a consistent estimator $\hat{\tau}$ of τ , with $\hat{\tau}/\tau \rightarrow_{\text{pr}} 1$. Show that $Z_n = \sqrt{n}(\hat{\mu} - \mu)/\hat{\tau} \rightarrow_d N(0, 1)$; you may check with Ex. 2.21.

(b) Show under these mild and very frequently met assumptions that

$$\Pr(\hat{\mu} - 1.96 \hat{\tau}/\sqrt{n} \leq \mu \leq \hat{\mu} + 1.96 \hat{\tau}/\sqrt{n}) \rightarrow 0.95.$$

In other words, the $\hat{\mu} \pm 1.96 \hat{\tau}/\sqrt{n}$ is an *approximate* or *asymptotic* 95 percent interval for μ . Note the grand generality here; this simple construction works in a large variety of situations, also in nonparametric setups, cases with dependent data, etc.

(c) The simplest interesting application of this standard recipe is for the unknown mean ξ of a population. Verify via the CLT of Ch. 2 that $\sqrt{n}(\bar{Y} - \xi)/\hat{\sigma} \rightarrow_d N(0, 1)$. Hence the t-based interval $\bar{Y} \pm 1.96 \hat{\sigma}/\sqrt{n}$, for which we have very precise probability computations under normality, is large-sample correct even if the data are not at all normal.

(d) Suppose (X, Y, Z) is trinomial (n, p, q, r) , with $p + q + r = 1$. Construct an approximate 90 percent confidence interval for $d = q - p$, then for $\gamma = r/(p + q)$. Check that you see how similar and not complicated tasks can be tended to in the examples of Ex. 2.50.

Ex. 3.12 *Confidence intervals for the standard deviation, outside normality.* Consider i.i.d. data $Y_1, \dots, Y_n$, from a distribution with mean ξ and standard deviation σ , from which we compute the classical sample mean $\hat{\xi} = \bar{Y}$ and empirical variance $\hat{\sigma}^2$ of (1.6). Here we illustrate the general large-sample methods by building confidence intervals for σ , with no assumptions on the distribution of the data, like normality. The only mild assumption we make is a finite fourth moment, in order for $\hat{\sigma}$ to have a clear limit distribution. See Figure 3.2 for 100 simulated confidence intervals, all attempting to capture the true value, here $\sigma = 1$, for two different sample sizes.

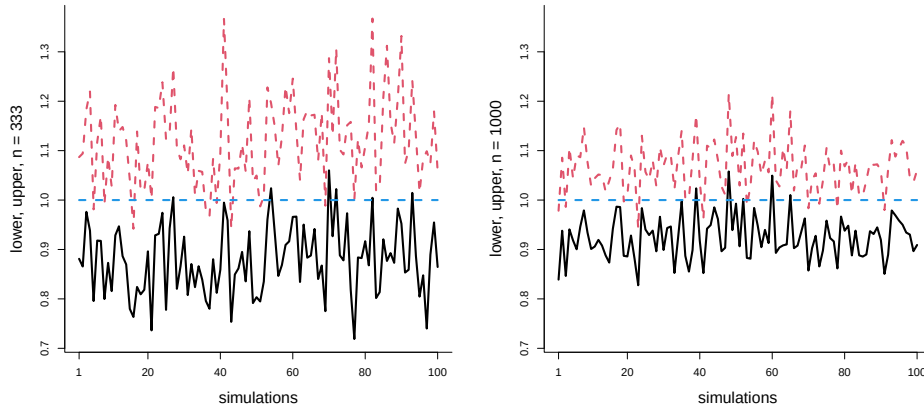


Figure 3.2: Simulations, with datasets from the unit exponential, displaying lower and upper confidence points for 90 percent intervals; the intervals attempt to cover the true value $\sigma = 1$, and will succeed about 90 percent of the time. Left panel: with $n = 333$; right panel: with $n = 1000$.

(a) Make sure you understand and can show that $\hat{\xi}$ and $\hat{\sigma}$ are consistent for ξ and σ , from the LLN theorems of Ch. 2.

(b) Via results from Ex. 2.48 and the delta method of Ch. 2, show that

$$\begin{pmatrix} \sqrt{n}(\hat{\xi} - \xi) \\ \sqrt{n}(\hat{\sigma} - \sigma) \end{pmatrix} \rightarrow_d N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3 & \frac{1}{2} + \frac{1}{4}\gamma_4 \end{pmatrix} \right),$$

in terms of skewness $\gamma_3 = E\{(Y - \xi)/\sigma\}^3$ and kurtosis $\gamma_4 = E\{(Y - \xi)/\sigma\}^4 - 3$.

(c) Show that $\hat{\gamma}_4 = (1/n) \sum_{i=1}^n \{(Y_i - \bar{Y})/\hat{\sigma}\}^4 - 3$ is consistent for γ_4 , and use this to construct an approximate 90 percent confidence interval for σ . Note that this is a nonparametric procedure, totally free of other distributional assumptions, like normality; if one assumes normality, as an extra condition, one may do more, of course.

(d) For an illustration, consider the unit exponential distribution; show that the standard deviation is 1 and that the kurtosis is $\gamma_4 = 6$. Simulate a suitably high number of datasets of size $n = 333$ from this distribution. For each simulated dataset, compute $\hat{\gamma}_4$, to check how close it is to γ_4 , along with the approximate 90 percent confidence interval for σ . Construct a version of Figure 3.2 (left panel for $n = 333$ and right panel for $n = 1000$). Examine in particular the coverage of your intervals: how often do they contain the correct σ ? Use simulations to check how big n must be, in order for $\hat{\gamma}_4$ to be inside $[5.5, 6.5]$ with probability at least 95 percent.

answer: very
big; empirical
kurtosis
converges slowly

(e) Generate a dataset of size $n = 333$ from the unit exponential, and construct an approximate 90 percent confidence ellipsoid on your screen for (ξ, σ) . Check if it contains the true values.

Ex. 3.13 *Inference for the normal density, via large-sample methods.* Consider again i.i.d. data $Y_1, \dots, Y_n$ from a normal density $f(y, \xi, \sigma)$. In Ex. 3.10 we constructed an estimator $\hat{f}(y)$ with the property that $\log \hat{f}(y)$ is exactly unbiased for $\log f(y, \xi, \sigma)$, for any given n . Here we instead work out the large-sample approximations for the direct estimator $f^*(y) = f(y, \hat{\xi}, \hat{\sigma})$, having plugged in the usual empirical mean and empirical standard deviation.

(a) We have seen in Ex. 2.53 that $(\sqrt{n}(\hat{\xi} - \xi), \sqrt{n}(\hat{\sigma} - \sigma))$ tends to (A, B) , independent zero-mean normals with variances σ^2 and $\frac{1}{2}\sigma^2$. For fixed y , write $x = (y - \xi)/\sigma$. Use the delta method to explain that (i) $\sqrt{n}(\log \hat{\sigma} - \log \sigma) \rightarrow_d (1/\sigma)B$; (ii) $\sqrt{n}((y - \hat{\xi})/\hat{\sigma} - (y - \xi)/\sigma) \rightarrow_d (1/\sigma)(-A - xB)$; (iii) $\sqrt{n}\{(y - \hat{\xi})^2/\hat{\sigma}^2 - x^2\} \rightarrow_d (1/\sigma)(-2xA - x^2B)$. Then combine these to show that

$$\sqrt{n}\{\log f^*(y) - \log f(y, \xi, \sigma)\} \rightarrow_d (1/\sigma)(-B + xA + \frac{1}{2}x^2B),$$

a zero-mean normal with variance $\tau^2 = x^2 + (\frac{1}{2}x^2 - 1)^2 = 1 + \frac{1}{4}x^4$.

(b) Show that the log-unbiased estimator $\log \hat{f}(y)$ is close enough to this $\log f^*(y)$ to make the scaled difference $\sqrt{n}\{\log f^*(y) - \log \hat{f}(y)\}$ go to zero in probability; explain that the two estimators therefore have the same limit distribution.

(c) Explain how a pointwise 90 percent confidence band can be constructed for the log-density, of the type $\log f^*(y) \pm 1.645 \hat{\tau}(y)$, which then may be transformed back to the density scale.

Ex. 3.14 *Variance of the variance estimator.* Let $Y_1, \dots, Y_n$ be i.i.d., with mean ξ , variance σ^2 , and finite kurtosis $\gamma_4 = E(Y_i - \xi)^4/\sigma^4 - 3$.

(a) With $A = \sum_{i=1}^n (Y_i - \bar{Y})^2$, show that $E A = (n-1)\sigma^2$. This says that $\hat{\sigma}^2 = A/(n-1)$ is unbiased for the variance, regardless of the underlying distribution.

(b) If the Y_i are actually normal, then $A \sim \sigma^2 \chi_m^2$, with $m = n-1$. Show that $\text{Var } \hat{\sigma}^2 = 2\sigma^4/(n-1)$.

(c) Outside normality, work out an expression for $\text{Var } A$, and show that

$$\text{Var } \hat{\sigma}^2 = \left(3 + \gamma_4 - \frac{n-3}{n-1}\right) \frac{\sigma^4}{n} = \frac{2\sigma^4}{n-1} + \frac{\gamma_4 \sigma^4}{n}.$$

Simulate a high number of samples of size $n = 12$ from the t distribution t_m , with say $m = 6$, and ‘verify’ the formula.

(d) Check: the finite-sample formula matches the large-sample limit for $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$.

Ex. 3.15 *The binomial, the normal approximation, and confidence intervals.* Consider Y_n , a binomial (n, p) . To invoke the CLT it is practical to use the $Y_n = X_1 + \dots + X_n$ representation, in terms of X_i being i.i.d. Bernoullis.

(a) Use the CLT of Ch. 2 to deduce that the normalised variable

$$W_n = \frac{\sum_{i=1}^n (X_i - p)}{\{\text{Var} \sum_{i=1}^n (X_i - p)\}^{1/2}} = \frac{Y_n - np}{\{np(1-p)\}^{1/2}} = \frac{\hat{p} - p}{\{p(1-p)/n\}^{1/2}}$$

converges to the standard normal $N(0, 1)$ with increasing n . Discuss briefly the relevance of the skewness result from Ex. 1.3 in light of the limiting normality.

(b) With $\hat{p}_B = (Y + 1)/(n + 2)$, show that the difference between W_n and $W_{n,B} = \sqrt{n}(\hat{p}_B - p)/\{p(1-p)\}^{1/2}$ is so small, for large n , that $W_{n,B}$ must have the same normal limit. The confidence intervals we construct below, based on $\hat{p}$, can therefore alternatively be based on $\hat{p}_B$.

(c) Verify from the above that $\Pr(-1.96 \leq W_n \leq 1.96) \rightarrow 0.95$ as sample size increases, and use this to construct an interval, based on having observed $Y_n = y$ in a given experiment with known n , which covers the true p with probability approximately 95 percent. binomial
confidence

(d) There are actually several constructions of such confidence intervals, with this property. Considering the basic estimator $\hat{p} = Y/n$ again, write $\sigma_n^2 = p(1-p)/n$ for its variance, and $\hat{\sigma}_n^2 = \hat{p}(1-\hat{p})/n$ for its estimated variance. Verify that both asymptotic
equivalence

$$W_n = (\hat{p} - p)/\sigma_n \quad \text{and} \quad W'_n = (\hat{p} - p)/\hat{\sigma}_n$$

tend to the standard normal in distribution. Now show that the arguments above, used for W'_n in lieu of W_n , lead to the confidence interval $\hat{p} \pm 1.96 \hat{\sigma}_n$ instead. Exemplify, with $n = 100$, for the three cases $y = 22$, $y = 55$, $y = 77$, where you compute both versions of the 95 percent confidence interval for p .

(e) Suppose certain details related to your applied research project require that you compute the probability p that $L \leq 1.33 R$, where

$$L = \{(G_1/G)(G_2/G)(G_3/G)(G_4/G)\}^{1/4}, \quad R = \{(G_5/G)(G_6/G)(G_7/G)\}^{1/3},$$

in terms of $G_1, \dots, G_8$ being i.i.d. from the χ_{12}^2 distribution (the chi-squared with degrees of freedom equal to 12), and $G = \sum_{i=1}^8 G_i$. Since it's hard to find an exact formula, or an exact answer in other ways, you *simulate* a high number sim of such vectors $(G_1, \dots, G_8)$, and check for each simulation whether the event just described takes place or not. How large should sim be, in order for your simulation based estimate of p to be correct to three decimal places? Carry out such simulations and thus find p . Display also a histogram of simulated L/R .

Ex. 3.16 *Confidence intervals via the delta method.* The delta method of Ex. 2.49 can be used to derive approximate normality of various estimators, after which the generic recipe of Ex. 3.11 may be used to form confidence intervals. Once limiting normality has been established, there are automatically limiting normality statements and uses for transformed parameters.

(a) Suppose $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_d N(0, \kappa^2)$, for some parameter μ of interest. Confidence intervals then take the form $\hat{\mu} \pm (c/\sqrt{n})\hat{\kappa}$. With a smooth transformation to $\gamma = g(\mu)$ and $\hat{\gamma} = g(\hat{\mu})$, explain that we have $\sqrt{n}(\hat{\gamma} - \gamma) \rightarrow_d g'(\mu)\kappa N(0, 1)$. If $\gamma = g(\mu)$ is monotone, with an inverse $\mu = g^{-1}(\gamma)$, explain that one via back-transformation can form the alternative interval

$$g^{-1}(\hat{\gamma} \pm (c/\sqrt{n})g'(\hat{\mu})\hat{\kappa}).$$

Explain also how this simplifies for variance stabilising transformations, where $g'(\mu)\kappa = 1$.

(b) Illustrate the above for a binomial (n, p) situation, where we know from Ex. 2.50 that $\sqrt{n}(\hat{\gamma} - \gamma) \rightarrow_d N(0, 1)$, with $\gamma = g(p) = 2 \arcsin(p^{1/2})$: construct the interval on the scale of γ , and then back-transform to the p scale. Check numerically how close such intervals are to the standard ones, for a few combinations of n and binomial outcome y .

(c) Show via Taylor expansion that $g^{-1}(\hat{\gamma} + \varepsilon) = \hat{\mu} + (g^{-1})'(\hat{\gamma})\varepsilon + O(\varepsilon^2)$ for small ε . Deduce that the transformation intervals above become large-sample equivalent to the $\hat{\mu} \pm (c/\sqrt{n})\hat{\kappa}$ we started out with.

(d) So which g should one perhaps use? Suppose g_a is one such transformation, with a tweaking parameter a . Argue that the best choices of a are those for which $g_a(\hat{\mu}) - g_a(\mu)$ are close to having zero-mean normal distributions. Play a bit with this idea, for the case of having normal approximation confidence intervals for σ in the $N(\xi, \sigma^2)$ situation. The starting point here would be $\sqrt{n}(\hat{\sigma} - \sigma) \rightarrow_d N(0, \frac{1}{2}\sigma^2)$; show that this is associated with approximate intervals of the type $\hat{\sigma}\{1 \pm (z/\sqrt{n})/\sqrt{2}\}$. Then attempt the transformation $g_a(\sigma) = \sigma^a$; the task is to check accurate normality for the appropriate $\sqrt{n}(\hat{\sigma}^a - \sigma^a)$. You should find that $a = 1/3$ is good for this purpose. Explain then how confidence intervals can be set from this. Check if these become more accurate, for small n , than the direct ones pointed to above.

Quantiles and sample quantiles

Ex. 3.17 Uniform ordering. Consider $U_1, \dots, U_n$ i.i.d. from the uniform distribution on the unit interval. Order these, to $U_{(1)} < \dots < U_{(n)}$.

(a) Explain that in order to have $U_{(i)} \in [u, u + du]$, one needs $i - 1$ points in $[0, u]$, 1 point in $[u, u + du]$, and $n - 1$ points in $(u + du, 1]$. Show from this trinomial setup that $U_{(i)}$ has density

$$g_i(u) = \frac{n!}{(i-1)!1!(n-i)!} u^{i-1}(1-u)^{n-i} \quad \text{for } u \in (0, 1).$$

Explain that $U_{(i)} \sim \text{Beta}(i, n - i + 1)$, see Ex. 1.22, and explain from this that

$$E U_{(i)} = p_i = i/(n+1) \text{ and that } \text{Var } U_{(i)} = p_i(1-p_i)/(n+2).$$

(b) Explain further that $U_{(i)} \leq p$ is equivalent to the binomial count Z of datapoints falling in $[0, p]$ being at least i , and that this gives rise to the following equation:

$$\frac{n!}{(i-1)!(n-i)!} \int_0^p u^{i-1}(1-u)^{n-i} du = \sum_{j=i}^n \binom{n}{j} p^j (1-p)^{n-j}.$$

(c) With $i < j$, show that $(U_{(i)}, U_{(j)})$ has density

$$g_{i,j}(u, v) = \frac{n!}{(i-1)!1!(j-i-1)!1!(n-j)!} u^{i-1}(v-u)^{j-i-1}(1-v)^{n-j}$$

connection to
Beta
distributions

Beta and
binomial
connection

for $u < v$. The idea behind the reasoning, and the ensuing notation, is that in order to see $U_{(i)} \in [u, u + du]$ and $U_{(j)} \in [v, v + dv]$, there is a multinomial situation, with five boxes $[0, u], [u, u + du], [u + du, v], [v, v + dv], [v + dv, 1]$, inside which we need to find $i - 1, 1, j - i - 1, 1, n - j$ datapoints.

(d) For an i.i.d. uniform sample $U_1, \dots, U_n$ on $[0, 1]$, consider the uniform range $R_n = U_{(n)} - U_{(1)}$, where we know that $U_{(1)} \sim \text{Be}(1, n)$. Show that given $U_{(1)} = u$, $U_{(n)}$ can be represented as $u + Z$, where Z is the maximum of another uniform sample, of size $n - 1$, on $[0, 1 - u]$. Use this to show that the c.d.f. of R_n can be expressed as $H_n(r) = nr^{n-1}(1 - r) + r^n = nr^{n-1} - (n - 1)r^n$, and show that this is the $\text{Be}(n - 1, 2)$ distribution.

(e) In general, if $Y_1, \dots, Y_n$ are i.i.d. from a density f , show that the joint density for the full order statistic vector $(Y_{(1)}, \dots, Y_{(n)})$ is $n! f(y_{(1)}) \cdots f(y_{(n)})$, on the set where $y_{(1)} < \cdots < y_{(n)}$. In particular, for order statistics from the uniform distribution, show that the joint density of $(U_{(1)}, \dots, U_{(n)})$ is flat and equal to $n!$ on the set $u_{(1)} < \cdots < u_{(n)}$.

(f) Use this, in conjunction with Ex. 1.20, to demonstrate that

connection to
exponential and
Dirichlet

$$\begin{aligned} (U_{(1)}, U_{(2)}, \dots, U_{(n)}) &=_d (D_1, D_1 + D_2, \dots, D_1 + \cdots + D_n) \\ &=_d (V_1/S, (V_1 + V_2)/S, \dots, (V_1 + \cdots + V_n)/S), \end{aligned}$$

with $V_1, \dots, V_n, V_{n+1}$ being i.i.d. from the unit exponential, with sum $S = V_1 + \cdots + V_{n+1}$, and $(D_1, \dots, D_n, D_{n+1})$ is a flat Dirichlet $(1, \dots, 1, 1)$. The differences $D_i = U_{(i)} - U_{(i-1)}$ are called the *spacings*. We use ‘ $=_d$ ’ to signal equality in distributions. Show that this leads to the representation of the order statistics process as

$$U_{([nq])} = \sum_{i=1}^{[nq]} V_i / \sum_{i=1}^{n+1} V_i \quad \text{for } 0 \leq q \leq 1. \quad (3.5)$$

Here $[nq]$ is the largest integer less than or equal to nq . We have seen in Ex. 2.14 that $U_{([nq])} \rightarrow_{\text{pr}} q$, via consistency properties of the empirical c.d.f. Now prove this result again, via the representation (3.5) and the law of large numbers.

Ex. 3.18 Ratios of ordered uniforms. Let $U_1, \dots, U_n$ be an i.i.d. sample from the uniform distribution on the unit interval, and order these into $U_{(1)} < \cdots < U_{(n)}$. From these form the ratios

$$V_1 = U_{(1)}/U_{(2)}, \quad V_2 = U_{(2)}/U_{(3)}, \quad \dots, \quad V_{n-1} = U_{(n-1)}/U_{(n)}, \quad V_n = U_{(n)}/1.$$

(a) Show that the inverse transformation leads to the representation

$$U_{(n)} = V_n, \quad U_{(n-1)} = V_n V_{n-1}, \quad \dots, \quad U_{(2)} = V_n V_{n-1} \cdots V_2, \quad U_{(1)} = V_n V_{n-1} \cdots V_2 V_1.$$

(b) Find the joint probability density for $(V_1, \dots, V_n)$, and show in fact that these are independent, with

$$V_1 \sim \text{Beta}(1, 1), \quad V_2 \sim \text{Beta}(2, 1), \quad \dots, \quad V_{n-1} \sim \text{Beta}(n - 1, 1), \quad V_n \sim \text{Beta}(n, 1).$$

(c) Independently of details above, we know from Ex. 3.17 that $U_{(i)}$ has a $\text{Beta}(i, n - i + 1)$ distribution. Explain that this $\text{Beta}(i, n - i + 1)$ can be represented as a product of different independent Beta variables.

(d) It is of course a somewhat cumbersome simulation recipe for generating a uniform sample, but it is a useful exercise, opening doors $\mathcal{E}$ minds to fruitful generalisations: For $n = 10$, say, generate ordered uniform samples of size n in your computer via the representation above, in terms of products of Beta variables. Carry out some checks to see that each single $U_{(i)}$ then has the right distribution, i.e. as described in (c).

(e) Work with the following generalisation of the construction above: Let $X_1, \dots, X_n$ be an i.i.d. sample from the distribution with density $f(x) = ax^{a-1}$, i.e. a $\text{Beta}(a, 1)$. Again form the ratios $V_i = X_{(i)}/X_{(i+1)}$ as above, leading to $X_{(i)} = V_i V_{i+1} \cdots V_n$. Show that the V_i are again independent, now with $V_i \sim \text{Beta}(ai, 1)$.

Ex. 3.19 *Min and max of two uniforms.* Suppose Y_1, Y_2 are i.i.d. from a density $f(y)$, and order them, to $V_1 < V_2$.

(a) Show that (V_1, V_2) has joint density $2f(v_1)f(v_2)$, on the set where $v_1 < v_2$.

(b) Then consider the special case of two datapoints from the uniform distribution on the unit interval, ordered to $V_1 < V_2$. Show that $R = V_1/V_2$ is another uniform on the unit interval, and that $W = V_2 - V_1$ is a $\text{Beta}(1, 2)$. Show that $\Pr(V_2 - V_1 \leq c) = \Pr(V_2 - V_1 > c) = \frac{1}{2}$, for $c = 1 - 1/\sqrt{2} = 0.2929$. (This is actually the answer to a specific question related to the best performance of the Wild Binary Segmentation method for change detection.)

(c) Find also the joint distribution for (R, W) here.

Ex. 3.20 *The sample median.* Let $Y_1, \dots, Y_n$ be i.i.d. from a positive density f with true median $\theta = F^{-1}(\frac{1}{2})$.

(a) Suppose for simplicity that n is odd, say $n = 2m + 1$. Show that M_n has density of the form

$$g_n(y) = \frac{(2m+1)!}{m!m!} F(y)^m \{1 - F(y)\}^m f(y).$$

(b) Show then that the density of $Z_n = \sqrt{n}(M_n - \theta)$ can be written in the form $h_n(z) = g_n(\theta + z/\sqrt{n})/\sqrt{n}$. Prove that

$$h_n(z) \rightarrow (2\pi)^{-1/2} 2f(\theta) \exp\{-\frac{1}{2}4f(\theta)^2 z^2\},$$

where the Stirling approximation formula of Story #103 may be of use. The limit is the density $h(z)$ of the normal $N(0, \tau^2)$, with $\tau = \frac{1}{2}/f(\theta)$. We have hence proved $Z_n \rightarrow_d N(0, \tau^2)$, by Scheffé's lemma; see Ex. 2.6.

(c) So when is the sample mean best, and when might the sample median be the better estimator, when it comes to estimating the centre point θ of a symmetric density? Explain that this for large n is a matter of the ratio

$$\rho = \frac{\sigma}{\frac{1}{2}/f(\theta)} = 2\sigma f(\theta),$$

where σ is the standard deviation for f . Argue that if $\rho < 1$, then the sample mean is best, and that if $\rho > 1$, then the sample median is the best.

(d) Compare the limiting distributions for the sample mean and the sample median for the normal density, the double exponential density $\frac{1}{2} \exp(-|y|)$, and the Cauchy density $(1/\pi)/(1 + y^2)$.

(e) Consider the t distribution, with degrees of freedom ν , see Ex. 1.50. Find an expression for the ratio $\rho = \rho(\nu)$, plot $(\nu, \rho(\nu))$ in a diagram, and comment. Show that $\rho(\nu)$

approaches $(2/\pi)^{1/2} = 0.7979$ for large ν . Show that for $\nu < 4.678$, there is roughness at the top, and the median is best; whereas for $\nu > 4.678$, there density is smoother at the top, and the mean is best.

(f) Carry out a similar analysis for the binormal symmetric mixture model $f = \frac{1}{2}N(-a, 1) + \frac{1}{2}N(a, 1)$. For which values of a is the sample median a better estimator of the centre point than the sample mean?

Ex. 3.21 Sample quantiles. Suppose $Y_1, \dots, Y_n$ are i.i.d. from some distribution with positive density f and c.d.f. F . The sample median estimates the population median $F^{-1}(0.50)$, and similarly the sample quantile $Q_n(q)$, at any prescribed level $q \in (0, 1)$, estimates the population quantiles $F^{-1}(q)$. Built-in functions like `quantile(data, 0.33)` in R find such sample quantiles directly, so users do not need the cumbersome linear interpolation fiddling between the two ordered observations coming closest to nq , or to care too much about ties in the data due to rounding-off errors. This exercise finds limit distributions for $\sqrt{n}\{Q_n(q) - F^{-1}(q)\}$. sample quantiles

(a) Suppose $U \sim \text{unif}$ and let $Y = F^{-1}(U)$. Show that Y has distribution F , and hence density f .

(b) Explain that the full order statistic vector $Y_{(1)} < \dots < Y_{(n)}$ may be represented via a correspondingly ordered sample of the uniform, as $F^{-1}(U_{(1)}) < \dots < F^{-1}(U_{(n)})$, with $U_{(i)}$ being the i th ordered observations in an i.i.d. sample $U_1, \dots, U_n$ from the uniform, studied in Ex. 3.17. In particular, $Y_{(i)}$ has the same distribution as $F^{-1}(U_{(i)})$.

(c) This also means that if we work out basic approximation results for the order statistics from the uniform, we are a modest delta method step away from similar results for the general case of a density f . In particular, suppose we manage to show $\sqrt{n}(U_{([nq])} - q) \rightarrow_d Z_q$, for some Z_q . Show that this implies $\sqrt{n}\{Q_n(q) - F^{-1}(q)\} \rightarrow_d c(q)Z_q$, with $c(q) = (F^{-1})'(q) = 1/f(\mu_q)$, writing $\mu_q = F^{-1}(q)$.

(d) To illustrate this point in a simple case first, show from what we already know in Ex. 3.20 that $\sqrt{n}(U_{([0.50n])} - 0.50) \rightarrow_d N(0, 0.50^2)$, for the uniform median. Then show for a general density f that $\sqrt{n}\{Q_n(0.50) - \mu\} \rightarrow_d N(0, 0.50^2/f(\mu)^2)$, with $\mu = F^{-1}(0.50)$ the population median. Here we used an exact expression for the density of the median.

(e) There are actually several other ways of proving this median of uniforms result. Such an alternative approach is to use the representation (3.5). Explain that $U_{([0.50n])} = A_n/(A_n + B_n)$, with A_n and B_n the averages of the first and second half of i.i.d. variables $V_1, \dots, V_{n+1}$ from the unit exponential. Use the CLT for the joint limit distributions of $\sqrt{n}(A_n - \frac{1}{2})$ and $\sqrt{n}(B_n - \frac{1}{2})$, and then use the delta method to land the $N(0, 0.50^2)$ limit.

(f) Then generalise to the case of any given quantile level q . Show first that $\sqrt{n}(U_{([nq])} - q) \rightarrow_d N(0, q(1-q))$, and then that the limit distribution is $N(0, q(1-q)/f(\mu_q)^2)$ for $\sqrt{n}\{Q_n(q) - \mu_q\}$.

(g) For the case of two quantiles jointly, like the lower and upper sample quartiles, show for the uniform case that with $q_1 < q_2$,

$$\begin{pmatrix} \sqrt{n}\{Q_n(q_1) - q_1\} \\ \sqrt{n}\{Q_n(q_2) - q_2\} \end{pmatrix} \rightarrow_d N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} q_1(1-q_1) & q_1(1-q_2) \\ q_1(1-q_2) & q_2(1-q_2) \end{pmatrix}\right).$$

Prove this via the explicit density for $(U_{(i)}, U_{(j)})$, given in Ex. 3.17, using the Scheffé Lemma of Ex. 2.6.

(h) It is also instructive to use representation (3.5) via i.i.d. unit exponentials. Prove the joint binormal limit distribution statement from this. Then generalise to the case of r quantiles, for levels $q_1 < \dots < q_r$. Show for the uniform case that there is a joint multivariate normal limit, with variances $q_j(1 - q_j)$ and covariances $q_j(1 - q_k)$ for $j < k$. Then carry out the transformation arguments needed to prove that for the case of an underlying positive density f , there is limiting joint normality for $\sqrt{n}(Q_{n,j} - \mu_j)$, where the limit has variances $q_j(1 - q_j)/f(\mu_j)^2$ and covariances $q_j(1 - q_k)/\{f(\mu_j)f(\mu_k)\}$ for $j < k$. See Ex. 10.31 for convergence of a full quantile process.

(i) So how variable are sample quantiles, how much bigger are variances out in the tails, compared to in the middle? Explain that the limiting standard deviation at quantile level q is $\text{sd}(q) = c(q)\{q(1 - q)\}^{1/2} = \{q(1 - q)\}^{1/2}/f(\mu_q)$. For a few densities, including the uniform, the normal, the exponential, compute and plot this $\text{sd}(q)$.

Ex. 3.22 Good and bad estimators. Suppose $X_1, \dots, X_n$ are i.i.d. from the density $f(x, \theta) = \exp\{-(x - \theta)\}$ for $y \geq \theta$, i.e. a unit exponential starting at parameter θ .

(a) Explain that we have $X_i = \theta + Y_i$, with the Y_i being i.i.d. from the unit exponential, and hence that the order statistics can be represented as $X_{(i)} = \theta + Y_{(i)}$, cf. Ex. 1.15.

(b) For the smallest and largest observations, show that $\hat{\theta}_A = X_{(1)} - 1/n$ and $\hat{\theta}_B = X_{(n)} - H_n$ are unbiased estimators of θ , with $H_n = 1 + 1/2 + \dots + 1/n$ the harmonic partial sum. Find their variances. Show that $\hat{\theta}_A$ is consistent, but that $\hat{\theta}_B$ is not.

Ex. 3.23 Smooth functions of quantiles. Suppose $Y_1, \dots, Y_n$ are i.i.d. from some distribution with continuous and positive density f . The inverse c.d.f. F^{-1} is then well-defined.

(a) Consider some smooth function $B(\mu_1, \dots, \mu_r)$ of the quantiles $\mu_j = F^{-1}(q_j)$. Via Ex. 3.21, identify first the multinormal limit $X \sim N_r(0, \Sigma)$ of the vector with components $\sqrt{n}(Q_{n,j} - \mu_j)$, with $Q_{n,j} = F_n^{-1}(\mu_j)$ the sample quantiles. Then establish that with $\hat{B} = B(Q_{n,1}, \dots, Q_{n,r})$, we have $\sqrt{n}(\hat{B} - B) \rightarrow_d c^t X \sim N(0, \tau^2)$, with $\tau^2 = c^t \Sigma c$. Here c is the vector of $c_j = \partial B(\mu_1, \dots, \mu_r)/\partial \mu_j$.

(b) Suppose $Y_1, \dots, Y_n$ are independent unit exponentials. With $p < q$, find the limit distribution of the ratio $\hat{\gamma} = F_n^{-1}(q)/F_n^{-1}(p)$. (xx check: we do have an exercise with quantiles and $F_n^{-1}(1 - \varepsilon) - F_n^{-1}(\varepsilon)$ for symmetric distributions? xx)

(c) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the Weibull distribution with c.d.f. $F(y) = 1 - \exp(-y^b)$. Show that the quantile function is $\mu_p = F^{-1}(p) = d(p)^{1/b}$, with $d(p) = -\log(1 - p)$. With $p < q$, find the limit distribution of

$$\sqrt{n} \left[\frac{Q_{n,q}}{Q_{n,p}} - \left\{ \frac{d(q)}{d(p)} \right\}^{1/b} \right].$$

Argue that with $R_n = Q_{n,q}/Q_{n,p}$, the solution $\hat{b}$ to $\log R_n = (1/b) \log\{d(q)/d(p)\}$ is a consistent estimator. Find the limit distribution of $\sqrt{n}(\hat{b} - b)$.

(d) Consider quantiles levels $p < q < r$ for i.i.d. observations $U_1, \dots, U_n$ from the uniform distribution on the unit interval. For empirical quantiles $Q_{n,p}, Q_{n,q}, Q_{n,r}$, explain via Ex. 3.21 that

$$\begin{pmatrix} \sqrt{n}(Q_{n,p} - p) \\ \sqrt{n}(Q_{n,q} - q) \\ \sqrt{n}(Q_{n,r} - r) \end{pmatrix} \rightarrow_d \begin{pmatrix} A \\ B \\ C \end{pmatrix} \sim N_3(0, \Sigma),$$

and identify the elements of Σ . From the empirical quantiles, form the two successive

ratios $R_n = Q_{n,q}/Q_{n,p}$ and $S_n = Q_{n,r}/Q_{n,q}$. Show that

$$\begin{pmatrix} \sqrt{n}(R_n - q/p) \\ \sqrt{n}(S_n - r/q) \end{pmatrix} \rightarrow_d \begin{pmatrix} R \\ S \end{pmatrix} = \begin{pmatrix} (1/p)\{-(q/p)A + B\} \\ (1/q)\{-(r/q)B + C\} \end{pmatrix}.$$

Do the algebra to verify that R and S have variances $(1/p)^2\{p(1 - q/p)^2 + q - p\}$ and $(1/q)^2\{q(1 - r/q)^2 + r - q\}$, and that they are in fact independent.

(e) A little factoid about the Cauchy: for R_n the interquartile range $Q_{n,0.75} - Q_{n,0.25}$, from an i.i.d. sample from that distribution, show that $\sqrt{n}(R_n - 2) \rightarrow_d N(0, \pi^2)$.

Ex. 3.24 Which order statistics interval contains the true median? Let $Y_1, \dots, Y_n$ be i.i.d. from a positive and smooth density f , with cumulative F . With $Y_{(1)} < \dots < Y_{(n)}$ the order statistics, which of the subintervals $(Y_{(i)}, Y_{(i+1)})$ will contain the true median, $\mu = F^{-1}(\frac{1}{2})$?

(a) Show that $p_i = \Pr\{\mu \in (Y_{(i)}, Y_{(i+1)})\} = \Pr\{\frac{1}{2} \in (U_{(i)}, U_{(i+1)})\}$, in terms of the order statistics from a uniform sample. We allow $i = 0, 1, \dots, n$, here, for the $n+1$ possibilities for which interval shall contain μ , writing $u_{(0)} = 0$ and $u_{(n+1)} = 1$.

(b) Given $U_{(i)} = u$, show that the distribution of $U_{(i+1)}$ is the same as that of $u + (1 - u)W$, where W is the smallest of $n - i$ observations from the uniform in the unit interval. Use this to show that

$$\begin{aligned} p_i &= \Pr\{U_{(i)} < \tfrac{1}{2} < U_{(i+1)}\} = \int_0^{1/2} \Pr\{U_{(i+1)} > \tfrac{1}{2} \mid U_{(i)} = u\} g_i(u) du \\ &= \int_0^{1/2} \left(\frac{\frac{1}{2}}{1-u}\right)^{n-i} \text{be}(u, i, n-i+1) du, \end{aligned}$$

involving a Beta density, as per Ex. 3.21. Show that this indeed leads to the explicit probability

$$p_i = \left(\frac{1}{2}\right)^{n-i} \frac{n!}{(i-1)!(n-i)!} \int_0^{1/2} u^{i-1} du = \binom{n}{i} \left(\frac{1}{2}\right)^n.$$

Hence we've reached the binomial probabilities, for a $\text{binom}(n, \frac{1}{2})$, via direct probability calculations. Try also to give a direct argument.

(c) Extend the above result to a general quantile $\mu_p = F^{-1}(p)$.

Moment matching methods

Ex. 3.25 *Moment matching estimators.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from some model $f(y, \theta)$, where $\theta = (\theta_1, \dots, \theta_p)^t$ is of dimension p . The *method of moments* consists in fitting the first p empirical moments to the theoretical ones. In detail, one computes

method of moments

$$M_1 = \bar{Y} = n^{-1} \sum_{i=1}^n Y_i, M_2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \dots, M_p = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^p,$$

and solves the p equations $M_1 = g_1(\theta), \dots, M_p = g_p(\theta)$, where $g_1(\theta) = E_\theta Y$, $g_2(\theta) = E_\theta \{Y - g_1(\theta)\}^2$, up to $g_p(\theta) = E_\theta \{Y - g_1(\theta)\}^p$.

(a) For one-parameter models, explain that this amounts to fitting the empirical and theoretical mean. If $Y_1, \dots, Y_n$ are i.i.d. $\text{geom}(p)$, see Ex. 1.25, use $E Y_i = 1/p$ to find the method of moments estimator for p . For another application, assume $Y_1, \dots, Y_n$ follow the distribution with c.d.f. y^θ on $[0, 1]$. Find the method of moments estimator for θ .

(b) For two-parameter models, explain that the method of moments means fitting the empirical mean and variance to the theoretical ones. If $Y_1, \dots, Y_n$ are i.i.d. $\text{Beta}(a, b)$, see Ex. 1.19, find expressions for the method of moments estimators for a, b .

(c) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the uniform distribution on $[0, b]$. Find the moment estimator for b . Then assume that the data are from a uniform on $[a, b]$, with both parameters unknown. Find the moment estimators.

(d) Define now $h_j(\theta) = E_\theta Y^j$ and $N_j = n^{-1} \sum_{i=1}^n Y_i^j$, for $j = 1, \dots, p$. The ‘method of direct moments’ is to solve the equations $h_j(\theta) = N_j$ for $j = 1, \dots, p$. For $p = 3$, set up the two systems of three equations with three unknowns, i.e. $g_j(\theta) = M_j$ and $h_j(\theta) = N_j$ for $j = 1, 2, 3$, and show that the solutions $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ are the same. Show in fact that the two methods, fitting centralised or direct methods, are identical, in the general case, for $p \geq 2$.

(e) For a given application one may choose one’s method based on convenience and practicality. Sometimes formulae for the direct moments are more easily found than those for the centralised moments. The case for using centralised moments is partly numerical safety; the M_j numbers may be much smaller than the N_j . Exemplify this by setting up the two equations, with two unknowns, both for the centralised moments and for the direct moments, for the case of the $\text{Beta}(a, b)$ distribution.

(f) Generate $n = 100$ data points via the equation $y_i = [\exp\{a(\xi + \sigma N_i)\} - 1]/a$, for say $(a, \xi, \sigma) = (0.33, 0.55, 0.77)$, with the N_i being standard normal. This is a skewed extension of the usual normal model, which corresponds to $a \rightarrow 0$ here. Find formulae for the first three moments for this distribution. From your data, use the method of moments to estimate the three parameters.

Ex. 3.26 *Moment fitting estimators for the Gamma distribution.* We now apply the moment matching principle of Ex. 3.25 to the Gamma model with parameters (a, b) , with density proportional to $y^{a-1} \exp(-by)$ for y positive; see Ex. 1.9, where we also give the mean, variance, skewness, kurtosis.

(a) With $\bar{Y}$ and V_n the usual sample mean and sample variance, find explicit formulae for the moment estimators $\hat{a}, \hat{b}$; show that $\hat{a} = \bar{Y}^2/V_n$ and $\hat{b} = \bar{Y}/V_n$.

(b) Use skewness and kurtosis formulae, in combination with Ex. 2.47, to show that

$$\begin{pmatrix} \sqrt{n}(\bar{Y} - a/b) \\ \sqrt{n}(V_n - a/b^2) \end{pmatrix} \rightarrow_d \begin{pmatrix} U \\ W \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} a/b^2 & 2a/b^3 \\ 2a/b^3 & (2 + 6/a)a^2/b^4 \end{pmatrix}\right).$$

Transform this, via the delta method, to find the limit distribution for the moment estimators.

Ex. 3.27 *Moment method estimation in a symmetric Beta.* Consider an i.i.d. dataset $Y_1, \dots, Y_n$ from the symmetric $\text{Beta}(a, a)$, with sample mean $\bar{Y}$ and sample variance $V_n = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$.

(a) Show that Y_i has mean $\frac{1}{2}$ and variance $\sigma^2 = \frac{1}{4}(2a+1)$. Explain that the moment method with the sample mean does not work in this situation, but that the moment method using the sample variance leads to $\hat{a} = 1/(8V_n) - \frac{1}{2}$. Show that it is consistent.

(b) Via patient integration, show that the kurtosis is

$$\text{kurt} = \int_0^1 \frac{(x - \frac{1}{2})^4}{\sigma^4} \frac{\Gamma(2a)}{\Gamma(a)^2} \{x(1-x)\}^{a-1} dx - 3 = -\frac{6}{2a+3}.$$

Then use Ex. 2.48 to show that $\sqrt{n}(V_n - \sigma^2) \rightarrow_d N(0, \tau^2)$, with

$$\tau^2 = \sigma^4(2 + \text{kurt}) = \frac{1/16}{(2a+1)^2} \frac{4a}{2a+3}.$$

(c) Use the delta method to find the limiting distribution of $\sqrt{n}(\hat{a} - a)$, and explain how this may be used to construct a 90 percent confidence interval for a . (xx later: compare variance to that of the ML method, with efficiency loss. check if the following is ok. xx)

$$\kappa^2 = (2a+1)^2 a / (2a+3).$$

Ex. 3.28 *Moment method estimators for the exponential family.* Consider an exponential family type model, studied in Ex. 1.57, with density of the form $f(y, \theta) = \exp\{\theta^t T(y) - k(\theta)\} h(y)$. Here $T(y) = (T_1(y), \dots, T_p(y))^t$ is a collection of data functions. As we saw in the exercise pointed to, many classical models are special cases. Now suppose $Y_1, \dots, Y_n$ are i.i.d. from such a model.

(a) Spell out the basic moment matching method, fitting empirical and model based moments for $Y_i, Y_i^2, \dots, Y_i^p$.

(b) The moment matching idea is however flexible enough to allow us to choose other data functions than those associated with $Y_i, \dots, Y_i^p$. Show that matching the moments of $T_1(Y_i), \dots, T_p(Y_i)$ corresponds to solving the equations

$$\bar{T}_j = \xi_j(\theta) = \partial k(\theta) / \partial \theta_j \quad \text{for } j = 1, \dots, p,$$

with $\bar{T}_j = n^{-1} \sum_{i=1}^n T_j(Y_i)$.

(c) Consider the setup of Ex. 3.26, with the $\text{Gam}(a, b)$ distribution. There we studied moment matching estimators based on sample mean $\bar{Y}$ and sample variance V_n . Show that the principle above, using the exponential family structure, leads to fitting $\bar{Y} = a/b$ and $n^{-1} \sum_{i=1}^n \log Y_i = \psi(a) - \log b$. We shall learn via likelihood theory of Ch. 5 that these are more precise than the moment based ones, under model conditions.

Ex. 3.29 *Mean and variance for background distribution for random sums.* Suppose there is a hidden background machine drawing first the number of dice N and then reporting the sum $Z = X_1 + \dots + X_N$ of outcomes from having thrown those N dice (without reporting N).

(a) For the individual random X_i , show that the mean and variance are $\xi = 3.50$ and $\sigma^2 = 35/12 = 2.9167$. Use results from Ex. 1.40 to find formulae for the mean and variance of Z .

(b) If $Z_1, \dots, Z_m$ are observed, from such a two-layer-random machine, with mean 33.33 and standard deviation 12.12, estimate the mean and standard deviation for N .

Quantile matching methods

Ex. 3.30 *Quantile fitting estimators.* An alternative to the method of moments, described in Ex. 3.25, is to fit empirical and theoretical quantiles (actually in several ways). If $Y_1, \dots, Y_n$ are i.i.d. from a density $f(y, \theta)$, for a parameter vector of length p , with quantiles $Q(r, \theta) = F^{-1}(r, \theta)$, we choose quantile levels $r_1 < \dots < r_p$, and solve the p equations $Q_n(r_j) = Q(r_j, \theta)$ with respect to the p unknown parameters, where $Q_n(r) = F_n^{-1}(r)$ is the empirical quantile.

method of
quantiles

(a) Suppose the distribution to be fitted has c.d.f. $F(y) = y^\theta$ on the unit interval. Find the estimator corresponding to fitting the empirical to the theoretical median. Starting with the limit distribution for the median, see Ex. 3.20, find the limit distribution for $\sqrt{n}(\hat{\theta} - \theta)$. More generally, find the estimator $\hat{\theta}_r$ corresponding to fitting the r level quantile, and then the limit distribution for $\sqrt{n}(\hat{\theta}_r - \theta)$. For large n , which value of r gives the most precise estimator?

(b) Consider $Y_1, \dots, Y_n$ from the location Cauchy density $f_0(y - \theta)$, with $f_0(x) = (1/\pi)/(1+x^2)$ the standard Cauchy. Its c.d.f. is $F_0(x) = \frac{1}{2} + (1/\pi) \arctan x$, see Ex. 1.18. Show that the r level quantile is $\mu + F_0^{-1}(r)$, and that this leads to the estimator $\hat{\mu}_r = Q_n(r) - F_0^{-1}(r)$. Find the limit distribution for $\sqrt{n}(\hat{\mu}_r - \mu)$. What quantile level r leads to the sharpest estimator?

(c) For data $Y_1, \dots, Y_n$ from a normal (ξ, σ^2) , find estimators for the two parameters using the median and interquartile range. Find the limit distributions. More generally, construct an estimator $\hat{\sigma}_r$ for σ based on the quantile difference $Q_n(1-r) - Q_n(r)$. Find the limit distribution and identify the best value of r .

Ex. 3.31 *Moment fitting and quantile fitting for the Weibull.* As a general illustration of moment and quantile fitting estimation methods, consider the Weibull distribution with c.d.f. $F(t) = 1 - \exp\{-(t/a)^b\}$ for $t \geq 0$, see Ex. 1.61.

(a) Take e.g. $(a, b) = (3.33, 1.44)$, and simulate $n = 100$ realisations. (i) Compute average and standard deviation for these, and compute estimates $(\hat{a}_m, \hat{b}_m)$. (ii) From 0.25 and 0.75 quantiles, fit the two relevant equations to compute $(\hat{a}_q, \hat{b}_q)$. Take the trouble to display three curves, the correct underlying cumulative hazard function $A(t) = (t/a)^b$ along with the two estimated versions.

(b) Repeat the experiment many times, to see how close the two $(\hat{a}, \hat{b})$ is to (a, b) . Also, as an instance of a focused question, check how close the two median estimates $\hat{m} = \hat{a}(\log 2)^{1/\hat{b}}$ are to the real median. Which of the two estimating schemes is best?

both methods
will lose to
maximum
likelihood, as
we learn in
Ch. 5

Minimum sum of squares and linear regression

Ex. 3.32 *Linear regression and least squares estimation.* Consider observed pairs (x_i, Y_i) for $i = 1, \dots, n$, where Y_i conditionally on covariate x_i is modelled to have mean $a_0 + bx_i$ and common variance σ^2 . This is classical linear regression, widely used in theoretical and applied statistics, most often analysed and used with the additional assumption that the Y_i are normally distributed. Importantly, the model is extended to the case of multiple covariates below, see Ex. 3.33, 3.34, with more to come regarding statistical testing, see e.g. Ex. 4.37.

(a) It is helpful to reparametrise the regression line from $a_0 + bx_i$ to $a + b(x_i - \bar{x})$. Show that minimising the sum of squares $Q(a, b) = \sum_{i=1}^n \{y_i - a - b(x_i - \bar{x})\}^2$ leads to

$$\hat{a} = (1/n) \sum_{i=1}^n y_i = \bar{y}, \quad \hat{b} = \sum_{i=1}^n (x_i - \bar{x}) y_i / M_n, \quad \text{with } M_n = \sum_{i=1}^n (x_i - \bar{x})^2.$$

(b) Show that $\hat{a}$ and $\hat{b}$ are unbiased, with zero covariance, and variances σ^2/n and σ^2/M_n .

(c) Let $Q_0 = \min_{a,b} Q(a, b) = \sum_{i=1}^n \{y_i - \hat{a} - \hat{b}(x_i - \bar{x})\}^2$ be the minimum sum of squares.

check Stories
#35 and #55

Show that

$$Q(a, b) = \sum_{i=1}^n \{y_i - a - b(x_i - \bar{x})\}^2 = Q_0 + n(\hat{a} - a)^2 + M_n(\hat{b} - b)^2.$$

Use this to show that $\hat{\sigma}^2 = Q_0/(n-2)$ is an unbiased estimator of σ^2 .

(d) We have found natural and unbiased estimators for a, b, σ^2 , without yet making assumptions of the underlying distributions for Y_i , beyond means and variance. Assume now, however, that the distributions are normal, so that $Y_i \sim N(a + b(x_i - \bar{x}), \sigma^2)$. Show that (i) $\hat{a} \sim N(a, \sigma^2/n)$; (ii) $\hat{b} \sim N(b, \sigma^2/M_n)$; (iii) $\hat{\sigma}^2 \sim \sigma^2 \chi_m^2/m$, with $m = n-2$; and (iv) that these three statistics are mutually independent. For this last point, it may be helpful to follow the line of proof for independence of sample mean and sample variance for normal data, used in Ex. 1.49, now with an orthonormal matrix A with first row $(1/\sqrt{n}, \dots, 1/\sqrt{n})$ and second row $((x_1 - \bar{x})/M_n^{1/2}, \dots, (x_n - \bar{x})/M_n^{1/2})$.

(e) Construct confidence intervals for b , for σ , and for $E(Y | x_0) = a + b(x_0 - \bar{x})$, the mean value at a given position x_0 .

Ex. 3.33 *Linear multiple regression and least squares.* The celebrated linear multiple regression model remains a cornerstone success story of theoretical and applied statistics. It is a bag of tools for investigating the extent to which covariates x influence the outcomes of certain interest variables Y . The standard formulation of the model is as follows. The data collected can be organised into (x_i, Y_i) , for individuals or objects $i = 1, \dots, n$, where $x_i = (x_{i,1}, \dots, x_{i,p})^t$ is of dimension p and y_i of dimension one. The model then postulates that

$$Y_i = x_i^t \beta + \varepsilon_i = x_{i,1} \beta_1 + \dots + x_{i,p} \beta_p + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

where the ε_i are i.i.d. from the normal $N(0, \sigma^2)$. Thus there are $p+1$ parameters at work here, the regression coefficients $\beta = (\beta_1, \dots, \beta_p)^t$ and the error distribution standard deviation σ . – Note that the very classical case of $y_i = a + bx_i + \varepsilon_i$, associated with a scatterplot of (x_i, Y_i) , is a special case; see Ex. 3.32.

(a) With Y the vector of Y_i , ε the vector of ε_i , and X the $n \times p$ matrix having $(x_{i,1}, \dots, x_{i,p})$ as its row i , show that

$$Y = X\beta + \varepsilon \sim N_n(X\beta, \sigma^2 I),$$

with I the $n \times n$ identity matrix. This is a practical and compact linear algebra version of the model formulation. We do assume that X is of full rank p , so that the symmetric matrix $X^t X$ is invertible. This amounts to there being at least p linearly independent covariate vectors in the X matrix; in particular, we must have $n \geq p$ to identify the β_j coefficients directly from data.

(b) The least squares estimator $\hat{\beta}$ is the minimiser of $Q(\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^n (Y_i - x_i^t \beta)^2$. Show that $\sum_{i=1}^n (Y_i - x_i^t \hat{\beta}) x_i = 0$. With Σ_n the $p \times p$ matrix $n^{-1} \sum_{i=1}^n x_i x_i^t = n^{-1} X^t X$, show that

$$\hat{\beta} = (X^t X)^{-1} X^t Y = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i Y_i.$$

Prove also that $Q(\beta) = Q_0 + n(\hat{\beta} - \beta)^t \Sigma_n (\hat{\beta} - \beta)$, with $Q_0 = \min_{\text{all } \beta} Q(\beta) = \sum_{i=1}^n \hat{\varepsilon}_i^2$, writing $\hat{\varepsilon}_i = Y_i - x_i^t \hat{\beta}$ for the estimated residuals.

(c) Show that $\hat{\beta}$ is unbiased and that its variance matrix can be written $\sigma^2(X^t X)^{-1} = (\sigma^2/n)\Sigma_n^{-1}$.

(d) Show also that $\hat{\beta}$ has a multinormal distribution, so that in fact $\hat{\beta} \sim N_p(\beta, (\sigma^2/n)\Sigma_n^{-1})$. This is the key result about the least squares estimators. We also need precise information for estimating σ ; see Ex. 3.34.

Ex. 3.34 *The residuals and their variances.* The setup is as in the previous Ex. 3.33, the $Y \sim N_n(X\beta, \sigma^2 I)$ linear regression model. Above we focused on the least squared method and the ensuing properties for the estimators of the regression coefficients, and found $\hat{\beta} \sim N_p(\beta, (\sigma^2/n)\Sigma_n^{-1})$. We also need to deal carefully with estimators of σ , the residual standard deviation, also since we encounter statistics of the type $(\hat{\beta}_j - \beta_j)/\hat{\sigma}$.

(a) From the basic $Y = X\beta + \varepsilon$ we may define the estimated residuals as

$$\hat{\varepsilon} = Y - X\hat{\beta} = (I - H)Y, \quad \text{where} \quad H = X(X^t X)^{-1}X^t,$$

the hat matrix

the so-called hat matrix, of size $n \times n$. Show that H is symmetric and idempotent, which means that $H^t = H$ and $H^2 = H$. This also implies $(I - H)H = 0$.

(b) Now consider the random minimum achieved by the $Q(\beta)$ which was used in the least squares operation,

$$Q_0 = \min\{Q(\beta) : \text{all } \beta\} = Q(\hat{\beta}) = \|Y - X\hat{\beta}\|^2 = \sum_{i=1}^n (Y_i - x_i^t \hat{\beta})^2.$$

The main result, arrived at below, is that $Q_0/\sigma^2 \sim \chi_m^2$, with degrees of freedom $m = n - p$, and that Q_0 is independent of $\hat{\beta}$. Show first that

$$\begin{pmatrix} X\hat{\beta} \\ \hat{\varepsilon} \end{pmatrix} = \begin{pmatrix} HY \\ (I - H)Y \end{pmatrix} \sim N_{2n}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I - H \end{pmatrix}\right).$$

In particular, these two random vectors are independent; also, $Q_0 = \|\hat{\varepsilon}\|^2 = Y^t(I - H)Y$ is consequently independent of $X\hat{\beta}$.

(c) Show that $(I - H)X = 0$, which implies $\hat{\varepsilon} = (I - H)Y = (I - H)(Y - X\beta) = (I - H)\varepsilon$ and hence $Q_0 = \varepsilon^t(I - H)\varepsilon$. We also reach the simple identity

$$\|\varepsilon\|^2 = \varepsilon^t H \varepsilon + \varepsilon^t (I - H) \varepsilon,$$

where the left-hand side is a $\sigma^2 \chi_n^2$ and the two terms on the right-hand side being independent. Show that the first term on the right-hand side is a $\sigma^2 \chi_p^2$. Via independence and a moment-generating function argument show then that $Q_0 \sim \sigma^2 \chi_{n-p}^2$ (this is essentially the basic trick of Ex. 1.54).

(d) With $Q_0 \sim \sigma^2 \chi_m^2$, with $\text{df} = m = n - p$, the natural estimator is $\hat{\sigma}^2 = Q_0/m$. Show that it is unbiased and find its variance. Note that things we have learned in Ex. 1.49 and other exercises related to the empirical variance for a normal sample carry over to the present linear regression model, simply with $n - p$ instead of $n - 1$.

Ex. 3.35 *Separate and joint confidence intervals.* Suppose a vector Y is observed from some $N_p(\theta, \Sigma)$, with unknown θ and known positive definite Σ . How can we set confidence intervals for the individual θ_j , for given linear combinations, and then for all linear combinations jointly?

(a) Consider a given linear combination $\gamma = c^t\theta = c_1\theta_1 + \cdots + c_p\theta_p$. Explain that $\hat{\gamma} = c^tY \sim N(\gamma, c^t\Sigma c)$, and that the canonical 95 percent confidence interval for γ is $\hat{\gamma} \pm 1.96\kappa(c)$, with $\kappa(c) = (c^t\Sigma c)^{1/2}$. Explain what this entails for the components $\theta_1, \dots, \theta_p$.

(b) Leaving the situation above aside for a minute, prove the following mathematical fact: for given x , and a given symmetric positive definite $p \times p$ matrix A , the maximum of $r(c) = c^tx / (c^tAc)^{1/2}$, taken over all c , is equal to $(x^tA^{-1}x)^{1/2}$. Do this by first seeing that the ratio $r(c)$ does not depend on the length of c , so the problem is equivalent to finding the maximum of c^tx over all c such that $c^tAc = 1$; you may proceed using Lagrange multipliers and a bit of linear algebra.

(c) Suppose now that one wishes to have intervals for *all* linear combinations, with confidence level 95 percent to hold jointly. To what extent does one need to stretch the $c^tY \pm 1.96\kappa(c)$ intervals, to secure such a joint-for-all 95 percent confidence? Show that

$$V = \max_{\text{all } c} \frac{|c^tY - c^t\theta|}{(c^t\Sigma c)^{1/2}} = \{(Y - \theta)^t\Sigma^{-1}(Y - \theta)\}^{1/2},$$

and that $V^2 \sim \chi_p^2$. Explain then that with $z_p = \Gamma_p^{-1}(0.95)^{1/2}$, involving the upper 0.05 quantile of the χ_p^2 , the intervals $I(c) = c^tY \pm z_p\kappa(c)$ have the desired property, that $\Pr\{c^t\theta \in I(c) \text{ for all } c\} = 0.95$. Check how this z_p factor grows with p . For $p = 5$, for example, the separate confidence intervals need to be made 1.697 times wider, in order to secure that the confidence level is 95 percent for all linear combinations.

Ex. 3.36 *Confidence intervals for key parameters in linear regression models.* Consider the general linear regression setup of Ex. 3.33–3.34, where $Y_i = x_i^t\beta + \varepsilon_i$ for $i = 1, \dots, n$, and the ε_i being i.i.d. $N(0, \sigma^2)$. The compact version of this is $Y \sim N_n(X\beta, \sigma^2 I_n)$. We use notation from these exercises below, with $\hat{\beta}$, $\hat{\sigma}^2 = Q_0/(n-p)$, etc.

(a) Construct a 90 percent confidence interval for σ .

(b) From the general results obtained above, show that $\hat{\beta}_j \sim N(\beta_j, \sigma^2 r_j/n)$, with r_j the diagonal (j, j) element of Σ_n^{-1} . Show from this that $t_j = \sqrt{n}(\hat{\beta}_j - \beta_j)/(r_j^{1/2}\hat{\sigma})$ has a t_m distribution, with degrees of freedom $m = n - p$, and construct a confidence interval for β_j from this. If a 99 percent confidence interval for β_j is outside zero, how can this be interpreted and used?

(c) More generally, consider $\gamma = c^t\beta = c_1\beta_1 + \cdots + c_p\beta_p$, a linear combination of the regression coefficients. Find the distribution of $\hat{\gamma} = c^t\hat{\beta}$, and show that the interval $c^t\hat{\beta} \pm t_{m,0.975}\hat{\kappa}(c)/\sqrt{n}$ covers the true $c^t\beta$ with probability 95 percent, when $\hat{\kappa}(c) = (c^t\Sigma_n^{-1}c)^{1/2}\hat{\sigma}$, with $t_{m,0.975}$ the upper 0.025 quantile of the t_m , with $m = n - p$ degrees of freedom. This may be used to assess the size of $\beta_1 - \beta_2$, of $\beta_1 - \frac{1}{2}(\beta_2 + \beta_3)$, and similar contrasts.

(d) The situation is different if one wishes to have confidence intervals for *all* linear combinations, jointly, with a guarantee that the overall coverage is e.g. 95 percent. Show using Ex. 3.35 that

$$V_n = \max_{\text{all } c} \frac{\sqrt{n}|c^t\hat{\beta} - c^t\beta|}{(c^t\Sigma_n^{-1}c)^{1/2}\hat{\sigma}} = \{n(\hat{\beta} - \beta)^t\Sigma_n(\hat{\beta} - \beta)\}^{1/2}/\hat{\sigma},$$

and that $V_n^2 \sim F_{p,m}$. Use this to show that the intervals $I_n(c) = c^t\hat{\beta} \pm z_{p,m}\hat{\kappa}(c)/\sqrt{n}$, with $z_{p,m}$ the 0.975 quantile of the $F_{p,m}$ distribution, have the property that $\Pr\{c^t\beta \in I_n(c) \text{ for all } c\} = 0.95$.

(e) Use the result about V_n to construct a confidence region, actually a confidence ellipsoid, for the full β vector.

Ex. 3.37 *Predicting the next y.* Suppose linear regression analysis has been carried out, for a given dataset $(x_1, y_1), \dots, (x_n, y_n)$. How can we predict what happens with Y_0 , associated with another covariate vector x_0 ? This could be for an individual outside the dataset, or in a time context, speculating about the next datapoint in a sequence. An illustration of methods given below is in Story #35.

(a) Assume the regression data are of the form $y_i \sim N(x_i^t \beta, \sigma^2)$ for $i = 1, \dots, n$, as with Ex. 3.33, and consider a new x_0 , for which the not yet unobserved Y_0 is independent of the other data, and with distribution $N(x_0^t \beta, \sigma^2)$. Show that $\hat{y}_0 = x_0^t \hat{\beta} \sim N(x_0^t \beta, x_0^t \Sigma_n^{-1} x_0 \sigma^2/n)$, and use this to form a confidence interval for the mean of Y_0 , as opposed to for Y_0 itself.

(b) Then show that

$$Y_0 - \hat{y}_0 \sim N(0, \sigma^2(1 + n^{-1} x_0^t \Sigma_n^{-1} x_0/n)).$$

Construct a prediction confidence interval for Y_0 based on this. Comment on the situation with a large n , and on the difference between confidence for Y_0 and its mean.

(c) Then consider the classic linear regression case with only one variable studied in Ex. 3.32, with $Y_i \sim N(a + b(x_i - \bar{x}), \sigma^2)$. For predicting a not yet observed $Y_0 \sim N(a + b(x_0 - \bar{x}), \sigma^2)$, demonstrate that

$$Y_0 - \{\hat{a} + \hat{b}(x_0 - \bar{x})\} \sim N(0, \sigma^2\{1 + 1/n + (x_0 - \bar{x})^2/M_n\}),$$

with $M_n = \sum_{i=1}^n (x_i - \bar{x})^2$. Show that this leads to the 95 percent confidence band

$$Y(x_0) \in \hat{a} + \hat{b}(x_0 - \bar{x}) \pm c \hat{\sigma} \{1 + 1/n + (x_0 - \bar{x})^2/M_n\}^{1/2},$$

for an appropriate range of x_0 values, with $c = t_{0.975, n-2}$ the t upper quantile. Illustrate this with a simulated dataset, to see both where the band is tight and where it becomes very broad. One learns that it is easier to predict outcomes related to x_0 closer to the middle of the covariate distribution.

Ex. 3.38 *Inference for mean by standard deviation.* Consider $Y_1, \dots, Y_n$ being i.i.d. from $N(\xi, \sigma^2)$. Estimation and precise inference for ξ and for σ separately is classic terrain, via the t and the chi-squared distribution, and we have seen in Ex. 3.8 how we can also deal with quantiles $\xi + z_q \sigma$ in a precise fashion. For more complicated parameters $g(\xi, \sigma)$ we might turn to large-sample approximations, as per Ex. 2.53. This exercise concerns however precise inference for the parameter $\kappa = \xi/\sigma$.

(a) We start with the following cleanly stated problem: suppose $T \sim t_m(a)$, with the noncentral t , as per Ex. 1.51, with known m but unknown a . With t_{obs} the observed value of T , we form a confidence interval by collecting together all values a for which t_{obs} is not surprising, defined by $I = \{a: 0.05 \leq G_m(t_{\text{obs}}, a) \leq 0.95\}$. Show that indeed $\Pr_a(a \in I) = 0.90$. Again, the $G_m(t, a)$ is a somewhat complicated function, but is numerically available in packages.

(b) Coming back to the normal sample, write as usual $\hat{\xi} = \bar{Y}$ and $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$, and form $\hat{\kappa} = \hat{\xi}/\hat{\sigma}$. Show that $X = \sqrt{n} \hat{\kappa} \sim t_m(\sqrt{n} \kappa)$, with $m = n - 1$ degrees of freedom. Show that

$$I = \{\kappa: 0.05 \leq G_m(\sqrt{n} \hat{\kappa}, \sqrt{n} \kappa) \leq 0.95\}$$

is a 90 percent confidence interval for κ . In practice we find the interval by computing and monitoring the $G_m(\sqrt{n}\hat{\kappa}, \sqrt{n}\kappa)$, as a function of κ .

(c) Choose your favourite (ξ, σ) , generate a normal sample of size $n = 100$, and compute (i) the interval $\hat{\kappa} \pm 1.645(1 + \hat{\kappa}^2)^{1/2}/\sqrt{n}$, from Ex. 2.53, and (ii) the exact interval from the recipe above. Compare and comment.

(d) Consider now a linear regression setup, as in Ex. 3.33–3.34, with independent $y_i = x_i^t \beta + \varepsilon_i$ for $i = 1, \dots, n$. With some given linear combination $c^t \beta$ of the regression components, consider $\kappa = c^t \beta / \sigma$. In the standard notation of the exercises pointed to, show that $\hat{\kappa} = c^t \hat{\beta} / \hat{\sigma}$ is such that $t = \sqrt{n}\hat{\kappa} \sim t_m(\sqrt{n}\kappa)$. Explain how a precise 90 percent confidence interval can be set for κ .

(e) In such a linear regression setup, suppose x_a and x_b are certain covariate values, outside of $x_1, \dots, x_n$, associated with unknown or not yet observed $Y_a = x_a^t \beta + \varepsilon_a$ and $Y_b = x_b^t \beta + \varepsilon_b$. Show that

$$p = \Pr(Y_b > Y_a) = \Phi(\kappa/\sqrt{2}), \quad \text{with } \kappa = (x_b - x_a)^t \beta / \sigma,$$

Explain that you via a clear 90 percent confidence interval for κ also have a 90 percent confidence interval for p . (xx brief pointer to New Haven Story #35, where nils will put in estimate and interval for p , the probability that 1976 is warmer than 1907, based on data from 1912 to 1971. more on the details there, prediction vs. the mean change. xx)

Ex. 3.39 *Linear regression outside normality.* The aim here is to show and appreciate that the classical coefficient estimators in linear regression setups are still approximately normal, even when the error terms distribution is not normal. That this is so is essentially thanks to basic large-sample theory, as partly summarised in Ex. 3.11, and specifically to the Lindeberg type theorems of Ch. 2; see in particular Ex. 2.41.

(a) We first deal with a simple setup with a single regression coefficient. Suppose $y_i = x_i \beta + \varepsilon_i$ for $i = 1, \dots, n$, with covariates x_i and error terms ε_i being i.i.d. from a zero-mean distribution with finite variance σ^2 . Show that the estimator minimising $Q_n(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2$ is $\hat{\beta} = \sum_{i=1}^n x_i y_i / M_n$, where $M_n = \sum_{i=1}^n x_i^2$. Show further that $\hat{\beta}$ is unbiased with variance σ^2 / M_n .

(b) Then consider $Z_n = M_n^{1/2}(\hat{\beta} - \beta)$. Show that it has zero mean and variance σ^2 , and that it can be written $\sum_{i=1}^n (x_i / M_n^{1/2}) \varepsilon_i$. Deduce that $\hat{\beta}$ is approximately normal, even if the ε_i are not normal, provided merely that $D_n = \max_{i \leq n} |x_i| / M_n^{1/2} \rightarrow 0$. If in particular $(1/n) \sum_{i=1}^n x_i^2$ stays bounded, then the natural condition is $(1/\sqrt{n}) \max_{i \leq n} |x_i| \rightarrow 0$.

(c) Then consider the general linear regression model $Y_i = x_i^t \beta + \varepsilon_i$ of Ex. 3.33, with the x_i being p -dimensional covariate vectors and β a p -dimensional vector of regression coefficients. We take the ε_i to be i.i.d. with mean zero and finite variance σ^2 , but do not stipulate normality. The least squares estimator is $\hat{\beta} = \Sigma_n^{-1} (1/n) \sum_{i=1}^n x_i Y_i$ with $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t$. It is unbiased with variance matrix $(\sigma^2/n) \Sigma_n^{-1}$, assumed to have full rank. Assume $\Sigma_n \rightarrow \Sigma$, a full rank matrix. Show that $Z_n = \sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N_p(0, \Sigma)$, under the condition $R_n = (1/\sqrt{n}) \max_{i \leq n} \|x_i\| \rightarrow 0$.

(d) Assume now that the x_i are drawn i.i.d. from a distribution over the covariate space, with finite variance matrix. Show that $R_n \rightarrow_{\text{pr}} 0$.

(e) Argue that the fine-tuned finite-sample confidence methods, developed in Ex. 3.33–3.34 under exact normality, continue to hold in the large-sample sense, even if the distributions are not normal. Give conditions for such results to hold.

check also with
Ex. 2.41

(f) Suppose the same type of phenomenon is studied in both Denmark and Sweden, with the same meaning for Y and covariates x_1, x_2, x_3 . This leads to regression estimators $\hat{\beta}_D$ and $\hat{\beta}_S$ for the two analyses. Construct confidence intervals for $d_j = \beta_{D,j} - \beta_{S,j}$, for $j = 1, 2, 3$.

Ex. 3.40 *Least squares estimation in other setups.* Least squares is a general principle, also outside linear regression. With $E(Y_i | x_i) = \xi_i(\theta)$, say, one minimises $Q(\theta) = \sum_{i=1}^n \{y_i - \xi_i(\theta)\}^2$.

(a) Let $Y_1, \dots, Y_n$ be i.i.d. Bernoulli with $\Pr(Y_i = 1) = p$, Find the least squares estimator, and comment. Similarly, with $Y_i \sim N(0, \sigma^2)$, do least squares with $\sum_{i=1}^n (Y_i^2 - \sigma^2)^2$. Discuss the case of $Y_i \sim \text{Pois}(a \exp(bx_i))$.

(b) Assume $y_1, \dots, y_n$ are i.i.d. and gamma distributed (a, b) , with known a . Find the least squares estimator $\hat{b}$. Then find the mean and variance of this estimator.

Ex. 3.41 *Estimating with autocorrelated data.* Situations dealt with above have involved different models and questions, but have assumed independence between observations. In various setups, including those of time series, there is correlation, however, with consequences for estimators and their precision. The brief discussion below has implications for reading off estimates for key quantities in Markov Chain Monte Carlo settings in Bayesian analysis; see Ch. 7.

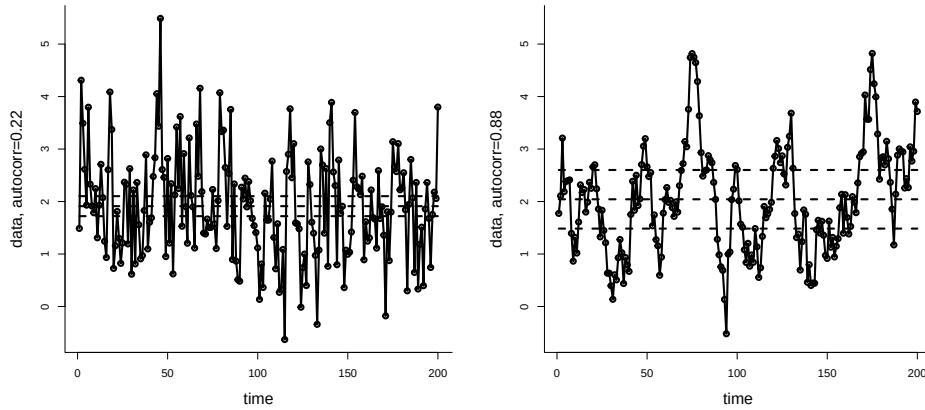


Figure 3.3: Data with autocorrelation, with underlying $\xi = 2.0$, $\sigma = 1.0$, and then $\rho = 0.22$ (left panel) and $\rho = 0.88$ (right panel). The horizontal lines are for the estimate and the 95 percent confidence intervals for ξ . With higher autocorrelation, data are more connected, and estimates have larger variances.

(a) Consider $V_1, \dots, V_n$ i.i.d. with mean zero and variance one. From these we form a new series, with autocorrelation $\rho \in (-1, 1)$, as follows. Let $W_1 = V_1$ and then $W_i = \rho W_{i-1} + (1 - \rho^2)^{1/2} V_i$ for $i = 2, \dots, n$. Show that these W_i have mean zero and variance one, but $\text{cov}(W_i, W_j) = \rho^{j-i}$ for $i < j$. A time series, with three crucial parameters mean ξ , variance σ^2 , and autocorrelation ρ , is then formed as $Y_i = \xi + \sigma W_i$.

(b) With such an autocorrelated time series $Y_1, \dots, Y_n$, the natural estimator of ξ is still the sample average $\hat{\xi} = \bar{Y}_n$, but its variance is affected. Show first in general terms that

one for any $a \in (-1, 1)$ has

$$n^{-1} \sum_{1 \leq i < j \leq n} a^{j-i} = \frac{a}{1-a} \left\{ 1 - \frac{1}{n} - \frac{1}{n} \frac{a}{1-a} (1 - a^{n-1}) \right\} = \frac{a}{1-a} \{1 + O(1/n)\}.$$

Show that this leads to

$$\text{Var } \bar{Y}_n = \frac{1}{n^2} \sum_{i,j} \text{cov}(Y_i, Y_j) = \frac{\sigma^2}{n} \left\{ 1 + 2 \frac{\rho}{1-\rho} + O(1/n) \right\} = \frac{\sigma^2}{n} \left\{ \frac{1+\rho}{1-\rho} + O(1/n) \right\}.$$

Conclude that $\bar{Y}_n$ is unbiased and consistent for ξ , but with this adjusted variance.

(c) We're not aiming here for maximum generality (which would take us into the terrain of stationary time series); the main point to convey is that positively autocorrelated data still allow consistent estimation and inference for key quantities, but with bigger variances, wider confidence intervals, etc. Assume therefore that the V_i above are normal, implying that $(W_1, \dots, W_n)$ and $(Y_1, \dots, Y_n)$ have multinormal distributions. In addition to the basic moment structure, with $E Y_i = \xi$, $\text{Var } Y_i = \sigma^2$, $\text{corr}(Y_i, Y_j) = \rho^{|j-i|}$, we can read off further variances and correlations, for Y_i^2 , for $Y_i Y_{i+1}$, etc. Show that $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is consistent for σ^2 , and that

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n-1} \frac{Y_i - \hat{\xi}}{\hat{\sigma}} \frac{Y_{i+1} - \hat{\xi}}{\hat{\sigma}}$$

is consistent for ρ . Show that this implies $\sqrt{n}(\bar{Y}_n - \xi)/\hat{g} \rightarrow_d N(0, \sigma^2)$, with factor $\hat{g}^2 = (1 + \hat{\rho})/(1 - \hat{\rho})$. Explain that this leads to these confidence intervals for ξ , adjusted for autocorrelation: $\hat{\xi} \pm 1.96 \hat{g} \hat{\sigma} / \sqrt{n}$. Construct a version of Figure 3.3, with $n = 200$ autocorrelated observations, with higher autocorrelation in the right panel, associated with wider confidence interval for the mean. Try also with e.g. $\rho = 0.99$, and comment.

(d) Statisticians A and B need to estimate a mean parameter ξ with a certain precision, from observations Y_i with this mean and variance σ^2 . A manages to get hold of i.i.d. data, and finds an estimate with a confidence interval after $n_A = 1000$ observations. B utilises a different setup for generating the data, and has to make do with a chain with autocorrelation $\rho = 0.90$. How much bigger does his sample size n_B need to be, in order to have the same precision as that obtained by his colleague?

Estimating a density

Ex. 3.42 *The kernel density estimator.* Suppose $X_1, \dots, X_n$ are i.i.d. from some smooth positive density f on the real line, with c.d.f. F . To estimate f , introduce a kernel function K , a density symmetric around zero, and for any bandwidth h write $K_h(u) = h^{-1} K(h^{-1}u)$. The kernel density estimator is

$$f_n(x) = n^{-1} \sum_{i=1}^n K_h(X_i - x).$$

Below we write $k_2 = \int u^2 K(u) du$ and $R(K) = \int K(u)^2 du$, both taken to be finite.

(a) Assuming that f has two smooth derivatives, show that the mean of $f_n(x)$ is

$$g_n(x) = \int K_h(y - x) f(y) dy = \int K(u) f(x + hu) du = f(x) + \frac{1}{2} k_2 f''(x) h^2 + o(h^3).$$

(b) For the variance, show that

$$\begin{aligned}\text{Var } f_n(x) &= n^{-1} \text{Var } K_h(X_i - x) = n^{-1} \left\{ \int K_h(y - x)^2 f(y) dy - g_n(x)^2 \right\} \\ &= (nh)^{-1} \int K(u)^2 f(x + hu) du - n^{-1} g_n(x)^2,\end{aligned}$$

with leading terms $R(K)f(x)/(nh) - f(x)^2/n$. Show that $f_n(x) \rightarrow_{\text{pr}} f(x)$ provided the bandwidths satisfy $h \rightarrow 0$ and $nh \rightarrow \infty$. If $h = c/n^\alpha$, what is the allowed range of α ?

(c) For the integrated squared error, show that

$$\text{mise}_n = \frac{1}{4} k_2 h^2 R(f'') + (nh)^{-1} R(K) + \text{smaller terms}.$$

Show next that the function $\frac{1}{4} a h^4 + b/(nh)$ is minimised at $h_0 = (b/a)^{1/5} n^{-1/5}$, with minimum value $(5/4) a^{1/5} b^{4/5} n^{-4/5}$. Hence deduce that the best bandwidth, in the sense of minimising the approximate mise, is

$$h_{0,n} = \{R(K)/k_2^2\}^{1/5} R(f'')^{-1/5} n^{-1/5},$$

with minimum mise

$$\min_h \text{mise} = (5/4) k_2^{2/5} R(K)^{4/5} R(f'')^{1/5} n^{-4/5} - R(f)/n + \text{smaller terms}.$$

(d) For the case of f being a normal $N(\xi, \sigma^2)$ density, using $K = \phi$ as the kernel function, show that

$$h_{0,n} = (4/3)^{1/5} \sigma n^{-1/5} = 1.0592 \sigma n^{-1/5}.$$

How well one can estimate the unknown f with the kernel method, for the case of a normal density, is seen through

$$\min_h \text{mise} = \frac{5}{4} \left(\frac{3}{4}\right)^{1/5} (1/\sqrt{2}) \phi(0) (1/\sigma) n^{-4/5} + O(n^{-1}) = 0.3329 (1/\sigma) n^{-4/5} + O(n^{-1}).$$

(e) The rule-of-thumb method is then to use the standard normal kernel, with bandwidth $h_n = 1.0592 \hat{\sigma}/n^{1/5}$, with $\hat{\sigma}$ the usual empirical standard deviation, or perhaps a robust version. One often sees that this is a reasonably clever and reasonably robust bandwidth, even though its precise formula involves an auxiliary normality assumption which is not meant to be trusted. For an illustration, reproduce a version of Figure 15.6 of Story #6, producing density estimates for the birthweights from $n_0 = 115$ nonsmoking and the $n_1 = 74$ smoking mothers.

(f) Suppose we need to estimate f not only in a given known position, but at $\mu = Q(\frac{1}{2})$ the unknown median. Show that $f_n(M_n) \rightarrow_{\text{pr}} f(\mu)$, with M_n the sample median, under the conditions above.

Ex. 3.43 *Mean or median for estimating the symmetry point?* Consider i.i.d. data $Y_1, \dots, Y_n$ from a symmetric density. For estimating the symmetry location θ , we learned in Ex. 3.20 that the mean $\bar{Y}_n$ is better than the median M_n provided $\rho < 1$, and vice versa for $\rho > 1$, where $\rho = 2\sigma f(\theta)$. So why not let the data decide?

(a) Consider $\hat{\rho} = 2\hat{\sigma} f_n(\hat{\theta}_{\text{pil}})$, with f_n the kernel density estimator of Ex. 3.42, evaluated at some pilot estimate of θ . Show that $\hat{\rho} \rightarrow_{\text{pr}} \rho$ for growing n .

(b) Let then θ^* be $\bar{Y}_n$ if $\hat{\rho} \leq 1$ and M_n if $\hat{\rho} > 1$. As long as the true ρ is not precisely equal to 1, show that $\sqrt{n}(\theta^* - \theta)$ has the same limit distribution as does $\sqrt{n}(\bar{Y}_n - \theta)$, if $\rho < 1$, and as $\sqrt{n}(M_n - \theta)$, if $\rho > 1$.

More material, leftovers (as of 16 August 2026)

Ex. 3.44 *Random quantiles.* (xx might push this to Ch. 14. xx) Models for a random c.d.f. F on the unit interval may be built via their quantile functions $Q = F^{-1}$.

(a) Consider building a model for $Q_1 < Q_2 < Q_3$ as follows: (i) $Q_2 \sim \text{unif}(0, 1)$; (ii) given $Q_2 = q_2$, let independently $Q_1 \sim \text{unif}(0, q_2)$ and $Q_3 \sim \text{unif}(q_2, 1)$. Show that (Q_1, Q_2, Q_3) has mean $(1/4, 2/4, 3/4)$.

(b) Show that the densities of Q_1 and Q_3 become $g_1(q_1) = -\log q_1$ and $g_3(q_3) = -\log(1 - q_3)$ on the unit interval.

(c) Simulate say $\text{sim} = 25$ random quantile functions Q on the unit interval, as piecewise linear bits, over $[0, 1/4]$, $[1/4, 2/4]$, $[2/4, 3/4]$, $[3/4, 1]$, with values Q_1, Q_2, Q_3 at $1/4, 2/4, 3/4$. Transform these to random c.d.f.s F on the unit interval, constructed as linear pieces viewed via $F(0) = 0$, $F(Q_1) = 1/4$, $F(Q_2) = 2/4$, $F(Q_3) = 3/4$, $F(1) = 1$.

(d) Generalise to the case of $Q_1 < \dots < Q_7$, thought of as random versions of the seven octiles: (i) $Q_4 \sim \text{unif}(0, 1)$; (ii) given Q_4 , let $Q_2 \sim \text{unif}(0, Q_4)$ and $Q_6 \sim (Q_4, 1)$, independently; (iii) given Q_2, Q_4, Q_6 , let Q_1, Q_3, Q_5, Q_7 be independent and uniform on the intervals $(0, Q_2), (Q_2, Q_4), (Q_4, Q_6), (Q_6, 1)$, respectively. Show that Q_j has mean $j/8$, and find the densities for each individual V_j .

(e) Simulate sim random quantile functions Q , via linear pieces, as above; then transform to random c.d.f.s $F = Q^{-1}$, piecewise linear with $F(Q_j) = j/8$.

Ex. 3.45 *Comparing estimators for a skewed distribution on the unit interval.* Consider i.i.d. observations $Y_1, \dots, Y_n$ from the model with c.d.f. $F(y, \theta) = y^\theta$ on the unit interval, with θ an unknown positive parameter. Various estimators may be constructed and then compared in terms of precision.

(a) We start with the moment estimator. Show that the density is $f(y, \theta) = \theta y^{\theta-1}$, and that its mean is $\theta/(\theta + 1)$. Show that the moment method estimator becomes $\hat{\theta}_m = \bar{Y}/(1 - \bar{Y})$, in terms of the sample average $\bar{Y}$.

(b) Explain that the q level quantile is $y_q = q^{1/\theta}$, and that quantile matching leads to the estimator $\hat{\theta}_q = \log q / \log Q_{n,q}$, with $Q_{n,q}$ the empirical quantile; a special case is the median matching estimator $\hat{\theta}_{0.50} = -\log 2 / M_n$, in terms of the median.

(c) Show that $Z_i = -\log Y_i \sim \text{Expo}(\theta)$. Argue that moment matching on this scale leads to the estimator $\hat{\theta}_e = 1/\bar{Z}$, with $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$.

(d) To compare the three estimators we find their limit distributions. Use the CLT to identify the limit of $\sqrt{n}\{\bar{Y} - \theta/(\theta + 1)\}$, and then the delta method to establish that

$$\sqrt{n}(\hat{\theta}_m - \theta) \rightarrow_d N(0, \tau_m^2), \quad \text{where } \tau_m^2 = (1 + \theta)^4 [\theta/(\theta + 2) - \{\theta/(\theta + 1)\}^2].$$

(e) For the quantiles, show first that $\sqrt{n}(Q_{n,q} - q^{1/\theta}) \rightarrow_d V$, where V has variance $q(1 - q)/\theta^2 q^{2-2/\theta}$. Again via the delta method, establish that

$$\sqrt{n}(\hat{\theta}_q - \theta) \rightarrow_d -\frac{\theta^2}{\log q} \frac{1}{q^{1/\theta}} V \sim N(0, \tau_q^2), \quad \text{where } \tau_q^2 = \frac{\theta^2}{(\log q)^2} \frac{1 - q}{q}.$$

Examine this function numerically and find that the best value for q , leading to the smallest variance, is $q^* = 0.203$, with variance $1.544 \theta^2$.

(f) For the transformed data estimator, show $\sqrt{n}(\hat{\theta}_e - \theta) \rightarrow_d N(0, \theta^2)$. Now compare the precision of the three estimators, and comment.

Notes and pointers

(xx to come. we point to various matters, genesis of crucial concepts, and also point to chapters ahead. explain that yes, we've touched and used CLT and delta method and a bit more here, but with details and more material to come in Ch 4. xx)

(xx simultaneous confidence intervals: can point to Bonferroni method. in notes: [Hjort \(1988b\)](#). xx)

(xx unbiased estimators: more, and UMVU, in Ch9. xx)

(xx CLT for binomials: associated with the famous names de Moivre (who showed a version of this in 1733) and Laplace (who had a clearer and more general proof in 1812). xx)

Briefly genesis of неравенство Маркова, the Markov, and the неравенство Чебышёва, the Chebyshev (often anglicised to Chebyshev, but his name was really Чебышёв). mention [Kahneman et al. \(2020\)](#).

I.4

Testing, sufficiency, power

In previous chapters we have learned about classes of distributions, their parameters, ways of estimating these from data, along with assessment of precision and confidence intervals. The present chapter goes on to the fundamental statistical reporting tool of hypothesis testing. Statistical testing of hypotheses are data-based rules for when to reject (and hence, when not to reject) a hypothesis about the parameters of a model. Theory is developed to construct such tests in quite general setups. These are constructed to have a certain significance level, like 0.05, the intended low probability of rejecting the hypothesis if it is in fact true. It also has a power function, the probability of rejection as a function of how far the model parameters might be from the hypothesis. We learn the basics of the Neyman–Pearson theory for optimal testing, and see it panned out for many situations, including the general setting of exponential families. The fruitful concept of sufficiency, the notion that a lower-dimensional vector of summaries contains all statistical information about a model, is also developed in this chapter, with implications for both estimation and testing. Parts of the theory of this chapter are generalised and extended in Chapters 5, 8, 9.

Key words: ancillarity, completeness, conditional tests, exponential family, factorisation theorem, Neyman–Pearson optimality, power, p-value, sufficiency, testing, Wald ratios

In the broad context of analysing data from models, the previous chapter dealt generally speaking with estimators of relevant parameters, their precision, and comparisons, leading in particular to confidence intervals. The present chapter develops the methods and applications further, in the direction of *statistical testing of null hypotheses*. Studying such tests involves their interpretation, construction, properties, performance, along with connections to confidence and yet other themes.

Consider in general terms data y stemming from a model with a parameter vector $\theta = (\theta_1, \dots, \theta_p)$, and suppose one wishes to test the null hypothesis H_0 that θ is inside a well-defined subset Θ_0 of the full parameter region Θ . Precisely what constitutes a null hypothesis is a matter of scientific and statistical context, often reflecting the intentions and the overall aims of the data collection and its analysis. The null hypothesis is typically a statement, concerning the nature of the mechanisms studied, the incorrect rejection of which one attempts to avoid. A simple illustration is testing whether a particular regression coefficient is equal to zero (so H_0 could be ‘ $\beta_3 = 0$ ’), or testing whether two parameters, perhaps for two groups, are equal (so H_0 would be ‘ $\theta_A = \theta_B$ ’).

A *test* for such a hypothesis is a rule saying ‘ H_0 is to be rejected if data y fall in the set R ’, along with the complementary statement ‘ H_0 is not rejected if data y fall in the set R^c ’. We talk here of *the rejection region* R and *the acceptance region* R^c . A fundamental aspect of such a test to look for is its *significance level*, typically meant to be a relatively small probability, like 0.05 or 0.01 or even smaller. We say that the test has level α provided

the level of a
test

$$\Pr_{\theta}(\text{reject } H_0) \leq \alpha \quad \text{for all } \theta \in \Theta_0. \quad (4.1)$$

With level 0.01 one is guaranteed such a low chance of falsely claiming that H_0 is wrong, if it is indeed correct. So for the illustrations briefly alluded to above, when statisticians after careful analyses reject ‘ $\beta_3 = 0$ ’ or ‘ $\theta_A = \theta_B$ ’, therefore going on to claim that their alternatives are valid, with ‘ $\beta_3 \neq 0$ ’ and ‘ $\theta_A \neq \theta_B$ ’, these claims are seen as trustworthy (and may make it into publications), since the probability of these claims being false is so low.

Often tests are carried out via appropriate *test statistics*, as when constructing a $T = T(y)$, a function of the data y , with the property that T ought to be inside some normal and well-understood range under H_0 conditions (we shall meet many cases of determining the *null distribution* of such test statistics), but bigger, if H_0 is wrong. In such cases, the rejection region takes the form $R = \{y: T(y) \geq t_0\}$, with t_0 the rejection threshold, chosen to have $\Pr_{\theta}(T(Y) \geq t_0) \leq \alpha$ for all θ congruent with H_0 .

test statistic,
rejection
threshold

Conceptually and operationally, there is a certain direction implied when setting up an hypothesis H_0 and its alternative. Rejecting H_0 , with a test with low level testing level α , leads to a positive claim about the alternative being true; the observed data have landed in a region which if H_0 were true has low probability. On the other hand, *not rejecting* the null hypothesis should not be seen as ‘verifying’ H_0 ; one needs to be content with the not so bold statement ‘the observed data do not provide sufficient evidence for claiming that H_0 does not hold’.

We are also keenly interested in *the power* of a test, which is the detection chance

$$\text{pow}(\theta) = \Pr_{\theta}(\text{reject } H_0) \quad \text{as a function of } \theta. \quad (4.2)$$

So for $\theta \in \Theta_0$, the power is low, say bounded by the traditional 0.05, but interest is then behaviour of the power function in the alternative parameter domain $\Theta - \Theta_0$. Thus some tests are *stronger* than other tests with the same level, and we learn recipes for constructing such in the exercises below. The power of a test clearly depends on the quality of data, and typically of the sample size, as we shall see in exercises. Thus detecting that $\theta_A \neq \theta_B$, for parameters of two groups, in a setup where there really is a difference, becomes more likely with more data.

tests have
power

Below we also define, discuss, and use *p-values*, which are commonly quoted in most branches of applied statistics work, typically to indicate how clear a potential finding is. The idea is to quantify how unlikely it is, to observe what has actually been observed, if some relevant null hypothesis H_0 is actually true. If the test is set up to reject the null if an appropriate test statistic T is sufficiently large, we’re after $\Pr_{\theta}(T \geq t_{\text{obs}})$ under H_0 conditions, with t_{obs} the observed T for the given dataset. Some care is needed since that probability might depend on parameters under H_0 . The more careful version is

the p-value

$$p = \max\{\Pr_{\theta}(T \geq t_{\text{obs}}) : \theta \in \Theta_0\}. \quad (4.3)$$

A small p-value, like $p \leq 0.01$, casts serious doubt on H_0 , since the observed t_{obs} is so unlikely. A rephrasing of the testing scheme, with significance level say 0.01, is to reject H_0 if $p \leq 0.01$.

A classic result for testing theory is the Neyman–Pearson Lemma, which in an idealised setup with just two possible densities identifies the most powerful method for testing one density against the other; see Ex. 4.8–4.10. This sharpens further questions for more general setups, and we manage to find optimal tests in a variety of setups, including for the broad exponential family class. This also necessitates exploring and developing certain themes of independent interest and use, those of sufficiency, ancillarity, completeness, conditional testing; see the string of exercises starting with Ex. 4.19. A core idea in that terrain is that of data compression; for various models, a low-dimensional vector of data summaries contains all relevant statistical information.

(xx then one paragraph with pointers to other chapters and perhaps to a few of the stories. xx)

Testing, testing

Ex. 4.1 *Testing a null hypothesis.* Here we introduce the notion of null hypotheses and their testing in a few simple setups.

(a) The probability $p = \Pr(A)$ of a certain event is meant to be $p_0 = 0.33$, if the machinery around it works as it should. To test this one carries out the relevant experiment $n = 100$ times, and the event takes place $y = 44$ times. Should you reject the 0.33 hypothesis? Show that with $Y \sim \text{binom}(n, p)$, the statistic $T = (Y - np_0)^2 / \{np_0(1 - p_0)\}$ is approximately a χ_1^2 , under the null assumption that $p = p_0$. Show also that W tends to be bigger than a χ_1^2 , if $p \neq p_0$. Show that when using T as test statistic, the p-value of (4.3) becomes $p = \Pr_{p_0}(T \geq t_{\text{obs}})$, with t_{obs} the value of T seen with $y_{\text{obs}} = 44$. Compute this p-value, and decide whether the $p = 0.33$ hypothesis should be rejected at the 0.05 level.

(b) Suppose (X, Y, Z) is trinomial with sum n and probabilities (p, q, r) ; see Ex. 1.5. For concreteness, suppose a theory holds that $(p, q, r) = (0.20, 0.30, 0.50)$, and that $(X, Y, Z) = (42, 47, 111)$ is observed. Is this enough to reject the theory in question? Show that

$$W = (X - np)^2/(np) + (Y - nq)^2/(nq) + (Z - nr)^2/(nr)$$

has mean equal to 2, and that it is approximately a χ_2^2 , using the multidimensional CLT for (X, Y, Z) . Explain how this may be used to test the theory mentioned, and carry out the test. This is actually the classic Pearson chi-squared test for multinomials; see Ex. 5.25 for details, generalisations, and discussion.

(c) You're rolling your die, but it takes you as many as $y_{\text{obs}} = 15$ rolls to get your first '6'. Does this make you suspect that the probability p of a '6' is not $1/6$, but lower? With testing level $\alpha = 0.05$, what is the set of suspicious outcomes of Y , the number of throws to get the first '6'?

Ex. 4.2 *Connections from confidence intervals to testing.* Though confidence intervals and testing are two different reporting tools, when summarising inference, there are clear connections. Suppose μ is a parameter of inference, perhaps a function of model parameters, for which we can build both confidence intervals and tests.

(a) Suppose one needs to test the one-point null hypothesis that $\mu = \mu_0$, a given value, and that $[L, U]$ is a 99 percent confidence interval. Show that the test consisting in rejecting, if μ_0 is outside this interval, has level 0.01.

(b) Suppose on the other hand that there is a well-defined 0.01 level test procedure for testing $\mu = \mu_0$, against $\mu \neq \mu_0$, for each candidate value μ_0 . Gather together in a set A all the μ_0 values which are not rejected by the corresponding 0.01 level test. Show that $\Pr_\mu(\mu \in A) = 0.99$, making A a 99 percent confidence region.

Ex. 4.3 *Confidence intervals for quantiles.* Let $Y_1, \dots, Y_n$ be i.i.d. from a continuous density, positive on its sample space. How can we construct confidence intervals for the median $\mu = F^{-1}(\frac{1}{2})$, and more generally for quantiles $\mu_q = F^{-1}(q)$? There are several approaches here, but here we give the basic method via the empirical c.d.f.

(a) Let F_n be the empirical c.d.f. for the data, see Ex. 3.9. If μ_0 is the true median, show that $F_n(\mu_0)$ is a simple B_n/n , with $B_n \sim \text{binom}(n, \frac{1}{2})$, and that this implies $W_n(\mu_0) = \sqrt{n}\{F_n(\mu_0) - \frac{1}{2}\}/\frac{1}{2}$ being approximately a standard normal. Argue that a natural 0.05 level test for $\mu = \mu_0$ is to accept the hypothesis provided $|W_n(\mu_0)| \leq 1.96$.

(b) Following the general testing-to-confidence connection of Ex. 4.2, show that the associated 95 percent confidence interval becomes $\text{ci}_n = \{\mu: |F_n(\mu) - \frac{1}{2}| \leq 1.96/(2\sqrt{n})\}$. In other words, we may read off the interval from a plot of F_n , without knowing or taking on board the details of the exact or approximate distribution of sample quantiles, as with Ex. 3.21.

(c) Generalise to the case of any quantile $\mu_q = F^{-1}(q)$. Show that the recipe above leads to the confidence interval $\text{ci}_n = \{\mu_q: |F_n(\mu_q) - q| \leq z_\alpha\{q(1-q)\}^{1/2}/\sqrt{n}\}$, where $\Pr(|N(0,1)| \leq z_\alpha) = \alpha$, the confidence level. For an illustration of these methods, check Story #6, where we plot the empirical c.d.f. and read off confidence intervals for quantiles $F^{-1}(q)$ at levels 0.10, 0.50, 0.90, for the weight of mothers pre pregnancy.

(d) Discuss ways in which more accurate confidence intervals can be constructed, using the exact binomial distribution; for the median, for example, $nF_n(\mu) \sim \text{binom}(n, \frac{1}{2})$. This leads to slight modification of the horizontal bands when reading off intervals from the empirical c.d.f.

Ex. 4.4 *t testing, one and two samples.* Testing the mean based on a sample of normal observations is a recurring problem, in several guises, and with the famous t test being the canonical procedure; details are given below. We also go through the basics for testing the difference of means for two normal samples. Due to the connections to confidence intervals discussed in Ex. 4.2 we also find accurate confidence intervals for the key parameters. Beyond their concrete relevance and repeated use in these standard setups, the t testing procedures are important since similar constructions can be worked with in large classes of more complicated setups, but then typically with approximations to key distributions, rather than the exact solutions found under these classic strict modelling assumptions.

(a) Suppose $X_1, \dots, X_n$ are i.i.d. $N(\xi, \sigma^2)$, so far assuming σ to be known, and that one wishes to test $H_0: \xi = \xi_0$ against the alternative that $\xi \neq \xi_0$, where ξ_0 is some appropriate given value, like zero. Using the exact distribution of $\bar{X}$, cf. Ex. 1.2, show that $Z = (\bar{X} - \xi_0)/(\sigma/\sqrt{n}) = \sqrt{n}(\bar{X} - \xi_0)/\sigma$ is standard normal under H_0 , and that $|Z|$ will tend to be bigger than a normal if H_0 is not true. Explain that the test which rejects H_0 when $|Z| > 1.96$ has level 0.05.

(b) For the more realistic case of σ not being known, the natural construction is the t statistic $t = \sqrt{n}(\bar{X} - \xi_0)/\hat{\sigma}$, with $\hat{\sigma}$ the empirical standard deviation. Show from Ex. 1.50

that $t \sim t_{n-1}$ under H_0 , and write down a precise 0.05 level test.

(c) Suppose now that the mean ξ for the population of X_i is to be compared with the mean η for another population, where we have i.i.d. data $Y_1, \dots, Y_m \sim N(\eta, \sigma^2)$. So the task is to test H_0 that $\xi = \eta$. Show first that $\bar{Y} - \bar{X} \sim N(\eta - \xi, \sigma^2(1/m + 1/n))$. To build a t statistic we need an estimator for the variance. Writing $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ for the empirical deviances for samples 1 and 2, show that with

$$\hat{\sigma}^2 = c_1 \hat{\sigma}_1^2 + c_2 \hat{\sigma}_2^2, \quad \text{using } c_1 = (n-1)/(n+m-2), c_2 = (m-1)/(n+m-2),$$

we have $\hat{\sigma}^2 \sim \sigma^2 \chi_{n+m-2}^2 / (n+m-2)$, independent of $\bar{X} - \bar{Y}$. Conclude that $t = (\bar{X} - \bar{Y})/R \sim t_{n+m-2}$ under H_0 , where $R = \hat{\sigma}(1/m + 1/n)^{1/2}$.

(d) Use these building blocks to also construct a 90 percent confidence interval for $d = \xi - \eta$. You may check Ex. 8.24 for confidence for the Mahalanobis distance $|d|/\sigma$.

see Story #80
for methods
needed when
the variances
are not assumed
equal

Wald tests

Ex. 4.5 *Wald tests.* Here we present the basics of so-called Wald tests, a general and practical way of forming tests via approximate normality. Such tests are used very routinely when looking for and reporting findings about regression coefficients in all standard regression models; see Story #6 for an illustration. The *Wald tests* can be constructed almost immediately, from the confidence interval construction of Ex. 3.11, via the interval-to-test connection of Ex. 4.2, but we tend to a few details to allow for potential simplifications of assumptions. – Assume $Y_1, \dots, Y_n$ comprise the data (not necessarily i.i.d.), from a suitable model with vector parameter θ . Suppose further that $\mu = \mu(\theta)$ is a focus parameter, for which we need to test $\mu = \mu_0$, for a given null value μ_0 .

(a) Suppose there is an estimator $\hat{\mu}$ with the property that $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_d N(0, \tau^2)$, and that there is a consistent estimator $\hat{\tau}$ for this limit spread τ . Show as with Ex. 3.11 that $W_n = \sqrt{n}(\hat{\mu} - \mu)/\hat{\tau} \rightarrow_d N(0, 1)$. Show that the arguments go through, with a limiting standard normal, for $W_{n,0} = \sqrt{n}(\hat{\mu} - \mu_0)/\hat{\tau}_0$, at the null hypothesis, as long as $\hat{\tau}_0$ is consistent for τ at this null position; also, technically speaking, we only need to establish $\sqrt{n}(\hat{\mu} - \mu_0) \rightarrow_d N(0, \tau^2)$ at the null hypothesis. Conclude that the test which rejects $\mu = \mu_0$ when $|W_{n,0}| \geq 1.96$ has level approaching 0.05 (and of course similarly with other chosen testing levels; level 0.01 corresponds to threshold 2.576, etc.).

(b) Explain that $p = \Pr(|N(0, 1)| \geq W_{n,0,\text{obs}})$ is an approximation to the exact p-value.

(c) Suppose $X \sim \text{binom}(100, p)$, with a need for testing $p = 0.33$. Set up a Wald test, and compute the p-value, if $x_{\text{obs}} = 44$. Suppose then that there is an additional $Y \sim \text{binom}(100, q)$, and that one wishes to test $p = q$. Set up a Wald test, and compute the p-value, if (X, Y) are observed to be $(44, 55)$.

Ex. 4.6 *When confidence intervals for two parameters overlap.* Suppose confidence intervals for parameters a and b overlap. It might sound plausible that the null hypothesis $a = b$ will then not be rejected. To check aspects of this and related questions, consider a prototype setup where $\hat{a} \sim N(a, 1)$ and $\hat{b} \sim N(b, 1)$ are independent.

(a) With confidence level $1 - \alpha$, show that the canonical intervals are $\hat{a} \pm z_0$ and $\hat{b} \pm z_0$, with $z_0 = \Phi^{-1}(1 - \alpha/2)$, e.g. the standard 1.96 for level 0.95. Show that overlapping intervals corresponds to $|\hat{d}| \leq 2z_0$, where $\hat{d} = \hat{b} - \hat{a} \sim N(b - a, 2)$.

(b) Consider the test of $a = b$ which rejects when the intervals for a and b are disjoint. Show that this test has considerably lower significance level than α ; for 95 percent intervals this level is 0.0056.

(c) Yes, it is possible to have overlapping 95 percent intervals and yet reject $a = b$ at level 5 percent. Identify the range of $\hat{d}$ where this happens.

Ex. 4.7 Consider the alternative. If your null hypothesis H_0 is wrong, and therefore ought to be rejected by a clever test procedure, in which ways might it be wrong, and to what extent?

(a) For a simple illustration, suppose $\bar{x} = 2.22$ and $\bar{y} = 2.66$, sample means from $N(\xi_1, \sigma^2)$ and $N(\xi_2, \sigma^2)$, for sample sizes 100, for simplicity with known $\sigma = 1$. (i) Test the H_0 , that the two means are equal; (ii) give a 95 percent confidence interval for $\delta = \xi_2 - \xi_1$. The point is that (ii) is more informative than the yes-no answer to the testing question.

(b) Suppose X and Y are binomial, of size $n = 100$, with probabilities p and q . You observed $(x_{\text{obs}}, y_{\text{obs}}) = (33, 55)$. Test $H_0: p = q$, and, in addition, find a confidence interval for $\rho = q/p$ (use the delta method), and comment.

(c) Suppose $X_1, \dots, X_{10}$ are independent, with $X_j \sim N(\xi_j, 1)$, and with H_0 equality of means to be tested. Give first an interpretation of λ , where $\lambda^2 = \sum_{j=1}^{10} (\xi_j - \bar{\xi})^2$, where $\bar{\xi}$ is the average of the ξ_j . Explain that rejecting H_0 if $Z = \sum_{j=1}^{10} (X_j - \bar{X})^2 > \gamma_{0.95,9} = 16.919$ is a 0.05 level test. If the data deliver $z_{\text{obs}} = 22.22$, you reject equality; what can you say about λ ?

see Ex. 8.7 for a full CD for λ

(d) The alternatives to H_0 should be taken into account also for constructing a test that might be better than others. For the little scenario above, suppose one knows a priori that if H_0 is wrong, it must be because one of the means is different from the rest. Explain that if H_0 is wrong in this fashion, then

$$W = \max_{j \leq 10} |X_j - \bar{X}_{(-j)}|$$

ought to be able to detect this, with $\bar{X}_{(-j)}$ the mean of the other nine X_ℓ . Carry out simulations to find the upper 0.05 point in its null distribution. Then compare the power functions of the Z and the W test, against alternatives of the form $(\xi, \dots, \xi + d, \dots, \xi)$, with d the amount with which the single ξ_j in question is different from the rest. You should find an explicit formula for the power function $\pi_Z(d)$ for Z , whereas $\pi_W(d)$ needs to be simulated.

(e) For the testing of equal means situation above, play with your code for other number of groups m than 10. You should find that for m moderate or big, the alternative-conscious W test solidly outperforms the overall-modus Z . This type of alternatives, where H_0 can be wrong only in terms of a few percent of parameters studied can be significantly different from the rest, or perhaps from a common null, are called *sparse*.

sparse alternatives

Neyman–Pearson optimal testing

Ex. 4.8 The Neyman–Pearson Lemma: the basics. Suppose data y come from a density f , where there are just two possibilities: either $f = f_0$, which is the null hypothesis to be tested, or $f = f_1$, the alternative. Here there's an optimal strategy, made clear by the so-called Neyman–Pearson Lemma, part of the 49-page landmark paper [Neyman and Pearson \(1933\)](#). For simplicity of presentation we consider the continuous case, where f_0 and f_1 are densities on the relevant sample space $\mathcal{Y}$ (which can be multidimensional). – A test function is a $T: \mathcal{Y} \rightarrow [0, 1]$, with $T(y)$ the probability of rejecting f_0 if the data take on value y . This setup even allows the possibility of an element of randomisation, as

in ‘if y turns out be 3.33 I throw some coins and I reject H_0 with probability 0.77’. Once in a blue while this might be of relevance, with discrete data, but in practice such a test function $T(y)$ takes on only values 1, for a rejection set R , and 0, for the complementary acceptance set R^c .

(a) Show that the probability of rejecting the null, if the null is true, can be written $\Pr_{f_0}(\text{reject}) = \int f_0 T \, dy$.

(b) For a given testing level α , like 0.01, let T^* be the test which rejects when $f_1(y)/f_0(y) \geq c$, with c tuned such that

$$\Pr_{f_0}(T^* \text{ rejects}) = \int_{y: f_1(y)/f_0(y) \geq c} f_0(y) \, dy = \alpha.$$

Let T be any other test function with the same level α . Show that the power difference at f_1 can be written

$$\pi_{T^*}(f_1) - \pi_T(f_1) = \int f_1(T^* - T) \, dy = \int (f_1 - cf_0)(T^* - T) \, dy.$$

the Neyman–
Pearson Lemma

(c) Show that among all possible tests, with level α , the T^* has the strongest detection power at position f_1 .

Ex. 4.9 *The Neyman–Pearson Lemma: more details.* Here we tend to some further details, related to the Neyman–Pearson Lemma and its proof, in Ex. 4.8.

(a) For two positive densities f_0 and f_1 defined on the same sample space, as with the Neyman–Pearson Lemma, consider the event $A_c = \{f_1(Y)/f_0(Y) \geq c\}$. Show that the function $p(c) = \Pr_{f_0}(A_c) = \int_{f_1(y) \geq cf_0(y)} f_0(y) \, dy$ is a continuous and monotone function, starting at 1 and ending at 0, when c travels through $[0, \infty)$. Hence deduce that there for given α really is a unique c in the Neyman–Pearson recipe.

(b) Illustrate the $p(c) = \Pr_{f_0}(A_c)$ in a few concrete situations, including (i) $f_0 \sim N(0, 1)$ and $f_1 \sim N(1, 1)$; (ii) $f_0 \sim \text{Gam}(2.2, 3.3)$ and $f_1 \sim \text{Gam}(3.3, 2.2)$. (iii) $f_0 \sim \text{Gam}(3.3, 2.2)$ and $f_1 \sim \text{Gam}(2.2, 3.3)$. For these situations, find the rejection threshold c such that the test rejecting when $f_1(y)/f_0(y) \geq c$ has significance level 0.05.

(c) For the three cases above, determine also the power at f_1 . Note that the power at f_1 when testing f_0 is typically different from the power at f_0 when testing f_1 .

Ex. 4.10 *The Neyman–Pearson Lemma: applications.* For the simple two-possibilities setup we learn from the Neyman–Pearson Lemma that there is a clear recipe for setting up the optimal test for $f = f_0$ against $f = f_1$. Here are some examples.

(a) Suppose $Y \sim N(\theta, 1)$. Show that the optimal test of level $\alpha = 0.01$, for testing $\theta = 0$ vs. $\theta = 1.234$, is to reject if $Y \geq z_{0.99} = 2.326$, the upper 0.01 point of the standard normal.

(b) In this situation, verify that one finds the very same optimal 0.01 level test, for any alternative point $\theta_1 > 0$. Hence the $Y \geq z_{0.99}$ test is *uniformly most powerful*, against all positive alternative.

UMP, uniformly
most powerful

(c) Generalise this to the case of data $Y_1, \dots, Y_n$ being i.i.d. from the normal $N(\theta, \sigma^2)$, with known σ . Show that the test which rejects $\theta = 0$ against $\theta > 0$ when $Z_n = \sqrt{n}\bar{Y}/\sigma > z_{1-\alpha}$ is uniformly most powerful, among all tests with level α ; here $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$. Find its power function, and draw it in a diagram, for $\theta_0 = 1.234$, $\sigma = 1$, and for $n = 10, 20, 30$.

(d) Let $Y_1, \dots, Y_n$ be i.i.d. from the $N(\theta_0, \sigma^2)$ distribution, with θ_0 known. Consider the problem of testing $\sigma = \sigma_0$ against $\sigma > \sigma_0$, at level say 0.01, where σ_0 is a prescribed null value. Show that the test which rejects when $V_n = \sum_{i=1}^n (Y_i - \theta_0)^2 / \sigma_0^2 \geq \gamma_{n,0.99}$, the 0.99 quantile of the χ_n^2 , is uniformly optimal. Find its power function.

(e) Consider f_0 , the standard normal, and $f_1(y) = \frac{1}{2}\sqrt{2}\exp(-\sqrt{2}|y|)$; they have both zero mean and unit variance. Find the optimal test for f_0 against f_1 , with level 0.05, and find its detection power at f_1 . Then do the opposite, constructing the best test at level 0.05 for f_1 against f_0 , and find the power at f_0 .

(f) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the exponential $\theta \exp(-\theta y)$. Find the strongest 0.10 level test for $\theta = \theta_0$ against an alternative $\theta_1 > \theta_0$. Your test will not depend on the θ_1 , as long as it is to the right of θ_0 ; hence this test is uniformly most powerful against these alternatives.

Ex. 4.11 *Setting up the test via null simulations.* Consider the simple two-possibilities Neyman–Pearson setup, with data Y from a density f , where one tests $f = f_0$ vs. $f = f_1$, for given densities f_0 and f_1 .

(a) Writing y_{obs} for the observed data, the strongest test consists in rejecting when $T(y_{\text{obs}}) = f_1(y_{\text{obs}})/f_0(y_{\text{obs}}) \geq c$, with c finetuned to have level say $\alpha = 0.05$. Explain that to actually find the c , in a perhaps complicated situation, one may simulate a high number of $T(y^*) = f_1(y^*)/f_0(y^*)$, with y^* drawn from f_0 , and then set c equal to the 0.95 quantile in this distribution.

(b) Explain that the Neyman–Pearson test may be defined and constructed using any monotone smooth transformation of the density ratio. In the simulation recipe above one may e.g. use simulations of $V(y) = \log f_1(y) - \log f_0(y)$, which might be numerically easier and safer, and then reject if the observed $V(y_{\text{obs}})$ is larger than then 0.95 quantile point among the simulated $V(y^*)$.

(c) As an illustration, suppose the data are $Y_1, \dots, Y_n$, with $n = 12$, and that one needs to test that their density f is standard normal, vs. the alternative that f is a t_ν , with $\nu = 10.0$ degrees of freedom. Show that the strongest test, with level 0.05, is to reject when $V = \sum_{i=1}^n \{\log f_\nu(Y_i) - \log \phi(Y_i)\} \geq c$, with $c = 0.636$.

(d) Explain that this null distribution simulation scheme also works in each case where one tests a specified $H_0: f = f_0$, also when the alternative hypothesis might be a bigger space than merely a given f_1 , as long as the test statistic has been decided upon.

Ex. 4.12 *Density ratios and optimal testing: the normal and the Cauchy.* The Neyman–Pearson recipe is to reject when the density ratio $f_1(y)/f_0(y)$ is sufficiently big. This pans out differently in different situations, as illustrated here.

(a) For a single observation Y , consider testing $f_0 = N(0, 1)$ against $f_1 = N(\theta_1, 1)$, with θ_1 positive. Show that

$$\frac{f_1(y)}{f_0(y)} = \frac{\exp\{-\frac{1}{2}(y - \theta_1)^2\}}{\exp(-\frac{1}{2}y^2)} = \exp(\theta_1 y - \frac{1}{2}\theta_1^2).$$

Verify that this is a monotone function in y , regardless of the value of $\theta_1 > 0$. Argue that ‘reject f_0 provided Y is big enough’ becomes the uniformly optimal test. Exhibit the rejection threshold in $Y \geq c$, if the significance level is to be 0.10.

(b) The situation is rather different for the case of the Cauchy density $f(y, \theta) = (1/\pi)\{1 +$

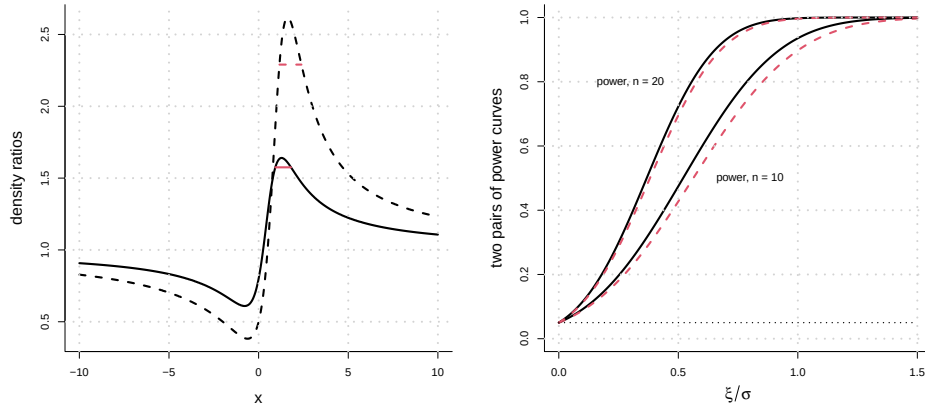


Figure 4.1: Left panel: for the Cauchy density model, see Ex. 4.12, ratios $f(y, \theta_1)/f(y, 0)$ (full line) and $f(y, \theta_2)/f(y, 0)$ (slanted) are shown, for alternative values $\theta_1 = 0.50$ and $\theta_2 = 1.00$ to the null. Also indicated are the rejection intervals $[a_1, b_1]$ and $[a_2, b_2]$, for the optimal tests against θ_1 and θ_2 . Right panel: two pairs of power. Testing $\xi = 0$ against $\xi > 0$, see Ex. 4.13, for a normal sample of size 10 (lower pair) and 20 (upper pair), at test level 0.05, as a function of ξ/σ . The lower power is for the t test (slanted, red); the upper power if for the normal based test (full, black), which requires that σ is known.

$(y - \theta)^2\}^{-1}$. Suppose we wish to test $\theta = 0$ versus a positive θ_1 . Show that

$$\frac{f_1(y)}{f_0(y)} = \frac{f(y, \theta_1)}{f(y, 0)} = \frac{1 + y^2}{1 + (y - \theta_1)^2},$$

and draw this function in a diagram, for a few values of θ_1 .

(c) For a concrete illustration, work through the alternative cases $\theta_1 = 0.50$ and $\theta_2 = 1.00$, for each case finding the rejection interval, say $[a_1, b_1]$ for the first and $[a_2, b_2]$ for the second, to give the optimal test, of level 0.10. (Answers: $[0.933, 1.804]$ for θ_1 ; $[1.161, 2.388]$ for θ_2 .) Construct a version of Figure 4.1, left panel. – The point is that these rejection regions are different; the optimal test depends on the specific alternative, and there can be no uniformly optimal test.

(d) Again for the sake of concreteness, compute the optimal possible power, for any 0.10 level test, at $\theta_1 = 0.50$ and at $\theta_2 = 1.00$. Compare these powers with that of the simple test $Y \geq 3.078$. – There are two drastic differences here, between the simple normal and the non-simple Cauchy; comment on the following notes. (i) The log-density-ratio $R(y, \theta_0, \theta_1) = \log f(y, \theta_1) - \log f(y, \theta_0)$ is monotone, for the normal, and not at all monotone for the Cauchy. (ii) There is a simple one-dimensional sufficient statistic, in the case of $Y_1, \dots, Y_n$ from the normal, whereas no such statistic exists for the Cauchy; see Ex. 4.19.

(e) The above deals with the simple stylised case of $n = 1$ datapoint. Check what happens with $n = 2$, and comment on what you find.

Ex. 4.13 *The t test and its power.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from the $N(\xi, \sigma^2)$, with testing of $\xi = 0$ required against $\xi > 0$. This is simple and standard, in the case of σ being known, but needs the t test in the case of σ being unknown and estimated from the data, as we have seen in Ex. 4.4. Setting up the test uses the relevant t distribution,

from $t = \sqrt{n}\bar{Y}/\hat{\sigma} \sim t_{n-1}$, but for studying the power function also the noncentral t distribution is required.

(a) Consider first the case of σ being known. Show that $z = \sqrt{n}\bar{Y}/\sigma$ is standard normal under the null, and that the 0.05 level test becomes that of rejecting when $z \geq z_0 = \Phi^{-1}(0.95) = 1.645$.

(b) Show that at a given $\xi > 0$, we have $z \sim N(\sqrt{n}\xi/\sigma, 1)$, and that this leads to the power function $\pi_{n,N}(\xi/\sigma) = \Phi(\sqrt{n}\xi/\sigma - z_0)$. Compute and display this power function for the case of $n = 10$ and $n = 20$, as with the black full curves of Figure 4.1 (right panel).

(c) Then consider the more complex situation where σ is not known, needing the empirical standard deviation $\hat{\sigma}$ of (1.6). We have seen in Ex. 1.50 that

$$t = \frac{\sqrt{n}\bar{Y}}{\hat{\sigma}} = \frac{\sqrt{n}\bar{Y}/\sigma}{\hat{\sigma}/\sigma} \sim \frac{N(0,1)}{(\chi_{df}^2/df)^{1/2}},$$

with numerator and denominator being independent, and where the degrees of freedom is $df = n - 1$. The probability density $g_{df}(x)$ and cumulative distribution $G_{df}(x)$ of this t_{df} distribution are moderately complicated, see the exercise mentioned, but that does not concern us much, as long as we can consult a table or run an algorithm to find associated quantiles and probabilities. Show hence that the t test, with level 0.05, must consist of rejecting when $t \geq t_0 = G_{df}^{-1}(0.95)$. Using `qt(0.95,df)` in R, we find 1.833 and 1.729 for $n = 10$ and $n = 20$. Check that `qt(0.95,df)` becomes close to 1.645 as df increases, and explain why.

(d) For the power of the t test, show that $\pi_{n,t}(\xi/\sigma) = \Pr_{\xi/\sigma}(t \geq t_0)$, where

$$t = \frac{\sqrt{n}\bar{Y}}{\hat{\sigma}} \sim \frac{N(\sqrt{n}\xi/\sigma, 1)}{(\chi_{df}^2/df)^{1/2}},$$

again with numerator and denominator independent. Explain that the power function therefore can be written

$$\pi_{n,t}(\xi/\sigma) = \Pr_{\xi/\sigma}(t \geq t_{0,m}) = 1 - G_m(t_{0,m}, \sqrt{n}\xi/\sigma),$$

with $G_m(x, \lambda)$ denoting the c.d.f. for this noncentral t with degrees of freedom m and noncentrality parameter λ ; see Ex. 1.51. This function is complicated, but can easily be found numerically via e.g. `pt(x,m,lambda)` in R. Construct a version of Figure 4.1, right panel, perhaps with other sample sizes than 10, 20. Comment on your findings.

(e) Describe how the two-sided tests and power functions pan out here, when $\xi = 0$ is to be tested against $\xi \neq 0$. Make a corresponding version of Figure 4.1, right panel, with the relevant two-sided power functions.

Ex. 4.14 *Establishing that a parameter is close to zero.* Suppose δ is some effect parameter, where the scientific context relates to establishing that it is close to zero. In such situations it might be natural to test $H_0: |\delta| \geq \varepsilon$, for some known small threshold ε , against the alternative that $\delta \in (-\varepsilon, \varepsilon)$. In some applications the δ might be the difference between two effect parameters, aiming to infer that these are close. This is turning the tables, somehow, compared to the more traditional setups, where the hypothesis to be tested is that a parameter is close to zero, against the alternative that it is some distance away.

sometimes
called
bioequivalence

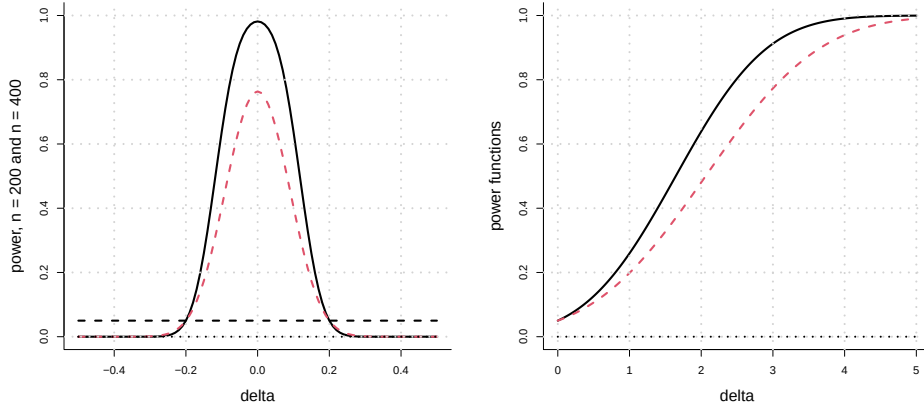


Figure 4.2: Left panel: power functions for the bioequivalence tests of Ex. 4.14, with $n = 200$ (slanted curve) and $n = 400$ (full curve), and threshold $\varepsilon = 0.20$. Right panel: limiting local power functions for the two tests of Ex. 4.15, for $\theta \leq \theta_0$ against $\theta > \theta_0$, in terms of the local alternatives parameter δ , with $\theta_n = \theta_0 + \delta/\sqrt{n}$. The more powerful test is based on $\bar{Y}$, the other on the median M_n .

(a) To study the main features of such a situation, consider a prototype setup with i.i.d. observations $Y_1, \dots, Y_n$ being $N(\delta, 1)$. Show that $\sqrt{n}\bar{Y} \sim N(\sqrt{n}\delta, 1)$, which implies $Z_n = n\bar{Y}^2 \sim \chi_1^2(n\delta^2)$.

(b) Explain that it is natural to reject the null, and hence claim that $|\theta|$ is small, if $|\sqrt{n}\bar{Y}| \leq c$, or equivalently $Z_n \leq c^2$, for c calibrated to reach testing level e.g. $\alpha = 0.05$. Show that $\Pr_\delta(|\sqrt{n}\bar{Y}| \leq c) = \Gamma_1(c^2, n\delta^2)$, in terms of the c.d.f. for the noncentral χ_1^2 .

(c) Explain that to have significance level 0.05, we need $\Gamma_1(c^2, n\varepsilon^2) = 0.05$. Find the full power curve $\pi_n(\delta) = \Pr_\delta(Z_n \leq c^2)$, and comment on the level of its maximum. Construct a version of Figure 4.2, left panel, for threshold $\varepsilon = 0.20$, for sample sizes $n = 200$ and $n = 400$.

Power, local power, average power

Ex. 4.15 *Power and local power: a particular case.* This exercise studies a prototype situation in some detail; the type of calculations and results will be seen to be rather similar in a long range of different situations. Let $Y_1, \dots, Y_n$ be i.i.d. data from $N(\theta, \sigma^2)$. One wishes to test $H_0: \theta = \theta_0$ vs. the alternative that $\theta > \theta_0$, where θ_0 is a known value (e.g. 3.14). Two tests will be considered, based on respectively $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ and the median M_n .

(a) For a given value of θ , prove that $\sqrt{n}(\bar{Y}_n - \theta) \rightarrow_d N(0, \sigma^2)$ and $\sqrt{n}(M_n - \theta) \rightarrow_d N(0, (\pi/2)\sigma^2)$. Note that the first result is immediate and actually holds with exactness for each n ; the second result requires more care and follows from Ex. 3.20.

(b) Working under the null hypothesis $\theta = \theta_0$, show that

$$Z_n = \frac{\sqrt{n}(\bar{Y}_n - \theta_0)}{\hat{\sigma}} \rightarrow_d N(0, 1), \quad Z_n^* = \frac{\sqrt{n}(M_n - \theta_0)}{(\pi/2)^{1/2}\hat{\sigma}} \rightarrow_d N(0, 1),$$

where $\hat{\sigma}$ is any consistent estimator of σ .

(c) With $z_{0.95} = \Phi^{-1}(0.95) = 1.645$, conclude from the above that the two tests that reject H_0 provided respectively

$$\bar{X}_n > \theta_0 + z_{0.95}\hat{\sigma}/\sqrt{n} \quad \text{and} \quad M_n > \theta_0 + z_{0.95}(\pi/2)^{1/2}\hat{\sigma}/\sqrt{n},$$

have the required asymptotic significance level 0.05; $\alpha_n = \Pr\{\text{reject } H_0 \mid \theta = \theta_0\} \rightarrow 0.05$. (There is one such α_n for the first test, and one for the other; both converge however to 0.05.)

(d) Our object is then to study the *local power*, the chance of rejecting the null hypothesis under alternatives of the type $\theta_n = \theta_0 + \delta/\sqrt{n}$, i.e. close to the null value. In generalisation of (b), show that local power

$$Z_n = \frac{\sqrt{n}(\bar{Y}_n - \theta_0)}{\hat{\sigma}} \rightarrow_d N(\delta/\sigma, 1), \quad Z_n^* = \frac{\sqrt{n}(M_n - \theta_0)}{(\pi/2)^{1/2}\hat{\sigma}} \rightarrow_d N((\pi/2)^{1/2}\delta/\sigma, 1),$$

where the convergence in question takes place under the indicated $\theta_0 + \delta/\sqrt{n}$ parameter values. Here we need to generalise the result of Ex. 3.20, to the case where θ_n moves with n . Writing F_n for the distribution of Y_i , you may do this by first showing that M_n has the same distribution as $F_n^{-1}(M_n^0) = \theta_0 + \delta/\sqrt{n} + \sigma\Phi^{-1}(M_n^0)$, where M_n^0 is the median for an i.i.d. sample from the uniform, and then using the delta method.

(e) Use these results to show that

$$\begin{aligned} \pi_n(\delta) &= \Pr\{\text{reject with } \bar{Y}_n \text{ test} \mid \theta_0 + \delta/\sqrt{n}\} \rightarrow \Phi(\delta/\sigma - z_{0.95}), \\ \pi_n^*(\delta) &= \Pr\{\text{reject with } M_n \text{ test} \mid \theta_0 + \delta/\sqrt{n}\} \rightarrow \Phi((2/\pi)^{1/2}\delta/\sigma - z_{0.95}), \end{aligned}$$

for the two power functions. Draw these in a diagram, and compare, as with Figure 4.2, right panel.

(f) Assume one wishes n to be large enough to secure that the power function is at least at level β for a certain alternative point θ_1 . Using the local power approximation, show that the required sample sizes are respectively

$$n_A \doteq \frac{\sigma^2}{(\theta_1 - \theta_0)^2} (z_{1-\alpha} + z_\beta)^2 \quad \text{and} \quad n_B \doteq \frac{\sigma^2/c^2}{(\theta_1 - \theta_0)^2} (z_{1-\alpha} + z_\beta)^2$$

for tests A (based on the mean) and B (based on the median), with $c = \sqrt{2/\pi}$. With level $\alpha = 0.05$, compute these sample sizes for the case of $\beta = 0.95$ and $\theta_1 = \theta_0 + \frac{1}{2}\sigma$.

(g) One sometimes defines the ARE, the asymptotic relative efficiency of test B with respect to test A, as the ARE

$$\text{ARE} = \lim n_A(\theta_1, \beta)/n_B(\theta_1, \beta),$$

the limit in question in the sense of alternatives θ_1 coming closer to the null hypothesis at speed $1/\sqrt{n}$. Show that the ARE in this particular situation becomes $c^2 = 2/\pi = 0.6366$; test A needs only ca. 64 percent as many data points to reach the same detection power as B needs.

Ex. 4.16 *Power and local power: general results.* In Ex. 4.15 we examined two particular estimators and tests, for the mean parameter in the normal distribution, and found precise limit distributions and local power functions, for alternatives of order $1/\sqrt{n}$ from the null hypothesis. Here we go through the appropriate generalities. The setup is that of data, of sample size n , informing us about a key parameter θ , where the null hypothesis

$H_0: \theta = \theta_0$ is to be tested against the alternative $\theta > \theta_0$, with θ_0 some specified value. We assume the data lead to one or more estimators $\hat{\theta}$, with the basic premise that at the null parameter θ_0 , we have $A_n = \sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \kappa^2)$, for the appropriate κ .

(a) With $\hat{\kappa}$ an estimator for κ , under H_0 conditions, show that $Z_n = \sqrt{n}(\hat{\theta} - \theta_0)/\hat{\kappa}_0 \rightarrow_d N(0, 1)$ under H_0 . Explain that the test which rejects when $Z_n \geq z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$ has asymptotic testing level α , i.e. $\alpha_n = \Pr_0(Z_n \geq z_{1-\alpha}) \rightarrow \alpha$.

(b) We now make two more assumptions, concerning natural smoothness in a $O(1/\sqrt{n})$ size local neighbourhood around θ_0 : (i) if the true parameter behind the data is $\theta_n = \theta_0 + \delta/\sqrt{n}$, then the same limit $N(0, \kappa^2)$ obtains for $A'_n = \sqrt{n}\{\hat{\theta} - (\theta_0 + \delta/\sqrt{n})\}$; (ii) the estimator $\hat{\kappa}$ still converges to the same κ in probability. Show that $Z_n \rightarrow_d N(\delta/\kappa, 1)$, and that this gives a limiting local power function,

$$\pi_n = \Pr(\text{reject} | \theta_0 + \delta/\sqrt{n}) \rightarrow \pi(\delta) = \Phi(\delta/\kappa - z_{1-\alpha}).$$

Show that this limit power function has derivative $\phi(z_{1-\alpha})/\kappa$ at zero.

(c) The above result is useful in its own right, providing an easy approximation to the power function for alternatives not far from the null hypothesis value. It may also be used for comparing different tests, built as above, for different estimators, say $\hat{\theta}_j$. Assume the conditions above hold for these, with $\sqrt{n}(\hat{\theta}_j - \theta_0) \rightarrow_d N(0, \kappa_j^2)$, leading to estimators $\hat{\kappa}_j$ for κ_j and to test statistics $Z_{n,j} = \sqrt{n}(\hat{\theta}_j - \theta_0)/\hat{\kappa}_j$. Explain that the best estimators lead to the best tests, with derivatives $\phi(z_{1-\alpha})/\kappa_j$ at zero. Connect also what we find here with the ARE of Ex. 4.15

Ex. 4.17 Optimal average power. Suppose Y is observed, perhaps a full vector, from a density f , where one wishes to test the null hypothesis $f = f_0$, a given density. As we saw in Ex. 4.12, there are cases where there is no uniformly optimal test, against all or a subset of alternatives; the optimal test at f_1 might be different from the one at f_2 . In one-parameter models, this is caused by the log-density-ratio not being monotone.

(a) Consider alternative densities $f_1, \dots, f_m$, given nonnegative weights of importance $w_1, \dots, w_m$. These may be taken to have sum 1. The *weighted average power* of a test T , at these points and with these weights, is

$$\bar{\pi}_T = \sum_{j=1}^m w_j \pi_T(f_j) = \sum_{j=1}^m w_j \int f_j(y) T(y) dy,$$

with $\pi_T(f_j)$ the power at f_j . Let $\bar{f}(y) = \sum_{j=1}^m w_j f_j$. Show that this average power is maximised, among all test functions $T(y)$ with level α at the null, by the $T^*(y)$ which rejects H_0 when $\bar{f}(y)/f_0(y) \geq c$, with c tuned to give rejection probability α at the null.

(b) For a one-parameter model $f(y, \theta)$, consider testing of $\theta = \theta_0$ against $\theta > \theta_0$. For any test function $T(y)$ with rejection level α at the null, consider in general terms the weighted average power

$$\bar{\pi}_T = \int_{\theta > \theta_0} \pi_T(\theta) dw(\theta),$$

with $\pi_T(\theta) = \int f(y, \theta) T(y) dy$ the power of the test at position θ . Show that $\bar{\pi}_T = \int_{\theta > \theta_0} \bar{f}(y) T(y) dy$, featuring the density $\bar{f}(y) = \int_{\theta > \theta_0} f(y, \theta) dw(\theta)$. This is the model density averaged over all alternatives to the null, as weighted by the $dw(\theta)$ measure.

(c) Show that the test maximising this weighted average power is rejecting the null if $\bar{f}(y)/f(y, \theta_0) \geq c$, with c tuned to have null level α .

(d) Suppose θ is one-dimensional and that the log-density $\log f(y, \theta)$ is monotone in θ . Explain that there then is a de facto uniformly optimal test for $\theta = \theta_0$ vs. $\theta > \theta_0$, and that the average power optimal test worked with here will be identical to this one, as long as the weight function is positive only for $\theta > 0$.

(e) Consider the $N(\xi, \sigma^2)$ model, and assume there is a need to test $f = N(0, 1)$ against the alternative that $\xi > 0$, or $\sigma > 1$, or both. Show first that the NP test against alternatives (1.1, 2.2) and (2.2, 3.3) are indeed different, so there is no uniformly most powerful test. Then if a weight function $w(\xi, \sigma)$ is specified, in the domain $\xi > 0, \sigma > 1$, show that the average power optimal test consists in rejecting if

$$V(y_1, \dots, y_n) = \frac{\bar{f}(y_1, \dots, y_n)}{f_0(y_1) \cdots f_0(y_n)} = \int_{\xi > 0, \sigma > 1} \left\{ \prod_{i=1}^n \frac{f(y_i, \xi, \sigma)}{f(y_i, 0, 1)} \right\} dw(\xi, \sigma)$$

is larger than the threshold c , with this finetuned to be the upper quantile of its null distribution. As an illustration, suppose $n = 10$ and that the weight function is flat on $[0, 4] \times [1, 6]$. Simulate from $W(y_1, \dots, y_n) = \log V(y_1, \dots, y_n)$ under the null hypothesis, as per Ex. 4.11, to find the rejection threshold, to have significance level 0.01.

(f) There are connections here to Bayesian perspectives, with distributions over parameters, see Ch. 7, and in Bayesian parlance the $\bar{f}$ is the predictive density for the data. Check also with Ex. 7.38 for approximations of relevance for carrying out testing as per the methods above. Invent an example where such approximations are used.

Sufficiency, factorisation theorem, completeness

Ex. 4.18 Sufficiency. Flip a coin ten times and record the number of heads. It is intuitively clear that the number of heads in the ten tosses is as informative about the unknown $\theta = \Pr(\text{heads})$ of the coin, as also knowing the exact ordering in which the heads and tails occurred. In fact, knowing the ordering in which the heads and tails occurred appears redundant for making inference about θ . This leads us to the notion of a *sufficient statistic*. Let X be data with a distribution from a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of distributions on the measurable space $(\mathcal{X}, \mathcal{B})$. In colloquial terms, the statistic $T = T(X)$ is sufficient for θ if for every t and θ , the conditional distribution of X given $T = t$ is the same for all values of θ . More formally, let $T : \mathcal{X} \rightarrow \mathcal{Y}$ be a statistic, then T is sufficient for $\mathcal{P}$ (or for θ) if there is a function $Q_t(B)$ of $B \in \mathcal{B}$ and $t \in \mathcal{Y}$, such that for each B and $\theta \in \Theta$, we have $Q_t(B) = P_\theta(B | T = t)$ for $P_\theta T^{-1}$ almost all t , that is

$$\int_A Q_t(B) dP_\theta T^{-1}(t) = \int_A P_\theta(B | T = t) dP_\theta T^{-1}(t) = P_\theta(B \cap T^{-1}(A)),$$

for each θ and for all A in the Borel- σ -algebra $\mathcal{B}(\mathcal{Y})$ on $\mathcal{Y}$.

(a) Let $X_1, \dots, X_n$ be i.i.d. Bernoulli(θ), and set $T = \sum_{i=1}^n X_i$. Show that for each t , $\Pr_\theta(X_1 = x_1, \dots, X_n = x_n | T = t) = 1/\binom{n}{t}$ on $\{(x_1, \dots, x_n) : x_i = 0, 1, \sum_{i=1}^n x_i = t\}$, and conclude that T is sufficient for θ .

(b) Let $Y_1, \dots, Y_n$ be i.i.d. Poisson(θ) and set $T = \sum_{i=1}^n Y_i$. Show that for each t

$$\Pr_\theta(Y_1 = y_1, \dots, Y_n = y_n | T = t) = \frac{t!}{n^t} \prod_{i=1}^n \frac{1}{y_i!},$$

on $\{(y_1, \dots, y_n) : y_i = 0, 1, 2, \dots, \sum_{i=1}^n y_i = t\}$, and conclude that T is sufficient for θ .

(c) Let $X_1, \dots, X_n$ be i.i.d. uniform on $(0, \theta)$ and set $T = \max_{i \leq n} X_i$. Show that the conditional distribution of $X_1, \dots, X_n$ given $T = t$ has density

$$f_\theta(x_1, \dots, x_n | t) = 1/t^n$$

on $\{(x_1, \dots, x_n) : 0 \leq x_i \leq t \text{ for all } i\}$, and conclude that T is sufficient for θ .

(d) Suppose that $Y \sim N(0, \sigma^2)$ and let $Z = \pm 1$ with equal probability be independent of Y . Show that $Y' = Z|Y|$ has the same distribution as Y . Consider the function

$$Q_t(B) = \frac{1}{2}I_B(t) + \frac{1}{2}I_B(-t), \quad B \in \mathcal{B}, t \geq 0,$$

and show, using that $Y' \sim Y$, that $|Y|$ is sufficient for σ^2 .

(e) Let X and Y be independent $\text{Expo}(\theta)$, and set $T = X + Y$. Let $U \sim \text{unif}(0, 1)$ be independent of X and Y , and define $X' = UT$ and $Y' = (1 - U)T$. Show that $(X, Y) \sim (X', Y')$. From this, show that T is sufficient for θ .

(f) Let X be a random vector with a distribution from $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Assume that T is sufficient for θ , and let $Q_t(B)$ be the function from the definition of sufficiency. Since $Q_t(B)$ is a version of $P_\theta(B | T = t)$, and X is a random vector, there exists a version of Q_t under $P_\theta T^{-1}$, Q_t^* , such that $B \mapsto Q_t^*(B)$ is a probability measure for all t [xx (see Notes and Pointers in the appendix) xx]. Given $T = t$, let X' be a random vector drawn from Q_t^* . Show that $X' \sim X$.

Ex. 4.19 *The factorisation theorem.* In Ex. 4.18 we found sufficient statistics through qualified guesswork. The factorisation theorem provides a recipe for finding sufficient statistics through (minor) manipulations of the density of a sample. Suppose that the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of distributions on $(\mathcal{X}, \mathcal{B})$ is dominated by a σ -finite measure μ . For each θ , let f_θ be the density of P_θ with respect to μ . The factorisation theorem says that the statistic $T: \mathcal{X} \rightarrow \mathcal{Y}$ is sufficient for $\mathcal{P}$ if and only if there exist nonnegative functions h and g_θ such that

$$f_\theta(x) = g_\theta(T(x))h(x),$$

for all $\theta \in \Theta$. We prove this theorem step by step.

(a) Assume that $\mathcal{P}$ is dominated by a probability measure P , and that for each $\theta \in \Theta$, the density $f_\theta = dP_\theta/dP$ is such that $f_\theta(x) = g_\theta(T(x))$, i.e., $h(x) = 1$ for all $x \in \mathcal{X}$. Show that T is sufficient for $\mathcal{P}$, and indeed also for the slightly larger family $\mathcal{P} \cup \{P\}$.

(b) For the converse of (a), assume that $\mathcal{P}$ is dominated by a probability measure P and that T is sufficient for $\mathcal{P} \cup \{P\}$. Let f_θ be a density of P_θ with respect to P , and show that $g_\theta(T(\cdot)) = E_P(f_\theta | T)(\cdot)$ is also a density of P_θ with respect to P , i.e. that $\int_B g_\theta(T(x)) dP(x) = P_\theta(B)$ for all $B \in \mathcal{B}$. –Taken together, (a) and (b) prove the factorisation theorem in the case that $\mathcal{P}$ is dominated by (at least) one of its members, which is not wholly unusual, consider the family $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$, for example.

(c) To prove the factorisation theorem in full generality, we use the result from Ex. [xx abc xx] in the appendix, which says that if $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ -finite measure μ , then there is a probability measure of the form $P = \sum_{j=1}^{\infty} a_j P_{\theta_j}$ equivalent to $\mathcal{P}$, i.e., $P_\theta(A) = 0$ for all θ if and only if $P(A) = 0$. Now, assume that for each $\theta \in \Theta$, P_θ has density $g_\theta(T(x))h(x)$ with respect to μ , and show that T is sufficient.

(d) Conversely, assume that $\mathcal{P}$ is dominated by μ and that T is sufficient. Show that P_θ has density $g_\theta(T(x))h(x)$ with respect to μ .

(e) Try to make sense of the colloquial definition of sufficiency given in Ex. 4.18, and show that it says the same thing as the formal definition of sufficiency.

Ex. 4.20 *Some more sufficiency.* In this exercise we employ the factorisation theorem on some concrete examples, and touch on some important facts about sufficiency.

(a) Use the factorisation theorem to find sufficient statistics for i.i.d. samples from, respectively, the Bernoulli(θ) distribution; the $N(\mu, 1)$ distribution; the $N(\mu, \sigma^2)$ distribution with both parameters unknown; the $\text{unif}(0, \theta)$ distribution; and, lastly, the $\text{unif}(\theta, \theta + 1)$ distribution. What is particular about this latter?

(b) Let X, Y be i.i.d. from $\mathcal{P} = \{\text{unif}(0, \theta) : \theta \geq 1\} \cup \{\text{unif}\{1, 2, \dots, \theta\} : \theta \in \mathbb{N}\}$. Show that $T = \max(X, Y)$ is sufficient for $\mathcal{P}$. Find also an expression for the function $Q_t(B)$ from the definition of sufficiency (see Ex. 4.18), and show that this function may be chosen so that $B \mapsto Q_t(B)$ is not a probability measure for all t .

(c) A sufficient statistic is not unique, and different sufficient statistics may provide varying degrees of data compression. At one extreme are sufficient statistics not providing any compression of the data. Show that the full sample is sufficient.

(d) Let $X_1, \dots, X_n$ be i.i.d. from an unknown distribution with c.d.f. belonging to a family of continuous c.d.f.s. Let $T = (X_{(1)}, \dots, X_{(n)})$ be the order statistics, and show that it is sufficient.

(e) Show that any one-to-one transformation of a sufficient statistic is sufficient.

(f) Let $X_1, \dots, X_n$ be i.i.d. from a member of $\mathcal{P} = \{N(0, \sigma^2) : \sigma > 0\}$ and consider the statistics $T_1 = (X_1, \dots, X_n)$, $T_2 = (X_1^2, \dots, X_n^2)$, $T_3 = (X_1^2 + \dots + X_k^2, X_{k+1}^2 + \dots + X_n^2)$, and $T_4 = X_1^2 + \dots + X_n^2$. Here, the data compression is increasing in the indices. Use the factorisation theorem to prove that all statistics are all sufficient.

Ex. 4.21 *Sufficiency in exponential families.* (xx need intro some technical but important things for the expo family. calibrate how and when we do sufficiency. xx)

(a) Suppose first, in general terms, that the joint density of $Y = (Y_1, \dots, Y_n)$ can be written $f_n(y, \theta) = g_n(T_1(y), \dots, T_p(y), \theta)h(y)$, for some functions of the data $T_1(y), \dots, T_p(y)$. Show that $(T_1(y), \dots, T_p(y))$ is sufficient. In cases where p is much smaller than n , there is significant reduction of the full data set to a small number of summary statistics.

(b) Suppose that Y stem from a member of the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where each P_θ has density of the exponential family form $f_\theta(y) = \exp\{\theta T(y) - k(\theta)\}h(y)$ with respect to some dominating measure μ . Show that $T = T(Y)$ is sufficient for the natural parameter θ ; and that the distribution of T has a density of the exponential family form, say $g_\theta(t) = \exp(\theta t - k(\theta))$, with respect to νT^{-1} , where $d\nu(y) = h(y) d\mu(y)$.

(c) Let $Y \sim N(0, \sigma^2)$. Write this normal density on exponential family form, and check the formula from (b) for the density $g_\theta(t)$ using the transformation methods of Ex. 1.12.

(d) Generalise (b) to the case where $Y_1, \dots, Y_n$ are i.i.d. from a distribution in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where each P_θ has density $f_\theta(y) = \exp\{\theta^t T(y) - k(\theta)\}h(y)$ with respect to some dominating measure μ . Show that $\bar{T} = (\bar{T}_1, \dots, \bar{T}_p)^t$, with $\bar{T}_j = (1/n) \sum_{i=1}^n T_j(Y_i)$ for $j = 1, \dots, p$, is sufficient for the natural parameter $\theta = (\theta_1, \dots, \theta_p)$; and that $\bar{T}$ has a distribution with density following the same exponential form with respect to an appropriate dominating measure. ($\bar{T}_1, \dots, \bar{T}_p$) is sufficient for $(\theta_1, \dots, \theta_p)$

(e) Here is a lemma that is useful for the following exercises. Let Pr_0 and Pr_1 be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let X and Y be random vectors with values in the measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$. Suppose that there are

functions $a: \mathcal{X} \rightarrow (0, \infty)$ and $b: \mathcal{Y} \rightarrow [0, \infty)$ such that $\Pr_1$ has density $a(x)b(y)$ with respect to $\Pr_0$. Show that the marginal distribution of Y under $\Pr_1$ is

$$\Pr_1 Y^{-1}(B) = \int_B b(y) E_0 \{a(X) | Y = y\} d\Pr_0 Y^{-1}(y), \quad B \in \mathcal{B},$$

and that for each $A \in \mathcal{A}$,

$$\Pr_1(X \in A | Y = y) = \frac{E_0 \{I_A(X)a(X) | Y = y\}}{E_0 \{a(X) | Y = y\}}, \quad \Pr_1 Y^{-1} \text{ a.e.},$$

that is, under $\Pr_1 Y^{-1}$ the right hand side is a version of the conditional distribution of X given $Y = y$.

(f) Let $\{P_{a,b}: (a,b) \in \Theta = \Theta_a \times \Theta_b\}$ be a family of distributions having densities of the exponential family form $f_{a,b}(y) = \exp\{a^t U(y) + b^t V(y) - k(a,b)\}h(y)$ with respect to μ . Use (e) to show that the marginal distribution of $V = V(Y)$ has density $\exp\{b^t v - k(a,b)\}$ with respect to the measure

$$d\lambda_a(v) = \exp\{-b_0^t v + k(a_0, b_0)\} E_{a_0, b_0} \{\exp((a - a_0)^t U) | V = v\} dP_{a_0, b_0} V^{-1}(v),$$

where (a_0, b_0) is some fixed point in Θ . The important thing here is that V has a density belonging to an exponential family, not the exact form of the measure with respect to which this is a density.

(g) Use (e) to show that the conditional distribution of U given $V = v$ has density

$$\exp\{a^t u - k_v(a)\}, \quad \text{with } k_v(a) = \log E_{a_0, b_0} \{\exp((a - a_0)^t U) | V = v\},$$

with respect to the measure $d\nu_v(u) = \exp(-a_0^t u) dP_{a_0, b_0}(U^{-1}(u) | V = v)$. In particular, the conditional density of U given $V = v$ does not depend on b .

conditional
sufficiency

(h) Show that, conditionally on $V = v$, U is sufficient for the family $\{P_a^{U|V=v} | a \in \Theta_a\}$. where $P_a^{U|V=v}$ is the conditional distribution of U given $V = v$.

Ex. 4.22 Minimal sufficiency. In Ex. 4.20 we saw that for any model there are many different sufficient statistics, often with some providing more compression of the data than others. Since the purpose of a sufficient statistic is to compress the data, this naturally leads to a search for a sufficient statistic providing the maximum amount of data compression. If such a statistic exists, it is said to be minimal sufficient. Here is the definition: Let T be sufficient for $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$. Then T is minimal sufficient if for any other sufficient statistic S , there is a measurable function g so that $P_\theta\{T = g(S)\} = 1$ for all $\theta \in \Theta$. Another way of saying this is that if T is such $S(x) = S(y)$ implies $T(x) = T(y)$ for any sufficient statistic S , then T is minimal sufficient.

minimal
sufficient,
maximal
compression

(a) Show that any one-to-one function of a minimal sufficient statistic is a minimal sufficient statistic.

(b) Suppose that T is minimal sufficient, and let S be some sufficient statistic. Show that the σ -algebra generated by T must be contained in the σ -algebra generated by S .

(c) Show that if $T(x) = T(y)$ if and only if

$$\theta \mapsto \frac{f_\theta(x)}{f_\theta(y)} \text{ is constant in } \theta,$$

then T is minimal sufficient. Here we assume, for simplicity, that $f_\theta(y) > 0$ for all y . If this is not the case, then instead consider the case where $T(x) = T(y)$ if and only if $y \in \{y: f_\theta(y) = f_\theta(x)h(x,y) \text{ for all } \theta \text{ and some } h > 0\}$.

(d) With $X \sim N(0, \sigma^2)$, show that the absolute value $|X|$ is minimal sufficient. Let $Y_1, \dots, Y_n$ be i.i.d. $N(\mu, \sigma^2)$, and show that $(\bar{Y}_n, S_n)$ is minimal sufficient, where $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ and $S_n = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Let $Z_1, \dots, Z_n$ be i.i.d. from a distribution with density $f_\theta(z) = \exp(-(z - \theta))$ for $z > \theta$ and $\theta \in \mathbb{R}$. Find a minimal sufficient statistic for θ .

(e) Let $g(x)$ be a positive and integrable function on $\mathbb{R}$, set $c(a, b)^{-1} = \int_a^b g(x) dx$, and define $f_{a,b}(x) = c(a, b)g(x)I_{(a,b)}(x)$, and suppose that $X_1, \dots, X_n$ are i.i.d. from the distribution with density $f_{a,b}(x)$. Find a minimal sufficient statistic for (a, b) .

(f) Let $X_1, \dots, X_n$ be i.i.d. from a distribution with density $f_\theta(x) = \frac{1}{2} \exp(-|x - \theta|)$ with $x, \theta \in \mathbb{R}$. Show that the order statistics are minimal sufficient.

Ex. 4.23 Ancillary statistics. Ancillarity is, in a sense, the opposite of sufficiency. If X stem from a member of $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$, a statistic $V = V(X)$ is ancillary if its distribution is the same for all θ . In other words, V by itself does not provide any information about θ . This does not, however, mean that V should be disregarded when making inference on θ . It just means that if you only learn $V = v$, you have not learned anything about θ . ancillary statistic

(a) Let $Y | (V = v) \sim N(\mu, \sigma_V^2)$ where V is Bernoulli $(\frac{1}{2})$. Why does it make the most sense to condition on $V = v$ when deriving a confidence interval for μ ?

(b) Let X and Y be independent $\text{unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, and set $T = \max(X, Y)$ and $V = |X - Y|$. Show that (T, V) is minimal sufficient, and that V is ancillary.

(c) With T and V as in (b), compute the probability $\Pr(T - V \leq \theta \leq T)$ and the conditional probability given $V = v$, $\Pr(T - \frac{1}{4}(1 + v) \leq \theta \leq T + \frac{1}{4} - \frac{3}{4}v | V = v)$. Which of these two confidence intervals makes the most sense to you?

Ex. 4.24 Completeness. Often, models are so harmoniously constructed that there are clear one-to-one connections between estimators (perhaps based on a set of summary statistics) and estimands, in the sense that there for each estimand is only one unbiased estimator. Clarifying such regularity leads to the concept of *completeness*, which turns out to be useful also when coming to conditional testing and optimal power in exercises below. Technically, suppose some vector $T = (T_1, \dots, T_p)^t$ has a distribution $f(t, \theta)$, with the property that $E_\theta h(T) = 0$ for all $\theta \in \Theta$ implies $\Pr_\theta\{h(T) = 0\} = 1$ for all θ , i.e. $h(t) = 0$ almost everywhere. We then say that T , or more formally its distribution, over the relevant parameter region, is complete. complete

(a) Let $X \sim \text{binom}(n, \theta)$, with $\theta \in (0, 1)$. Show that X is complete; zero is the only unbiased estimator of zero. (You may appeal to properties of power series.) Show that X is complete as long as the parameter region contains an open interval. Show similarly that if $X \sim \text{geom}(p)$, see Ex. 1.25, then X is complete, again requiring only that the parameter range for p contains an open interval.

(b) With $Y_1, \dots, Y_n$ i.i.d. from the uniform on $[0, \theta]$, consider $M = \max_{i \leq n} Y_i$. Show that if the parameter region is the full halfline $\theta > 0$, then M is sufficient and complete, but that M is not complete if it is a priori known that $\theta \geq 1.234$, i.e. with a restricted parameter range.

(c) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the uniform distribution over $[\theta - 1, \theta + 1]$. Show that $(Y_{(1)}, Y_{(n)})$, the smallest and largest, is sufficient, but not complete.

(d) Suppose $Y_1, \dots, Y_n$ are i.i.d. from the uniform distribution over $[a, b]$. Show that $(Y_{(1)}, Y_{(n)})$ is sufficient and complete. [xx check this xx]

(e) Consider Y being uniform on $\{1, \dots, \theta\}$, where $\theta \in \{1, 2, \dots\}$ is an unknown parameter. Show that Y is complete. When the parameter region is e.g. $\{4, 5, \dots\}$, however, show that Y is not complete.

(f) If T is complete, show that any one-to-one transformation variable $T' = a(T)$ is also complete.

(g) Consider $Y_1, \dots, Y_n$ an i.i.d. sample from the double exponential with density $f(y, \theta) = \frac{1}{2} \exp(-|y - \theta|)$. Show that the full set of order statistics $(Y_{(1)}, \dots, Y_{(n)})$ is sufficient, but not complete; do this, by exhibiting two different unbiased estimators of θ .

Ex. 4.25 Completeness for the exponential family. For the large class of exponential family models, see Ex. 1.57 and 4.21, there is a completeness lemma, as follows. Suppose $Y_1, \dots, Y_n$ are i.i.d. from the model $f(y, \theta) = \exp\{\theta^t T(y) - k(\theta)\} h(y)$, with $T = (T_1, \dots, T_p)^t$ and $\theta = (\theta_1, \dots, \theta_p)^t$ varying in an open set, then the vector of sample averages $(\bar{T}_1, \dots, \bar{T}_p)^t$ is not merely sufficient, as seen in Ex. 4.22, but also complete. We shall freely use this lemma; we give references in Notes & Pointers.

completeness
lemma for
exponential
family

(a) Let $Y_1, \dots, Y_n$ be i.i.d. from the $N(\xi, 1)$. Show that the full set $(Y_1, \dots, Y_n)$ is not complete, but that the sample mean $\bar{Y}$ is.

(b) Consider $Y_1, \dots, Y_n$ i.i.d. from the $\text{Gam}(a, b)$ model. Show that $(\sum_{i=1}^n Y_i, \sum_{i=1}^n \log Y_i)$ is sufficient and complete. Identify similarly a sufficient and complete pair of statistics for a sample from the $\text{Beta}(a, b)$.

(c) Consider an i.i.d. sample $Y_1, \dots, Y_n$ from the $N(\xi, \sigma^2)$. Show that $(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$ is sufficient and complete, and also that $(\bar{Y}, \hat{\sigma})$ is sufficient and complete. Suppose then that the variance is postulated to be equal to the squared mean, so that the sample is from $N(\theta, \theta^2)$. Construct two different unbiased estimators of θ , and show that this means that $(\bar{Y}, \hat{\sigma})$ is not complete. You may similarly construct two different unbiased estimators of θ^2 .

(d) Consider the linear regression model with $Y_i \sim N(a + bx_i, \sigma^2)$, studied in Ex. 3.32. With $\hat{a}, \hat{b}$ the least squares estimators, and $Q_0 = \sum_{i=1}^n \{Y_i - \hat{a} - \hat{b}(x_i - \bar{x})\}^2$, show that the log of the joint density can be written $-n \log \sigma - \frac{1}{2} \{Q_0 + n(\hat{a} - a)^2 + M_n(\hat{b} - b)^2\} / \sigma^2$, with $M_n = \sum_{i=1}^n (x_i - \bar{x})^2$. Write this in the exponential family fashion, with natural parameters $1/\sigma^2, a/\sigma^2, b/\sigma^2$. Argue via the general exponential family results that $(Q_0, \hat{a}, \hat{b})$ is both sufficient and complete. Extend this, arguments and results, to the general multiple linear regression model of Ex. 3.33.

this is the
log-likelihood
function, see
Ch. 5

Ex. 4.26 Basu's Lemma. Consider a setting with data Y from some parametric family, indexed by $\theta \in \Theta$, where a suitable T is sufficient and complete.

(a) Assume that a suitable statistic $Z = Z(Y)$ has a distribution not depending on the θ . Show that Z is independent of T . This is called Basu's Lemma. You may follow this path: start writing $\Pr_\theta(Z \in A) = p$, by assumption not depending on θ . Show that $p = E_\theta h(T)$, for all θ , where $h(t) = \Pr_\theta(Z \in A | T = t)$, another function not depending on θ . Explain that this implies $\Pr_\theta(Z \in A | T = t) = p$ for all θ .

Basu's lemma

(b) Consider the familiar setup with $Y_1, \dots, Y_n$ being i.i.d. from $N(\xi, \sigma^2)$. For σ fixed, show that the mean $\bar{Y}$ is sufficient and complete for ξ ; explain that all statistics Z with distribution not depending on ξ must be independent of $\bar{Y}$. Explain hence in particular that $\sum_{i=1}^n (Y_i - \bar{Y})^2$ is independent of $\bar{Y}$. This has been demonstrated in different ways in Ex. 1.54 and 1.49.

(c) In this normal sample setting, argue that the classic t ratio $t = \sqrt{n}(\bar{Y} - \xi)/\hat{\sigma}$, with

$\hat{\sigma}$ the sample standard deviation, is independent of $(\bar{Y}, \hat{\sigma})$.

(d) We have seen in Ex. 1.58 that the five-parameter binormal distribution is inside the exponential family class. With $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. from the binormal, show first that $(\bar{X}, \bar{Y}, \hat{\sigma}_1, \hat{\sigma}_2, R_n)$ is complete and sufficient, with R_n the empirical correlation coefficient. Show that R_n is independent of $\bar{X}, \bar{Y}, \hat{\sigma}_1, \hat{\sigma}_2$.

Optimal conditional testing

Ex. 4.27 *Conditional tests.* Suppose in general terms that data y are observed from a model with parameter θ , where the null hypothesis $H_0: \theta \in \Theta_0$ is to be tested, against the alternative that $\theta \notin \Theta_0$. Assume one computes $U = U(y)$ and $V = V(y)$. A *conditional test*, with respect to V , with level α , is then to find a rejection region $R(v)$, using the distribution of U given $V(y) = v$, with

conditional tests

$$\Pr_{\theta}\{U(Y) \in R(v) \mid V(Y) = v\} \leq \alpha \quad \text{for all } \theta \in \Theta_0.$$

Such tests are natural and important in multiparameter setups, as we shall see, and various constructions succeed in ‘reducing the dimensionality’ down to the analysis of a one-parameter family, where e.g. Neyman–Pearson more readily applies.

(a) Even when such a test has been constructed in a conditional modulus, ‘what is unlikely null behaviour of U given that $V = v$ ’, it may of course be translated or paraphrased without the conditioning: one rejects if $U(Y) \in R(V)$. Show that the test also has unconditional level α .

(b) From unconditional to conditional: Conditional tests as above have the form $T(U, V) = I(U \in R(V))$, built to have $E_{\theta}\{I(U \in R(v)) \mid v\} = \alpha$. Assume now that V is complete at the boundary $\partial\Theta_0$ of the null hypothesis parameter region; see Ex. 4.24. Show that if *any* test $T(U, V)$ has constant level α at this boundary, then it is a conditional test, with this level; $E_{\theta}\{T(U, V) \mid V = v\} = \alpha$ for all $\theta \in \bar{\Theta}_0$. (It might be useful to check the function $h(v) = E_{\theta}\{T(U, V) \mid V = v\} - \alpha$.)

Ex. 4.28 *Conditional tests: pairs of exponentials.* Suppose $X \sim \text{Expo}(a)$ and $Y \sim \text{Expo}(a + \delta)$, and that one wishes to test $\delta = 0$, i.e. equal distributions, against $\delta > 0$.

the arguments used in this simple setup are the same as for the much more general cases below

(a) Show that the joint density may be written $a(a + \delta) \exp(-az - \delta y)$, with $z = x + y$. Find the distribution of $Z = X + Y$, and show that the distribution of Y given $Z = z$ has the density

$$g_{\delta}(y \mid z) = \frac{\delta \exp(-\delta y)}{\int_0^z \delta \exp(-\delta y') dy'} = \frac{\delta \exp(-\delta y)}{1 - \exp(-\delta z)} \quad \text{for } 0 \leq y \leq z.$$

In particular, it does not depend on the θ . For the null hypothesis case of $\delta = 0$, show that $Y \mid z$ is uniform on $[0, z]$.

(b) The natural conditional 0.05 level test is then to first compute z , followed by rejection if $y \leq 0.05z$. Show that this procedure indeed has level 0.05, and that is the power optimal test among all conditional tests, using Y given z . Verify that this conditional test is the same as the unconditional test of rejecting when $R = Y/(X + Y) \leq 0.05$. Compute the power function of the $T^* = I(Y \leq 0.05Z)$ test (in the testing function parlance of Ex. 4.8), conditional on z , and unconditionally.

(c) At the boundary of the null, where $\delta = 0$, show that Z is complete. Show hence that *any* test with level 0.05 also must be a conditional on Z 0.05 level test, via Ex. 4.28.

(d) We know that $T^*(y, z) = I(y \leq 0.05z)$ is the most powerful conditional test with level 0.05; we now wish to extend this statement to T^* actually being the most powerful among *all* tests with level 0.05. For any competing test $T(Y, Z)$ with level 0.05, show, since it must be a z -conditional 0.05 level test, where it cannot beat T^* , that

$$E_{a,\delta} \{T^*(Y, Z) | z\} \geq E_{a,\delta} \{T(Y, Z) | z\} \quad \text{for all } \delta > 0, z > 0.$$

There is equality, to 0.05, at $\delta = 0$. Show from this that T^* is more powerful than such T , unconditionally; in suitable power function symbols, $\pi_{T^*}(a, \delta) \geq \pi_T(a, \delta)$ for all $\delta > 0$.

(e) Suppose now that there are m independent pairs, $X_i \sim \text{Expo}(a_i)$ and $Y_i \sim \text{Expo}(a_i + \delta)$, with sums $Z_i = X_i + Y_i$; there are hence $m+1$ parameters with $2m$ data points. Show that the optimal test is to reject when $U_m = Y_1 + \cdots + Y_m$ is small, given $z_1, \dots, z_m$. Explain how the null distribution of U_m can be evaluated via simulations. For an illustration, suppose three pairs (x_i, y_i) are observed: (0.927, 0.819), (1.479, 0.408), (3.780, 1.311). Carry out the test of δ , and compute the p-value.

Ex. 4.29 Conditional tests: normal. Consider a pair of normals, where interest lies in assessing their difference in means. This may of course be parametrised in different ways, but one natural way is $X \sim N(\theta, 1)$ and $Y \sim N(\theta + \delta, 1)$. One wishes to test $\delta = 0$ vs. $\delta > 0$, equivalent, of course, to testing equality of the means vs. $EY > EX$.

(a) Show that the joint density function can be written

$$\begin{aligned} f(x, y, \theta, \delta) &= (2\pi)^{-1} \exp\left[-\frac{1}{2}\{(x - \theta)^2 + (y - \theta - \delta)^2\}\right] \\ &= (2\pi)^{-1} \exp\left\{\theta z + \delta y - \frac{1}{2}x^2 - \frac{1}{2}y^2 - \frac{1}{2}\theta^2 - \frac{1}{2}(\theta + \delta)^2\right\}, \end{aligned}$$

where $z = x + y$, and with the main interaction between parameters and data being in the $\theta z + \delta y$ part.

(b) Show that (Y, Z) is a binormal, and set up its mean vector and variance matrix. Then use Ex. 1.44, or other algebraic methods, to show that $Y | (Z = z) \sim N(\frac{1}{2}(z + \delta), \frac{1}{2})$; in particular, its conditional distribution does not depend on θ .

(c) Through the conditioning on Z the testing problem has been reduced from a two-parameter to a one-parameter situation. For $Y | z \sim N(\frac{1}{2}(z + \delta), \frac{1}{2})$, show that the optimal test is to reject when $Y - \frac{1}{2}z > (1/\sqrt{2})c$, with $c = \Phi^{-1}(1 - \alpha)$ the standard normal quantile.

(d) Show that the above test, constructed to be optimal in the model for $Y | z$, is equivalent to that of rejecting when $D = Y - X > \sqrt{2}c$. – So the conditional test is an ordinary unconditional test in disguise, or vice versa, in this particular situation. One may argue directly that the D test is a natural one; the point of the extra arguments above is to establish that the D test is optimal. Similar comments apply to many cases below.

(e) Consider now m pairs of normal data, of the form $X_i \sim N(\theta_i, 1)$ and $Y_i \sim N(\theta_i + \delta, 1)$. Do the math, with the steps above. This involves the log joint density $\sum_{i=1}^m (\theta_i z_i + \delta y_i)$, with $z_i = x_i + y_i$. The optimal conditional test is then to reject when $\sum_{i=1}^k Y_i$ is big given $z_1, \dots, z_k$. Simplify the description of the final test.

Ex. 4.30 Conditional tests: Poisson. There are optimal tests for the proportionality factor for Poisson parameters.

(a) We start with a single pair of Poissons, x with mean θ , y with mean $\theta\gamma$. Show that the joint distribution becomes $\exp(-\theta - \theta\gamma)\theta^{x+y}\gamma^y/(x!y!)$. This inspires inspecting the distribution of y given $z = x + y$. Show that $y | z \sim \text{binom}(z, \gamma/(\gamma + 1))$.

also called the
likelihood
function

(b) To test $\gamma = 1$ against $\gamma > 1$, describe in details the natural conditional test which rejects when y is big, given $z = x + y$.

(c) Next consider independent Poisson pairs x_i, y_i for $i = 1, \dots, m$, where x_i has mean θ_i and y_i mean $\theta_i \gamma$. The model hence has $m+1$ parameters for the $2m$ observations, with γ the common multiplicative factor. Show that the joint distribution may be written

$$f = \exp \left[- \sum_{i=1}^m (\theta_i + \theta_i \gamma) + \sum_{i=1}^m \{ (x_i + y_i) \log \theta_i + y_i \log \gamma \} \right] \frac{1}{x_1! y_1! \cdots x_m! y_m!}.$$

With $z_i = x_i + y_i$, find the distribution of $y_i | z_i$, and also the distribution of $S = \sum_{i=1}^m y_i$ given $z_1, \dots, z_m$.

(d) Find the power optimal test for $\gamma = 1$ against $\gamma > 1$, among all those based on S given $z_1, \dots, z_m$.

(e) A mild but useful extension of the above is as follows. Assume $Y_0 \sim \text{Pois}(m_0 \theta_0)$ and $Y_1 \sim \text{Pois}(m_1 \theta_1)$, with m_0 and m_1 exposure levels, e.g. corresponding to observation time intervals. With interest being in the ratio parameter $\gamma = \theta_1 / \theta_0$, show that $Y_1 | (Z = z)$ is binomial $(z, m_1 \gamma / (m_0 + m_1 \gamma))$, and give an optimal test for $\gamma = 1$ vs. $\gamma > 1$.

Ex. 4.31 *Conditional tests: 2×2 tables.* For 2×2 tables there are optimal tests, as we learn here.

(a) Consider two binomials $y_0 \sim \text{binom}(m_0, p_0)$ and $y_1 \sim \text{binom}(m_1, p_1)$. The outcomes in such situations are often presented as a two-by-two table,

$$\begin{array}{cc} y_0, & m_0 - y_0 \\ y_1, & m_1 - y_1 \end{array}$$

Consider the so-called logistic parametrisation

$$p_0 = H(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} \quad \text{and} \quad p_1 = H(\theta + \gamma) = \frac{\exp(\theta + \gamma)}{1 + \exp(\theta + \gamma)}.$$

Show that $\theta = \log\{p_0/(1 - p_0)\}$ and $\theta + \gamma = \log\{p_1/(1 - p_1)\}$ in terms of the so-called log-odds. Show that the joint distribution can be written

$$f = \binom{m_0}{y_0} \binom{m_1}{y_1} \frac{\exp(\theta(y_0 + y_1))}{\{1 + \exp(\theta)\}^{m_0}} \frac{\exp(\gamma y_1)}{\{1 + \exp(\theta + \gamma)\}^{m_1}}.$$

(b) This inspires reaching inference for γ via the conditional distribution of y_1 given $z = y_0 + y_1$. Show that this distribution becomes

$$g_\gamma(y_1 | z) = \binom{m_0}{z - y_1} \binom{m_1}{y_1} \exp(\gamma y_1) / \sum_{y'_1 \leq \min(m_1, z)} \binom{m_0}{z - y'_1} \binom{m_1}{y'_1} \exp(\gamma y'_1)$$

for $y_1 = 0, 1, \dots, \min(m_1, z)$. In particular, this so-called excentric hypergeometric distribution depends on γ but not θ . We recognise the ordinary hypergeometric for $\gamma = 0$; see Ex. 1.67.

the excentric
hypergeometric

(c) Show that the optimal 0.05 level conditional test for the null hypothesis of equality, $p_0 = p_1$, is to reject when $y_1 > c(z)$, with $c(z)$ the highest number with $\sum_{0 \leq y_1 \leq c(z)} g_0(y_1 | z) \leq 0.95$.

(d) Assume there are k two-by-two tables, or pairs of binomials, with probabilities $p_{i,0} = H(\theta_i)$ and $p_{i,1} = H(\theta_i + \gamma)$, hence $k+1$ parameters for the $2k$ outcomes. Construct the optimal test for $\gamma = 0$, i.e. equality of probabilities $p_{i,0} = p_{i,1}$, vs. $\gamma > 0$. This is in fact the conditional test for $\sum_{i=1}^k y_{1,i}$ given $z_1, \dots, z_k$, with $z_i = y_{i,0} + y_{i,1}$.

this is used in
Story #13

Ex. 4.32 *The t test as an optimal conditional test.* Let $Y_1, \dots, Y_n$ be i.i.d. from the normal (ξ, σ^2) , where we wish to test $\xi = 0$ against $\xi > 0$. The canonical classical test, of level say 0.05, is based on $t = \sqrt{n}\bar{Y}/\sigma$, rejecting if $t \geq t_{n-1, 0.95}$, the upper 0.05 point of the t_{n-1} distribution; see also Ex. 3.7 and (3.4). We cannot use the Neyman–Pearson lemma directly to demonstrate optimality of the t test, however. One of several optimality properties may be derived via conditioning.

(a) Write $U = \sqrt{n}\bar{Y}$ and $V = \sum_{i=1}^n Y_i^2$, so that in particular $W = \sum_{i=1}^n (Y_i - \bar{Y})^2 = V - U^2$. Show that the joint density of the data can be written

$$\begin{aligned} f &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \xi)^2\right\} \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left\{\frac{\xi}{\sigma^2} \sqrt{n}U - \frac{1}{2} \frac{1}{\sigma^2} V - \frac{1}{2} \frac{\xi^2}{\sigma^2}\right\}. \end{aligned}$$

Note that the testing problem is equivalent to testing $\lambda = 0$ against $\lambda > 0$, with $\lambda = \xi/\sigma^2$, the mathematics indicating that this is a parameter easier to work with than ξ .

(b) Find the distribution of $U | (V = v)$, and show in particular that it depends on the parameters only via $\lambda = \xi/\sigma^2$. It is convenient here to work with

$$T = \frac{U}{\{W/(n-1)\}^{1/2}} = (n-1)^{1/2} \frac{U}{(V - U^2)^{1/2}}.$$

(c) Show that the power optimal test, among all tests based on $U | (V = v)$, is to reject when U is big, say $U \geq c(v)$, with $\Pr_0\{U \geq c(v) | V = v\} = 0.05$. Then show that this is actually the same as the t test.

(d) Show via arguments used in Ex. 4.28 and in other places above that the t test is not merely optimal among all conditional tests, for given level α , i.e. given V , but among *all* tests with level α .

Ex. 4.33 *Conditional tests: multiparameter exponential models.* In the rather simple situation of Ex. 4.28, with the exponential pair $X \sim \text{Expo}(a)$ and $Y \sim \text{Expo}(a + \delta)$, with sum $Z = X + Y$, we learned that (i) there is a clear level α conditional test for $\delta = 0$ vs. $\delta > 0$, in terms of $Y | Z$; (ii) that test is uniformly most powerful against all $\delta > 0$, among all conditional tests; and (iii) all other level α competitors are in fact also Z -conditional. Hence the winning test, reject if $Y \leq \alpha Z$, is the uniformly most powerful level α test. – We shall see now that the same arguments essentially go through for the wide class of all exponential families. Consider data Y from a density of the form $f(y, a, b) = \exp\{aU(y) + b^t V(y) - k(a, b)\}h(y)$, as in Ex. 4.21, with one-dimensional U and p -dimensional V . Suppose we need to test $a = a_0$ against $a > a_0$, for some given null hypothesis value a_0 .

(a) We have seen in the exercise pointed to that $U | (V = v)$ has a density depending on a but not b , and that it has an exponential form. Assume for simplicity that the distribution of U is continuous; mild formalistic additional arguments are required if the distribution is discrete. Deduce that there is a most powerful conditional level α test, say $T^*(y) = I\{U > c(V)\}$, with $c(v)$ determined from $\Pr_{a_0}(U > c(v) | V = v) = \alpha$, and with consequent power function $\pi(a, b) = \Pr_{a,b}\{U > c(V)\}$.

(b) Then consider *any* competing test $T(U, V)$ with level α . From $E_{a_0, b} T(U(Y), V(Y)) = \alpha$, for all b , use completeness in b of $V(Y)$ for fixed a_0 to prove that

$$E_{a_0, b} \{T(U(Y), V(Y)) | V(y) = v\} = \alpha$$

for all v (except perhaps in a region of probability zero), and for all b . Thus the T competitor is also a level α V -conditional test, and we have proved that the conditional test is uniformly most powerful among *all* level α tests.

(c) The theory extends fruitfully to the case of testing $H_0: a \leq a_0$ against $a > a_0$. Show that the test $T^* = I\{U > c(V)\}$ above, with $c(v)$ determined from $\Pr_{a_0}(U > c(v) | V = v) = \alpha$ at the boundary, is still of level α . Then show that this test is uniformly most powerful against all competing tests with constant level α at the boundary $a = a_0$; one says that such tests are *unbiased*. This latter very mild limitation is in order for the completeness argument to go through.

(d) Above we have worked with one-sided alternatives, $a = a_0$ vs. $a > a_0$. Work through the details when the alternative is two-sided, i.e. $a \neq a_0$; the tests are still based on $U | (V = v)$.

(e) When $U | (V = v)$ has a discrete distribution, the arguments still go through, but one cannot expect to find $c(v)$ with e.g. $\Pr_{a_0}\{U > c(v) | V = v\} = 0.05$. There are two ways out of this mild quandary. The first is to be satisfied with level 0.042, say, if that is how close one comes to 0.050, by appropriate choice of $c(v)$. The other, if one pedantically insists on 0.05, is to finetune $c(v)$ such that $\Pr_{a_0}\{U > c(v) | V = v\}$ is just below 0.05, and then identify the probability r such that

$$\Pr_{a_0}\{U > c(v) | V = v\} + r \Pr_{a_0}\{U = c(v) | V = v\} = 0.05,$$

So one rejects if $U > c(V)$, or, but then with probability r , if $U = c(V)$.

Ex. 4.34 Optimal t testing in linear regression. Consider the classic linear regression setup with $Y_i \sim N(a + bx_i, \sigma^2)$ for $i = 1, \dots, n$, see Ex. 3.32, and for simplicity of presentation here we assume the x_i have been centred, so $\bar{x} = 0$. The three estimators are then $\hat{a} = \bar{Y}$, $\hat{b} = \sum_{i=1}^n x_i Y_i / M_n$, with $M_n = \sum_{i=1}^n x_i^2$, and $\hat{\sigma}^2 = Q_0 / (n - 2)$, with $Q_0 = \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}x_i)^2$. Methods and results below extend those for the simpler t test of Ex. 4.32.

(a) With $V = \sum_{i=1}^n Y_i^2$, show that $Q_0 = V - n\hat{a}^2 - M_n\hat{b}^2$, and that the joint density may be written

$$\begin{aligned} L &= \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2} \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - a - bx_i)^2\right\} \\ &= \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2} \frac{V}{\sigma^2} + \frac{a}{\sigma^2} n\hat{a} + \frac{b}{\sigma^2} M_n\hat{b} - \frac{1}{2} \frac{na^2 + M_nb^2}{\sigma^2}\right\}, \end{aligned}$$

which is in the exponential family form, with natural parameters $1/\sigma^2$, a/σ^2 , b/σ^2 .

(b) Explain from the general exponential family testing theory that the uniformly most powerful test for $b = 0$ against $b > 0$, among all level α tests is to reject when $\hat{b}$ is large enough, in its conditional distribution given $(V, \hat{a})$. We need to find $c(V, \hat{a})$ such that

$$\Pr_{a,\sigma}(\hat{b} > c(V, \hat{a}) | V, \hat{a}) = \alpha$$

under $b = 0$. But this may be transformed to

$$W = \frac{\hat{b}}{(V - n\hat{a}^2)^{1/2}} = \frac{\hat{b}}{(Q_0 + M_n\hat{b}^2)^{1/2}} = \frac{\hat{b}/Q_0^{1/2}}{(1 + M_n\hat{b}^2/Q_0)^{1/2}} > \frac{c(V, \hat{a})}{(V - n\hat{a}^2)^{1/2}}.$$

Here W is a smooth increasing function of $t = \hat{b}/(\hat{\sigma}/M_n)$, the classic t ratio, which has the t_{n-2} distribution under $b = 0$. Argue that W is therefore independent of $(V, \hat{a})$, and

that the optimal test is to reject when $t > t_0$, the upper α in the t_{n-2} distribution. You have now shown that the traditional t test is optimal.

(c) Generalise the above to general linear multiple regression models, giving the result that the traditional t tests, for each of the β_j coefficients, are optimal.

Ex. 4.35 *A generalised Poisson distribution.* For a count variable Y , consider the model with point probabilities

$$f(y, \lambda, \gamma) = k(\lambda, \gamma)^{-1} \lambda^y / (y!)^\gamma \quad \text{for } y = 0, 1, 2, \dots,$$

where $k(\lambda, \gamma)$ is the normalisation constant $\sum_{y=0}^{\infty} \lambda^y / (y!)^\gamma$. For $\gamma = 1$ we're back to ordinary $\text{Pois}(\lambda)$, with $k(\lambda, 1) = \exp(\lambda)$. A regression version model of this type is used in Story #28, to assess potential overdispersion in Poisson counts.

(a) Pick some λ , and compute and display curves of the mean $\xi(\lambda, \gamma)$ and the variance-to-mean ratio $\rho(\lambda, \gamma)$, for an interval of γ around 1. Show that this ratio is decreasing in γ ; hence $\gamma < 1$ indicates overdispersion and $\gamma > 1$ underdispersion, relative to the Poisson. Also show that the mean of $\log(Y!)$ is decreasing in γ .

(b) Show that the distribution is of the exponential family form, and that the sufficient statistics, after having observed a sample $Y_1, \dots, Y_n$, is $T = \sum_{i=1}^n Y_i$ and $U = \sum_{i=1}^n \log(Y_i!)$. Show also that the joint distribution of these two must take the form

$$g_n(t, u) = \exp\{t \log \lambda - u\gamma - r_n(\lambda, \gamma)\} h_n(y_1, \dots, y_n),$$

for appropriate functions r_n and h_n .

(c) For an observed sample $Y_1, \dots, Y_n$, to test the Poisson assumption, against overdispersion, show that the optimal test is to reject when U is sufficiently small, given $T = t$. In other words, with level the classic 0.05, for example, we reject when $U \leq u_0(t)$, where $u_0(t)$ is the 0.95 quantile of the distribution of U given $T = t$, computed at $\gamma = 1$, i.e. under Poisson conditions.

(d) There is no table or simple formula for the distribution of $U | (T = t)$, but show that it depends on γ , but not λ . Show that under $\gamma = 1$, $(Y_1, \dots, Y_n) | (T = t)$ is a multinomial with count t and probabilities $(1/n, \dots, 1/n)$. Explain then how the distribution of $U | (T = t)$ may be simulated under Poisson conditions.

Ex. 4.36 *Testing for correlation.* Consider binormal i.i.d. pairs (X_i, Y_i) , with parameters $\xi_1, \xi_2, \sigma_1, \sigma_2, \rho$, as in Ex. 1.58. Suppose we need to test $\rho = 0$ vs. $\rho > 0$.

(a) Argue as in the exercise pointed to that the binormal distribution is inside the exponential family class, with natural parameters

$$\frac{1}{1-\rho^2} \left(\frac{\xi_1}{\sigma_1^2} - \frac{\rho \xi_2}{\sigma_1 \sigma_2} \right), \frac{1}{1-\rho^2} \left(\frac{\xi_2}{\sigma_2^2} - \frac{\rho \xi_1}{\sigma_1 \sigma_2} \right), \frac{1}{(1-\rho^2)\sigma_1^2}, \frac{1}{(1-\rho^2)\sigma_2^2}, \frac{\rho}{(1-\rho^2)\sigma_1 \sigma_2},$$

associated with data functions $X_i, X_i^2, Y_i, Y_i^2, X_i Y_i$. Explain that the testing problem is equivalent to testing $\lambda = 0$ vs. $\lambda > 0$, where $\lambda = \rho / \{(1-\rho^2)\sigma_1 \sigma_2\}$.

(b) Then argue from the general testing theory for the exponential class that there for given testing level α is a uniformly most powerful test, consisting in rejecting the null provided $A_n = n^{-1} \sum_{i=1}^n X_i Y_i$ is big enough, in its conditional null distribution given $(\bar{X}, n^{-1} \sum_{i=1}^n X_i^2, \bar{Y}, n^{-1} \sum_{i=1}^n Y_i^2)$. Formally, the rejection threshold c , which depends on these variables, is determined by

$$\Pr_0\{A_n > c | (\bar{X}, n^{-1} \sum_{i=1}^n X_i^2, \bar{Y}, n^{-1} \sum_{i=1}^n Y_i^2)\} = \alpha,$$

a generalised
Poisson

with the footscript 0 indicating probability under $\rho = 0$.

(c) Write $\hat{\sigma}_1^2 = n^{-1} \sum_{i=1}^n X_i^2 - \bar{X}^2$ and $\hat{\sigma}_2^2 = n^{-1} \sum_{i=1}^n Y_i^2 - \bar{Y}^2$ for the empirical variances. Explain that in the conditional situation, given values for $\bar{X}, \bar{Y}, \hat{\sigma}_1, \hat{\sigma}_2$, the requirement above can be transformed to $\Pr_0\{R_n > d \mid (\bar{X}, \bar{Y}, \hat{\sigma}_1, \hat{\sigma}_2)\} = \alpha$, with R_n the empirical correlation coefficient, see (2.11).

(d) Then explain that R_n is actually independent of $\bar{X}, \bar{Y}, \hat{\sigma}_1, \hat{\sigma}_2$, under the null, so that the optimal test becomes the simpler one, of rejecting if $R_n > d$, where $\Pr_0(R_n > d) = \alpha$.

(e) For this optimal test regime to be specified fully, it remains to find the null distribution of R_n . Explain that this null distribution does not depend on the actual values of $(\xi_1, \xi_2, \sigma_1, \sigma_2)$, which we therefore may take to be $(0, 0, 1, 1)$. For practical purposes one may simulate a high number of R_n and then read off the d quantile; explain also that $\sqrt{n}R_n \rightarrow_d N(0, 1)$, from Ex. 2.55, so that $z_{1-\alpha}/\sqrt{n}$ is an approximation to the upper α point. We show however in the following point that $T_n = m^{1/2}R_n/(1 - R_n^2)^{1/2}$ has a t_m distribution, under the null, with $m = n - 2$ degrees of freedom. With e.g. $n = 25$, find the 0.95 quantile of the R_n distribution.

(f) So let us find the null distribution for R_n . Write $s_x^2 = \sum_{i=1}^n (X_i - \bar{X})^2$, $s_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$, $s_{xy} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$, so that $R_n = s_{xy}/(s_x s_y)$. We may now use results from linear regression of the Y_i with respect to the X_i . Study in particular the least squares estimators $\hat{a}, \hat{b}$, minimising $Q(a, b) = \sum_{i=1}^n \{Y_i - a - b(X_i - \bar{X})\}^2$, as with Ex. 3.32. Using results from that exercise, show (i) that $\hat{a} = \bar{Y}$ and $\hat{b} = s_{xy}/s_x^2$; (ii) that $Q_0 = Q(\hat{a}, \hat{b}) = s_y^2 - s_x^2 \hat{b}^2$; and (iii) that $T = \hat{b}/\{(Q_0/m)^{1/2}/s_x\} = m^{1/2} s_x \hat{b}/Q_0^{1/2}$ has the t_m distribution, with $m = n - 2$ degrees of freedom. Show that this implies

$$R_n = \frac{\hat{b} s_x}{s_y} = \frac{s_x \hat{b}}{(Q_0 + s_x^2 \hat{b}^2)^{1/2}} = \frac{T/m^{1/2}}{(1 + T^2/m)^{1/2}}.$$

Solve for T to get the $T = m^{1/2}R_n/(1 - R_n^2)^{1/2}$, pointed to and used in the previous point. Go on to work out the density of R_n , via the density g_m for the t_m ; show that it becomes

$$h_n(r) = g_m\left(\frac{m^{1/2}r}{(1-r^2)^{1/2}}\right) \frac{m^{1/2}}{(1-r^2)^{3/2}} = \frac{\Gamma(\frac{1}{2}(m+1))}{\Gamma(\frac{1}{2}m)\sqrt{\pi}} (1-r^2)^{(m-2)/2}$$

for $r \in (-1, 1)$. As a curiosum, check that it is U-shaped for $n = 3$, uniform for $n = 4$, and then bell-shaped for $n \geq 5$.

Ex. 4.37 Inference for linear multiple regression. For linear regression setups, we point back to Ex. 3.33, 3.34, 3.36 for basic estimation schemes and associated confidence intervals. Results there may be used also for testing purposes.

(a) As with Ex. 3.36, use $\hat{\beta}_j \sim N(\beta_j, r_j \sigma^2/n)$, with r_j the (j, j) element of Σ_n^{-1} , to construct a test for $\beta_j = 0$. Put also up a test for $\beta_j = \beta_\ell$, for two different coefficients.

(b) Construct a test for $\sigma = \sigma_0$, for some given null value. For two regression datasets, with similarities making it relevant to compare their standard deviations, construct a test for $\sigma_1 = \sigma_2$.

(c) Consider some quantile $\gamma = x_0^t \beta + \sigma z_0$ in the Y distribution for an x_0 of interest. Set up a test for $\gamma = \gamma_0$, for some given γ_0 .

Ex. 4.38 How much of the variance is explained? In the linear regression model, the extent to which the covariates influence the outcomes may be assessed in several ways,

for an
illustration,
with full
confidence, see
Story #7

one of which is to decompose the variance of the outcomes into a covariate part and a ‘remaining variability’ part. Such assessments relate also to ‘signal plus noise’ viewpoints; how strong is the signal?

(a) We start out writing the regression model as

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \beta_0 + x_i^t \beta + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

again with the ε_i seen as i.i.d. $N(0, \sigma^2)$, and further assume that the covariates have been centred, having their means subtracted, so that $\sum_{i=1}^n x_{i,j} = 0$ for $j = 1, \dots, p$. This gives β_0 the interpretation as the overall mean of the Y_i . With $\Sigma_n = (1/n) \sum_{i=1}^n x_i x_i^t$ the empirical $p \times p$ variance matrix for the x_i , show that the least squares estimators become

$$\hat{\beta}_0 = \bar{Y} \sim N(\beta_0, \sigma^2/n), \quad \hat{\beta} = \Sigma_n^{-1} (1/n) \sum_{i=1}^n x_i Y_i \sim N_p(\beta, (\sigma^2/n) \Sigma_n^{-1}),$$

and that these two are independent.

(b) Write $\hat{\mu}_i = \hat{\beta}_0 + x_i^t \hat{\beta}$ for the model based estimate of the outcome at x_i . For the sum of squared residuals, show that $Q_0 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 - n \hat{\beta}^t \Sigma_n \hat{\beta}$. In other words,

$$V_n = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 + n \hat{\beta}^t \Sigma_n \hat{\beta},$$

a neat decomposition of the full variability of outcomes as a sum of squared residuals and the covariate part $n \hat{\beta}^t \Sigma_n \hat{\beta}$.

(c) Standard themes for the linear regression model are developed with analyses carried out conditional on the covariates. Allow now a change in this narrative, where the x_i are seen as having their own covariate distribution, with mean zero and variance matrix Σ_n . Show that a randomly selected outcome Y_i then has variance $\beta^t \Sigma_n \beta + \sigma^2$. Show that the covariate part of the full variability becomes

$$\rho = \frac{\beta^t \Sigma_n \beta}{\beta^t \Sigma_n \beta + \sigma^2} = \frac{\lambda}{\lambda + 1}, \quad \text{with } \lambda = \beta^t \Sigma_n \beta / \sigma^2.$$

With $\tilde{\sigma}^2 = Q_0/n$ (rather than the unbiased $\hat{\sigma}^2 = Q_0/(n - p - 1)$), show that this leads to

$$\tilde{\rho} = \frac{\hat{\beta}^t \Sigma_n \hat{\beta}}{\hat{\beta}^t \Sigma_n \hat{\beta} + \tilde{\sigma}^2} = \frac{V_n - \sum_{i=1}^n (Y_i - \hat{\mu}_i)^2}{V_n} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\mu}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

This is often called the coefficient of determination, or R^2 .

(d) To carry out precise inference for ρ , show first that $n \hat{\beta}^t \Sigma_n \hat{\beta} / \sigma^2 \sim \chi_p^2(n\lambda)$; check with Ex. 1.53. Then show that with $\hat{\lambda} = \hat{\beta}^t \Sigma_n \hat{\beta} / \hat{\sigma}^2$, we have

$$F = n \hat{\lambda} / p = n \hat{\beta}^t \Sigma_n \hat{\beta} / (p \hat{\sigma}^2) \sim F(p, m, n\lambda),$$

the noncentral F with $m = n - (p + 1)$ the degrees of freedom for $\hat{\sigma}^2$.

(e) Explain how this may be used to set confidence intervals for λ and hence for ρ . You may check with Ex. 8.23 to see this being used to construct full confidence distributions for the fraction ρ of variation explained by the covariates; for an illustration, see Story #6.

Ex. 4.39 *Testing and confidence for linear combinations of regression parameters.* Consider again the linear regression model with $Y_i \sim N(x_i^t \beta, \sigma^2)$, with the least squares estimator $\hat{\beta} \sim N_p(\beta, (\sigma^2/n)\Sigma_n^{-1})$.

(a) A useful little lemma is as follows. Let x be a vector of length p , with A a symmetric positive definite $p \times p$ matrix. Show that the maximum of $c^t x$, under the condition $c^t A c = 1$, is $x^t A^{-1} x$. You may show this via Lagrange multipliers and a bit of linear algebra.

(b) In the regression problem, show that we for each c have $c^t \hat{\beta} \sim N(c^t \beta, (\sigma^2/n) c^t \Sigma_n^{-1} c)$, and that

$$V_n(c) = \frac{\sqrt{n}(c^t \hat{\beta} - c^t \beta)}{\hat{\sigma}(c^t \Sigma_n^{-1} c)^{1/2}} \sim t_m,$$

with $m = n - p$ degrees of freedom. Explain how this can be used to test $\gamma(c) = c^t \beta = 0$, for any given c , and construct a confidence interval for such a $\gamma(c)$.

(c) Will the hypothesis $\gamma(c) = 0$ be rejected, in that case for which linear combinations? Show that

$$V_n^* = \max_{\text{all } c} V_n(c) = (\sqrt{n}/\hat{\sigma})\{(\hat{\beta} - \beta)^t \Sigma_n (\hat{\beta} - \beta)\}^{1/2},$$

and that $(V_n^*)^2$ has a Fisher distribution with degrees of freedom $(p, n - p)$. Use this to construct joint confidence intervals, of the type $c^t \beta \in c^t \hat{\beta} \pm d \hat{\sigma}(c^t \Sigma_n^{-1} c)^{1/2}/\sqrt{n}$, to hold jointly with confidence level 0.95, say.

Ex. 4.40 *Inference for ratios of standard deviations.* Suppose two independent samples, of sizes n_1 and n_2 , come from two populations, with standard deviations σ_1 and σ_2 . From the empirical standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$, form the ratio $R = \hat{\sigma}_1/\hat{\sigma}_2$, to be used for inference about the underlying ratio $\rho = \sigma_1/\sigma_2$.

(a) Suppose first that the two distributions are normal. We then saw in Ex. 1.55 that $R^2 = \rho^2 F$, where $F \sim F_{m_1, m_2}$, an F distribution with degrees of freedom $(m_1, m_2) = (n_1 - 1, n_2 - 1)$. Construct a 95 percent confidence interval for ρ based on this. Also give a 0.05 level test for the equality hypothesis $\sigma_1 = \sigma_2$. this gives c.i. for ρ , and tests for $\rho = 1$.

(b) Then consider inference for the ratio ρ outside the assumption of normally distributed data. From Ex. 3.12, find the representation

$$R = \frac{\hat{\sigma}_1}{\hat{\sigma}_2} \doteq \frac{\sigma_1}{\sigma_2} \frac{1 + (1/n_1)^{1/2}(\frac{1}{2} + \frac{1}{4}\gamma_{4,1})^{1/2}N_{n_1}}{1 + (1/n_2)^{1/2}(\frac{1}{2} + \frac{1}{4}\gamma_{4,2})^{1/2}N_{n_2}},$$

in terms of the kurtoses $\gamma_{4,1}$ and $\gamma_{4,2}$, where N_{n_1} and N_{n_2} are independent variables tending to standard normals as sample sizes increase. Use the delta method to deduce that $R/\rho \approx_d N(1, \tau^2)$, with $\tau^2 = (1/n_1)(\frac{1}{2} + \frac{1}{4}\gamma_{4,1}) + (1/n_2)(\frac{1}{2} + \frac{1}{4}\gamma_{4,2})$. For normal data, show that this matches the distributional approximation $F^{1/2} \approx_d N(1, 1/(2n_1) + 1/(2n_2))$.

(c) Construct an approximate 95 percent confidence interval for σ_1/σ_2 , valid also outside normal data. For an application, see the Bach and Reger Story #53.

Ex. 4.41 *Testing equality across groups.* Suppose a parameter has the same interpretation across groups, say θ_j for groups $j = 1, \dots, k$. associated with estimators, perhaps of different precision. How can we test the null hypothesis of no difference, i.e. $H_0: \theta_1 = \dots = \theta_k$?

(a) Suppose first that $Y_j \sim N(\theta_j, \sigma^2)$, with independence, and the same known precision, i.e. σ . Explain from classic results of Ex. 1.54–1.49, that $Z = \sum_{j=1}^k (Y_j - \bar{Y})^2 / \sigma^2 \sim \chi_{k-1}^2$, with $\bar{Y} = (1/k) \sum_{j=1}^k Y_j$ the natural estimator for the common parameter under H_0 . Explain that in the case of σ not being known, we still have $Z^* = \sum_{j=1}^k (Y_j - \bar{Y})^2 / \hat{\sigma}^2 \rightarrow_d \chi_{k-1}^2$, as long as $\hat{\sigma}$ is consistent, with growing information for each group.

(b) A useful generalisation of the basic results for i.i.d. normal data is as follows. Suppose $Y_j \sim N(0, 1/a_j)$, with the a_j positive and $a = \sum_{j=1}^k a_j$. Define $Y^* = \sum_{j=1}^k (a_j/a) Y_j$ and then $Z = \sum_{j=1}^k a_j (Y_j - Y^*)^2$. Show that $Z = \sum_{j=1}^k a_j Y_j^2 - a (Y^*)^2$, that it is independent of Y^* , and that its distribution is a χ_{k-1}^2 . The classical case, leading to the distribution of the empirical variance and then to the t test, corresponds to all $a_j = 1$.

(c) A natural variation of the previous point is as follows. Assume in general terms that $Y_j \sim N(\theta, \sigma_j^2)$, with independence, for $j = 1, \dots, k$. Show that the minimiser of $Q(\theta) = \sum_{j=1}^k (Y_j - \theta)^2 / \sigma_j^2$ is $\hat{\theta} = (\sum_{j=1}^k Y_j / \sigma_j^2) / (\sum_{j=1}^k 1 / \sigma_j^2)$, and that $Q_{\min} = Q(\hat{\theta})$ is χ_{k-1}^2 , independent of $\hat{\theta}$. Demonstrate also that of all unbiased estimators $\theta^* = \sum_{j=1}^k c_j Y_j$ of θ , $\hat{\theta}$, with weights proportional to inverse variances, has the smallest variance. Again, if the σ_j are consistently estimated, as opposed to known, explain that with $\theta^* = (\sum_{j=1}^k Y_j / \hat{\sigma}_j^2) / (\sum_{j=1}^k 1 / \hat{\sigma}_j^2)$, we still have $Q^* = \sum_{j=1}^k (Y_j - \theta^*)^2 / \hat{\sigma}_j^2$ tending to χ_{k-1}^2 .

(d) Consider correlations $\rho_1, \dots, \rho_k$ for k groups of binormal data, with sample sizes $n_1, \dots, n_k$, summing to the total n . Via Ex. 2.55, explain that $n_j^{1/2}(\hat{\zeta}_j - \zeta_j) \rightarrow_d N(0, 1)$, where $\hat{\zeta}_j = \frac{1}{2} \log\{(1 + \hat{\rho}_j)/(1 - \hat{\rho}_j)\}$ is Fisher's zeta transform. Deduce that $Q = \sum_{j=1}^k n_j (\hat{\zeta}_j - \zeta^*)^2 \rightarrow_d \chi_{k-1}^2$, with $\zeta^* = \sum_{j=1}^k (n_j/n) \hat{\zeta}_j$. This is accordingly a natural test statistic for $\rho_1 = \dots = \rho_k$.

(e) Situations similar to that of the previous point abound in applied statistics, when needing to compare more groups than two. Suppose $\hat{p}_j = Y_j/n_j$ is estimating a binomial p_j , for independent experiments, for $j = 1, \dots, k$. Then form $\hat{p} = \sum_{j=1}^k (n_j/n) \hat{p}_j$, and show that under $p_1 = \dots = p_k$ conditions, $Q = \sum_{j=1}^k n_j (\hat{p}_j - \hat{p})^2 \rightarrow_d p(1-p) \chi_{k-1}^2$, using p to denote the common value of the p_j . Construct a test for equality of the p_j .

Ex. 4.42 *Testing equality across groups, vector parameter case.* In Ex. 4.41 we developed a recipe for testing equality across groups for any one-dimensional parameter. Here we lift those methods to the case where the $\theta_1, \dots, \theta_k$ in question have dimension $p \geq 2$. We wish again to test equality of these.

(a) Assume $Y_j \sim N_p(\theta_j, \Sigma_j)$ for $j = 1, \dots, k$, with known positive definite variance matrices. Under H_0 of $\theta_1 = \dots = \theta_k$, show that $Q(\theta) = \sum_{j=1}^k (Y_j - \theta)^t \Sigma_j^{-1} (Y_j - \theta)$ is a χ_{kp}^2 , and that it is minimised by $\hat{\theta} = A^{-1} \sum_{j=1}^k \Sigma_j^{-1} Y_j$, with $A = \sum_{j=1}^k \Sigma_j^{-1}$. Show further (i) that $\hat{\theta} \sim N_p(\theta, A^{-1})$; (ii) that the full ensemble of $Y_j - \theta$ is independent of $\hat{\theta}$; and (iii) that the natural test statistic $Q_0 = Q(\hat{\theta}) = \sum_{j=1}^k (Y_j - \hat{\theta})^t \Sigma_j^{-1} (Y_j - \hat{\theta})$ is a $\chi_{(k-1)p}^2$.

(b) This Q_0 may be used as a test statistic for $\theta_1 = \dots = \theta_k$, with an exact chi-square null distribution, provided the Σ_j matrices are known. In various applications these are estimated from data, with $\hat{\Sigma}_j$ using data for group j . This leads to $\hat{\theta} = \hat{A}^{-1} \sum_{j=1}^k \hat{\Sigma}_j^{-1} Y_j$, with $\hat{A} = \sum_{j=1}^k \hat{\Sigma}_j^{-1}$. Show that if these $\hat{\Sigma}_j$ are consistent, then $Q^* = Q^*(\hat{\theta}) = \sum_{j=1}^k (Y_j - \hat{\theta})^t \hat{\Sigma}_j^{-1} (Y_j - \hat{\theta})$ tends to the $\chi_{(k-1)p}^2$, under the null.

(c) Suppose we have normal datasets with sample sizes n_j , yielding the usual parameter

estimates $(\hat{\xi}_j, \hat{\sigma}_j)$, for groups $j = 1, \dots, k$. From these compute $\hat{\xi} = \sum_{j=1}^k (n_j/n) \hat{\xi}_j$, the overall mean, and $\hat{\sigma} = \prod_{j=1}^k \hat{\sigma}_j^{n_j/n}$. Show that

$$Q = \sum_{j=1}^k \{n_j(\hat{\xi}_j - \hat{\xi})^2 + 2n_j(\log \hat{\sigma}_j - \log \hat{\sigma})^2\}$$

tends to $\chi_{2(k-1)}^2$ under the hypothesis of common (ξ, σ) . Write also down the Wilks test, based on attained log-likelihood maxima. (xx do both, for the mothers and babies story, for the three ethnic groups. xx)

(d) Consider a probability distribution $b = (b_1, \dots, b_s)$ for suitable outcomes $1, \dots, s$, and suppose we have multinomial estimators $\hat{b}_j = (\hat{b}_{j,1}, \dots, \hat{b}_{j,s})$ for each of groups $j = 1, \dots, r$, resulting from sample sizes $n_1, \dots, n_r$, with total sum n . In other words, $\hat{b}_j = (N_{j,1}, \dots, N_{j,s})/n_j$, for the multinomial experiment for group j . The hypothesis to be tested is that of equal probability distribution across groups. Explain that $\hat{b}_j \approx_d N_s(b, \Sigma/n_j)$, where Σ_j has $b_j(1 - b_j)$ on the diagonal and $-b_j b_\ell$ outside. Write Σ_0 for the $(s-1) \times (s-1)$ submatrix of Σ , needed since Σ does not have full rank, with $\hat{b}_0$ and $\hat{b}_{j,0}$ similarly being the shorter $(s-1)$ -length vectors, so that the general method above can be applied. Then show that the test statistic Q_0 from the general recipe above can be written

$$\sum_{j=1}^k n_j (\hat{b}_{j,0} - \hat{b}_0)^t \hat{\Sigma}_0^{-1} (\hat{b}_{j,0} - \hat{b}_0) = \sum_{j=1}^k n_j \sum_{r=1}^s \frac{(\hat{b}_{j,r} - \hat{b}_r)^2}{\hat{b}_r} = \sum_{j=1}^k \sum_{r=1}^s \frac{(N_{j,r} - E_{j,r})^2}{E_{j,r}},$$

which then tends to $\chi_{(k-1)(s-1)}^2$ under the null. You may check the Pearson statistic from Story #58, and check that this is precisely identical to the test above, for testing independence between the two factors ‘group’ and ‘bin’.

Ex. 4.43 One-way layout. The most prominent special case of the setup from Ex. 4.41 is that with an i.i.d. normal sample for each of the groups in question, with the same variance for all data. Consider therefore $Y_{j,1}, \dots, Y_{j,n_j}$ i.i.d. from $N(\xi_j, \sigma^2)$, for group j , and with overall sample size $n = \sum_{j=1}^k n_j$. This is the so-called normal one-way layout, with the chief hypothesis to be tested is $H_0: \xi_1 = \dots = \xi_k$.

(a) Let $\bar{Y}_j = (1/n_j) \sum_{r=1}^{n_j} Y_{j,r}$ be the group averages. Explain that we for the overall average have $\bar{Y} = (1/n) \sum_{j=1}^k \sum_{r=1}^{n_j} Y_{j,r} = \sum_{j=1}^k (n_j/n) \bar{Y}_j$. Show then that $Q = \sum_{j=1}^k n_j (\bar{Y}_j - \bar{Y})^2$ has $E Q = (k-1)\sigma^2 + \sum_{j=1}^k n_j (\xi_j - \bar{\xi})^2$, with $\bar{\xi} = \sum_{j=1}^k (n_j/n) \xi_j$. Explain that $Q \sim \sigma^2 \chi_{k-1}^2$, under equality of means, so that the test statistic $W = Q/\hat{\sigma}^2$ approximately has the χ_{k-1}^2 distribution under H_0 , as long as $\hat{\sigma}$ is consistent. This is already a satisfactory answer to the one-way layout testing problem, without further fine-tuning.

(b) In this setup further fine-tuning is available, however. Show first that

$$Q_0 = \sum_{j=1}^k \sum_{r=1}^{n_j} (Y_{j,r} - \bar{Y}_j)^2 \sim \sigma^2 \chi_{n-k}^2,$$

making $\hat{\sigma}^2 = Q_0/(n-k)$ the natural unbiased estimator of σ^2 . Explain also that Q_0 is independent of Q . Under the null of equal means, then, deduce that

$$F = \frac{Q/(k-1)}{Q_0/(n-k)} = \frac{Q/(k-1)}{\hat{\sigma}^2} = \frac{\sum_{j=1}^k n_j (\bar{Y}_j - \bar{Y})^2 / (k-1)}{\sum_{j=1}^k \sum_{r=1}^{n_j} (Y_{j,r} - \bar{Y}_j)^2 / (n-k)}$$

has the F distribution $F_{k-1, n-k}$. Checking whether an observed F is too big, compared to this null distribution, is then a way to assess the hypothesis $\xi_1 = \dots = \xi_k$.

illustrations:
see Stories #6,
#20

Ex. 4.44 *Testing equality of multinormal means.* (xx a brief thing, used in Story #45, can point to ML theory too. establish the following, then apply in two-three settings. xx) Suppose $A \sim N_p(a, \Sigma_1)$ and $B \sim N_p(b, \Sigma_2)$ are independent multinormal data vectors, with known variance matrices. How can we test $a = b$? Show that $W = (B - A)^t(\Sigma_1 + \Sigma_2)^{-1}(A - B) \sim \chi_p^2$ under the null.

Notes and pointers

(xx confidence intervals. testing. connections. power. Neyman–Pearson. point to Lehmann. and to later chapters, Ch. 8 for CDs. point to interplay between modelling, probability calculus, thinking, a bit of philosophy, and practice. xx)

A long generation of Norwegian statisticians, actuaries, and insurance mathematicians have grown up with the sentence “*Et utsagn hvis feilaktige benektelse vi fortrinnsvis ønsker å unngå kalles nullhypotesen*”, from the curriculum based on E. Sverdrup’s two books. Discuss how ‘the null hypothesis is the statement the incorrect negation of which is the error one preferably seeks to avoid’ can help you set up the H_0 , in a couple of applications.

(xx Wald tests: after A. Wald. xx)

(xx completeness in expo families: check BickelDoksum or Johansen or Brown or Schervish. perhaps requiring analytizing continuation arguments. point is that $a(\theta) = E_\theta h(\bar{T})$ is a super smooth functions with all derivatives smooth. xx)

The two-parameter generalised Poisson model of Ex. 4.35 is from Schweder and Hjort (2016, Examples 4.18, 8.16).

(xx God Fisk: for completeness lemma: check BickelDoksum or Johansen or Brown or Schervish. perhaps requiring analytizing continuation arguments. point is that $a(\theta) = E_\theta h(\bar{T})$ is a super smooth. xx)

(xx one-way layout: primarily to test equality of means, so ‘anova’ is arguably a slight misnomer. xx)

Basu’s lemma: from Basu (1955), see also Ghosh (2002).

ToDo notes, as of 16 August 2026:

Lots, though the chapter is shaping up. There are lots of ‘test $\theta = \theta_0$ against $\theta > \theta_0$ ’ prose in exercises, since it’s easiest and cleanest, with NP etc. But we need to say a couple of times that all of this generalises to $\theta \neq \theta_0$ etc.

the Lindqvist and Taraldsen things, for simulating from $U(x) | \{V(x) = v\}$. do the handball model from CLP.

Do the bioequivalence: test the H_0 that θ is outside $[-\varepsilon, \varepsilon]$ against the alternative that it is inside. Diffent type of tests, and different looking power function.

Do sample size things, and efficiency; $\pi_n(\theta) \doteq \Phi(\sqrt{n}(\theta - \theta_0)/\sigma - z_0)$, with informative derivative $\phi(0)\sqrt{n}/\sigma$ at θ_0 . Efficiency things.

Make an example or two, perhaps with Cauchy again, to see that the confidence region might not be an interval.

Point to Cox (1958) and also interviews with him regarding conditional stuff.

I.5

Maximum likelihood

Consider the joint density of a dataset Y from a parametric model, say $f_{\text{full}}(y, \theta)$. The likelihood function, a fundamental concept in parametric inference, is just this density, but seen as a function of θ , with y fixed at the observed dataset y_{obs} . In this chapter we go through the fundamental likelihood inference methods, in particular with results for the maximum likelihood estimator, the maximiser of the likelihood, and for the attained maximum value itself. It is useful to see this general likelihood theory as a special case of the more flexible machinery of minimum divergence methods, to be returned to in Ch. 6. The divergence associated with maximum likelihood theory is Kullback–Leibler. The likelihood methods are practical and versatile, as is demonstrated in several exercises and stories, also for say non-standard regression models. With models outside the most familiar ones, inference analysis essentially flows from being able to programme the log-likelihood function. Further material connected to likelihood theory are Cramér–Rao information inequalities, limiting normality and Wilks theorems. Crucially, the theory developed does not in general presuppose that the parametric model worked with is correct, as results are established both under and outside the precise model conditions.

Key words: Cramér–Rao lower bounds, Fisher information matrix, Kullback–Leibler, least false parameters, log-likelihood, maximum likelihood, regression models, sandwich matrix, score function, Wilks theorems

In earlier chapters we have met and worked with several methods for estimation parameters in different settings. Sometimes one estimates population parameters directly from data, like the mean, the median, the standard deviation, the skewness, the correlation, the median difference between groups, etc., without necessarily using parametric models. Very frequently, however, the most fruitful data analyses involve fitting some parametric model to the data, as in regression models, where regression coefficients are estimated and assessed to learn how covariates influence the main outcomes. In Chs. 3–4 we have already worked with methods associated with moment fitting, quantile fitting, and least squares for regression models, but the present chapter has a wider aim and applicability, specifically developing methodology for *minimum divergence function estimators*, and its primary special case, *maximum likelihood estimators*. There is extensive use of likelihood theory in Chs. 7, 8, 11, 12, 13.

In general terms, suppose data y have been modelled via a parametric model, leading to a joint probability density $f_{\text{full}}(y, \theta)$, with θ in some relevant parameter region Θ . Then

the *likelihood function* is $L(\theta) = f_{\text{full}}(y_{\text{obs}}, \theta)$, studied as a function of the parameters, with y held fixed at the observed y_{obs} . The maximum likelihood (ML) estimator is the value $\hat{\theta}$ maximising the likelihood, or equivalently the *log-likelihood function* $\ell(\theta) = \log f_{\text{full}}(y_{\text{obs}}, \theta)$. The simplest setup is that of i.i.d. observations $Y_1, \dots, Y_n$ to be fitted to some parametric $f(y, \theta)$, with θ a parameter vector inside its relevant parameter region. The log-likelihood function is then

the ML
estimator

log-likelihood
function

$$\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta). \quad (5.1)$$

The apparatus of likelihood estimation carries over to regression models too, where the density of Y_i given a covariate vector x_i is modelled via some $f(y_i | x_i, \theta)$.

Generally speaking there are two valid viewpoints when developing the required theory. The first takes *the model to be correct*, so there is a true parameter θ_0 to be estimated, with associated inference. The second takes the parametric model to be a *sensible approximation* to the real and unknown data-generating mechanism. With i.i.d. data, the model density $f(y, \theta)$ is consequently seen as an approximation to the real underlying density $g(y)$. Estimation and inference then involves the *least false parameter* θ_0 , in a suitable sense making $f(y, \theta_0)$ coming as close to the real $g(y)$ as possible.

the agnostic
view, with least
false parameter
values

A generic and powerful statistical idea is to estimate parameters by minimising a relevant distance or divergence from the true mechanism to the data. The arguably most important estimation method, the *maximum likelihood method*, is the special case associated with the *Kullback–Leibler divergence* $\text{KL}(g, f_\theta) = \int \log(g/f_\theta) g \, dy$. Apart from clarifying the concepts, defining parameters and estimators, in such ways, developing the theory amounts to establishing clear results for the behaviour and performance of these estimators and related data-driven tests, confidence methods, etc. This again involves limiting normality, assessing and estimating variability, and so on. All of this leads to a fruitful and indeed practical general theory, which also in new situations, with models constructed for new purposes, allows the statistician to estimate, to reach relevant confidence statements, test and compare, predict, etc. In brief, this chapter has these connected parts, with associated correlated groups of exercises.

more on this in
Ch. 6

(i) Motivating and building an apparatus for defining parameters and their estimators via maximised likelihoods, initially for i.i.d. setups.

(ii) Analysing ML methods involves studying score functions, information functions, and the Fisher information matrix, for general parametric models. This again relates to ‘information inequalities’, specifically of the Cramér–Rao type, establishing lower bounds for how small variances can be, with a given model and a given sample size n . The ML methods achieve these lower bounds, under model conditions, for growing n .

(iii) Methods are developed for deriving the basic properties, concerning limiting normality of estimators, chi-squared type results for profiled versions for focus parameters, and quadratic forms for classes of test criteria, via the mechanics of minimisers of random functions.

(iv) Importantly, the i.i.d. setup is then lifted to classes of regression models. Intriguingly, with attention to certain crucial details, these extensions turn out to be not too strenuous, partly with Lindeberg theorem arguments replacing CLT details for the simpler case. The general likelihood theory makes it relatively straightforward to define and work with not only multiple linear regression, but logistic regression, Poisson regression, gamma type regressions, and more.

Carrying out data analyses using tools from this chapter is often even surprisingly straightforward, even in new situations with new models. This is partly due the general well-working theory, but also to the modern conveniences of software packages for numerical optimisation, easy calculation of derivatives and second derivatives, and so on, after having programmed the basic empirical distance function to be used. For maximum likelihood methodology, in particular, this is showcased in Story #6, with a logistic full-data model for how covariates influences birthweights, in Story #28, with regression models for the number of bird species on islands outside Ecuador, and in Story #98, concerning models for time-to-failure of machine components. These applications are also meant to inspire the invention of new parametric models in new situations.

perhaps along
with context
relevant
distance
functions from
Ch. 6

Kullback–Leibler, likelihood functions

Ex. 5.1 *Maximising the log-likelihood.* Here we work through some examples of setting up the log-likelihood function and finding the ML estimator.

(a) Suppose $Y \sim \text{binom}(100, \theta)$ and that you observe $y_{\text{obs}} = 22$. Set up the log-likelihood function $\ell(\theta)$ and plot it. Show that its maximiser is $\hat{\theta} = 0.22$.

(b) Simulate ten values of Y from the binomial $(100, 0.25)$, and plot the resulting ten log-likelihood functions. Note how they vary, giving different ML estimates of the underlying true $\theta_0 = 0.25$. With your code, experiment a bit with different sizes of the binomial n . For bigger n , the peak is sharper; show this mathematically.

(c) With i.i.d. observations $Y_1, \dots, Y_n$ from some parametric $f(y, \theta)$, show that the log-likelihood function becomes $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$. With such data points from the two-parameter normal $N(\xi, \sigma^2)$ model, show that the log-likelihood function may be written

$$\begin{aligned} \ell_n(\xi, \sigma) &= \sum_{i=1}^n \left\{ -\log \sigma - \frac{1}{2}(y_i - \xi)^2 / \sigma^2 - \frac{1}{2} \log(2\pi) \right\} \\ &= -n \log \sigma - \frac{1}{2} \{ Q_0 + n(\xi - \bar{Y})^2 \} / \sigma^2 - \frac{1}{2} n \log(2\pi), \end{aligned} \quad (5.2)$$

with $\bar{Y}$ the sample mean and $Q_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Show that the ML estimators are $\hat{\xi} = \bar{Y}$ and $\hat{\sigma}^2 = Q_0/n$. Verify also that the log-likelihood maximum value is $\ell_{n,\max} = -n \log \hat{\sigma} - \frac{1}{2} n \{1 + \log(2\pi)\}$. Compared to the classical empirical variance $Q_0/(n-1)$, see Ex. 1.49, we learn that the ML estimator has a small negative bias, but that the difference is small for moderate to large n .

(d) With data $y_1, \dots, y_{100}$ from the exponential density $\theta \exp(-\theta y)$, set up the log-likelihood function and find a formula for its maximiser $\hat{\theta}$.

(e) Simulate 50 points from the uniform $[0, \theta]$, with true value $\theta_0 = 1$. Find and plot the log-likelihood function, and read off the ML estimate $\hat{\theta}$.

(f) Simulate 100 values y_i from the $\text{Gam}(a_0, 1)$ distribution (see Ex. 1.9), with e.g. $a_0 = 3.33$. Set up and plot the log-likelihood function. Set up a little experiment where you keep track of the ML estimators $\hat{a}$ in repeated experiments from the samme $\text{Gam}(a_0, 1)$.

Ex. 5.2 *The likelihood and log-likelihood functions.* Consider the one-parameter model with density $f(y, \theta) = \exp(-\theta y^{1/2})\theta/(2y^{1/2})$ for $y > 0$, and assume $n = 12$ data points have been observed:

0.233 0.334 0.067 0.148 0.007 0.639 0.017 0.298 0.030 0.120 0.061 0.063

(a) Show that $f(y, \theta)$ indeed is a density, write down the log-likelihood function $\ell_n(\theta)$, and show that it is maximised at $\hat{\theta} = 1/W_n$, with $W_n = n^{-1} \sum_{i=1}^n y_i^{1/2}$. Find also a formula for the Hessian at the maximum position, $\hat{J}_{\text{obs}} = -\ell_n''(\hat{\theta})$.

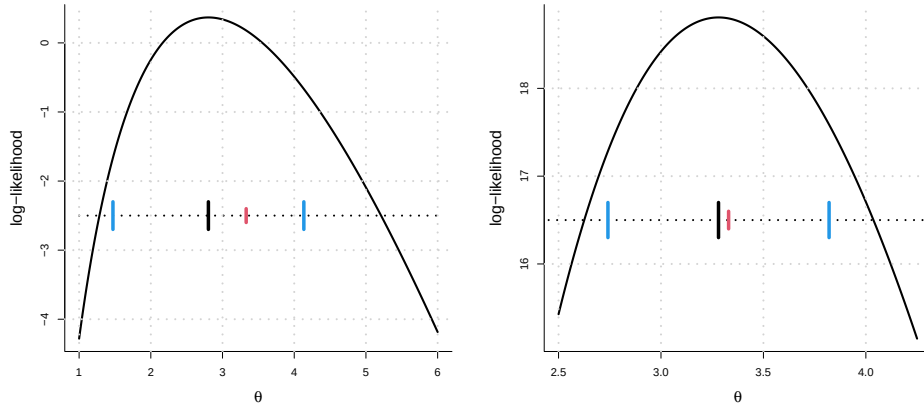


Figure 5.1: Left panel: the log-likelihood function $\ell_n(\theta)$ for the simple $n = 12$ dataset of Ex. 5.2. The maximum is attained at $\hat{\theta}_{12} = 2.803$. The blue bars indicate the 90 percent confidence interval coming from standard ML estimation theory, see Ex. 5.10, and is $[1.472, 4.134]$. The true value behind the data, $\theta_0 = 3.33$, is indicated as the red bar. Right panel: as for the left, but now with the bigger data set of $n = 100$ datapoints, from the same distribution, where the first 12 are as above. The ML estimate $\hat{\theta}_{100} = 3.281$ is closer to the truth. The minus second derivative $\hat{J}$ at the ML position has increased from 1.527 to 9.289, causing the ML based confidence interval to become considerably tighter, and is $[2.741, 3.821]$. The $\ell_n(\theta)$ is much sharper with $n = 100$; note the narrower range of θ to the right.

(b) Find the limit distribution of $\sqrt{n}(W_n - 1/\theta)$, using the CLT of Ch. 2, and use the delta method to find that the limit distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is $N(0, \theta^2)$. You may verify already that this is what comes out of the general ML theory developed below, see e.g. Ex. 5.10.

(c) Anticipating general ML theory to come, check again with Ex. 5.10, it will be seen that $\hat{\theta} \approx_d N(\theta, 1/\hat{J})$, under model conditions. Explain that this leads to approximate confidence intervals of the type $\hat{\theta} \pm z_0/\hat{J}^{1/2} = \hat{\theta}(1 \pm z_0/\sqrt{n})$, with z_0 the appropriate normal quantile. We've actually simulated these few data points above from the model, with true parameter $\theta_0 = 3.33$. Construct a version of Figure 5.1, left panel.

(d) In general the ML estimator might have a complicated distribution (though it is approximately normal, as we have seen here). In this particular model its precise distribution may be worked out, however; show that $\hat{\theta} \sim \theta (2n)/\chi_{2n}^2$. Use this to find a precise 90 confidence interval for θ , and compare to the approximation given above.

(e) To simulate data from this model, show that Y_i is equal in distribution to $(V_i/\theta_0)^2$, with the V_i i.i.d. from the unit exponential. Make a computer program to simulate n points from the $f(y, \theta_0)$ model, and which then finds the log-likelihood function, the ML estimator, and the approximate 90 percent confidence interval, as above. Run such a program for say 88 more points, forming a bigger dataset with $n = 100$ datapoints, and comment on what you find. Produce a version of Figure 5.1, right panel. Comment on the main features here, including that $\hat{J}$ becomes bigger with more data, yielding sharper confidence intervals. We would usually plot the log-likelihood and related aspects, as the

confidence curves of Ch. 8 for a shorter range of parameter values than in this right panel, but we here choose to plot using the same range for both $n = 12$ and $n = 100$.

Ex. 5.3 *The Kullback–Leibler divergence and the maximum likelihood method.* With i.i.d. data $Y_1, \dots, Y_n$ from a density g , to be approximated with a parametric f_θ , the particularly important *maximum likelihood estimation method* is worked with here, with $\hat{\theta}$ the maximiser of $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$. It will be seen as the natural cousin to the Kullback–Leibler divergence as a measure of distance from the true density to the parametric approximation. Development in exercises below give a precise description of how the method actually behaves.

(a) For two densities g and f , defined on a common support, the Kullback–Leibler distance, interpreted to be ‘from the first density to the second’, is

$$\text{KL}(g, f) = \int \log \left(\frac{g}{f} \right) g \, dy. \quad (5.3)$$

It is an important concept and tool for communication and information theory, as for probability theory and statistics. The $\log(g/f)$ term will be both positive and negative, in different parts of the domain. Show nevertheless that indeed $\text{KL}(g, f) \geq 0$, perhaps via the Jensen inequality, and that $\text{KL}(g, f) = 0$ only when the two densities are equal a.e. In Ex. 5.4 we learn more details about the KL distance, and look into illustrations, but here the main point is to see its close connection to ML estimation.

(b) To get to know the KL distance, work with normal distributions. (i) For $g = N(a, \sigma^2)$ and $f = N(b, \sigma^2)$, show that $\text{KL}(g, f) = \frac{1}{2}(b - a)^2/\sigma^2$, symmetric and a function of $|b - a|/\sigma$. (ii) For $g = N(0, \sigma_1^2)$ and $f = N(0, \sigma_2^2)$, show that $\text{KL}(g, f) = \frac{1}{2}h(\sigma_1^2/\sigma_2^2)$, with $h(u) = u - \log u - 1$ a nonnegative function, but not symmetric around 1. So KL is not in general symmetric, and it is fruitful to view it as the distance *from g to f* .

(c) Explain that the normalised log-likelihood $n^{-1}\ell_n(\theta)$ for each θ tends a.s. to the function $E_G \log f(Y, \theta) = \int g \log f_\theta \, dy$. When regularity conditions are in place to imply convergence of maximisers (we tend to such in exercises below), then, explain that $\hat{\theta} \rightarrow_{\text{pr}} \theta_0 = \theta_0(G)$, the minimiser of $\text{KL}(g, f_\theta)$. Explain that if the model is correct, then θ_0 is simply the true parameter; when the model is not perfect, we call the θ_0 the *least false parameter value*.

(d) As a simple illustration (to be returned to in Ex. 6.4), generate say $n = 100$ points from the uniform distribution, and use the parametric density $f(y, \theta) = \theta y^{\theta-1}$ on the unit interval. Write down and plot the log-likelihood function, and find the ML estimate. Also show that the implied best parameter value, if the data stem from some g , rather than from the model, is $\theta_0 = 1/E_g \log(1/Y)$.

Ex. 5.4 *The Kullback–Leibler distance: details and illustrations.* In Ex. 5.3 we have seen that the machinery of maximum likelihood is intimately related to the KL distance $\text{KL}(g, f) = \int g \log(g/f) \, dy$. Here we work on illustrations to learn more.

(a) For Poissons with parameters a and b , show that $\text{KL}(f_a, f_b) = a \log(a/b) + b - a$. For two gamma densities, with parameters $(c, 1)$ and $(d, 1)$, find $\text{KL}(g_c, g_d) = \log \Gamma(d) - \log \Gamma(c) - (d - c)\psi(c)$, with ψ the log-gamma function derivative.

(b) The KL distance is also perfectly well-defined and meaningful in higher dimension. Show that the KL distance from $N_p(\xi_1, \Sigma)$ to $N_p(\xi_2, \Sigma)$ can be expressed as $\frac{1}{2}\delta^2$, where $\delta = \{(\xi_2 - \xi_1)^t \Sigma^{-1}(\xi_2 - \xi_1)\}^{1/2}$ is the so-called Mahalanobis distance between the two populations.

(c) Start from $d(g, f) = -\int g \log\{1 + (f/g - 1)\} dy$, for densities which are not far from each other, and use Taylor expansion to find

$$\text{KL}(g, f) \approx \frac{1}{2} \int g(f/g - 1)^2 dy = \frac{1}{2} \left(\int f^2/g dy - 1 \right).$$

(d) As noted the KL distance is not symmetric, so ‘distance’ has a direction. In various statistical setups it makes sense to interpret $d(g, f)$ as the distance from ‘home density g ’ to ‘approximation candidate f ’. As also becoming clear from examples above, it’s somehow quadratic in nature, so when numbers are involved, measuring the KL distances, it would typically make more sense to give their square roots, as with $\{d(g, f_\theta)\}^{1/2}$, the degree of closeness of the parametric approximant f_θ to the ground truth g . Give in this fashion distances from f_a to f_b , then from f_b to f_a , for two Poissons.

Ex. 5.5 *KL approximation.* For the following cases the point is to set up a data generating density g , and then check how well a certain parametric family $f(y, \theta)$ does the approximation job. For each case, this tells us how well the ML can do its job, with enough data. For the various cases, find the minimiser, i.e. the best approximation; find the minimum square-root distance $d(g, f(\cdot, \theta_0))^{1/2}$ (since this gives a better picture than on the KL scale itself); and plot the true g alongside the parametric approximant.

(a) Let $g = 0.33 N(-1, 1) + 0.67 N(1, 1)$. Find the best normal approximation.

(b) Let g be a Gamma with parameters (2.22, 3.33). Find the best Weibull approximant, and also the best log-normal approximant. Similarly, start with a Weibull distribution, with parameters say (3.33, 2.22), and find the best Gamma distribution approximation.

(c) Let $g = 0.95 \text{Expo}(1) + 0.05 \text{Expo}(0.01)$, which roughly means that about five percent of the data come from a distribution which much higher mean than the mainstream exponential data. Find the best exponential model approximation, and also the best Gamma and Weibull approximations. Display the true g and these three best parametric approximations in the same diagram.

(d) Suppose data really come from $N(0.333, \sigma_1^2)$, with $\sigma_1 = 1.111$, where a statistician fits the simpler $N(0, \sigma^2)$ model. First, find out what happens to the ML estimator. Secondly, illustrate ‘what goes on’ by drawing e.g. ten samples of size $n = 50$ from the true density, and then display the ten versions of $n^{-1} \ell_n(\sigma)$, along with its limit $C(\sigma)$. Comment on your findings.

Score function, Fisher information matrix, Cramér–Rao lower bounds

Ex. 5.6 *Score functions and the Fisher information matrix.* Consider a parametric model with density $f(y, \theta)$ with respect to some measure μ , where $\theta = (\theta_1, \dots, \theta_p)^t$, the parameter of the model, is contained in some open parameter space $\Theta \subset \mathbb{R}^p$. Introduce score function

$$u(y, \theta) = \partial \log f(y, \theta) / \partial \theta \quad \text{and} \quad i(y, \theta) = \partial^2 \log f(y, \theta) / \partial \theta \partial \theta^t, \quad (5.4)$$

called the *score function*, and is a p -dimensional column vector, and the *information function*, a $p \times p$ matrix. For the theory developed in this exercise, these partial derivatives are assumed to exist and, they must be continuous in θ . We also assume that the *support* for the distribution, the smallest closed set for which the density is positive, does not depend on θ . Cases falling outside such assumptions are, for example, the uniform distribution on an unknown interval $[0, \theta]$.

(a) Show that if $\theta \mapsto (\partial/\partial\theta)f(y, \theta)$ is continuous for each y , and each element of $(\partial/\partial\theta\partial\theta^t)f(y, \theta)$ is bounded by an integrable function not depending on θ , then the first partial derivative can be passed under the integral sign in $\int f(y, \theta) d\mu(y)$. Deduce conditions under which we may pass also the second partial derivative under the same integral sign.

first Bartlett
identity

(b) Assuming conditions alluded in (a) hold, with derivatives of first and second order allowed to take inside the integral. Show that the score function has mean zero,

$$E_\theta u(Y, \theta) = \int f(y, \theta) u(y, \theta) d\mu(y) = 0.$$

second Bartlett
identity

Let next $K(\theta) = \text{Var}_\theta u(Y, \theta)$ and $J(\theta) = -E_\theta i(Y, \theta)$. Show that indeed $K(\theta) = J(\theta)$. This matrix is *the Fisher information matrix* for the model. It provides a measure of how much information about a parameter a data set provides. Note that the calculation of both $J(\theta)$ and $K(\theta)$ is taking place under the assumption that the model is actually correct.

(c) Consider a parametric density $f(y, \theta)$, with score function $u(y, \theta) = \partial \log f(y, \theta)$ and information matrix $J(\theta) = \text{Var}_\theta u(Y, \theta)$; see Ex. 5.6 for more on these. Show here, with appropriate conditions on third derivatives, that

$$\text{KL}(f(\cdot, \theta), f(\cdot, \theta + \varepsilon)) = \frac{1}{2} \varepsilon^t J(\theta) \varepsilon + O(\|\varepsilon\|^3).$$

(d) For the exponential model, with density $\theta \exp(-\theta y)$, find the score function, and compute the Fisher information function in two ways. The second (derivative) way of computing the Fisher information here is quite simple.

(e) Consider the general exponential family, with its natural parametrisation $f(y, \theta) = \exp\{\theta^t T(y) - k(\theta)\} h(y)$, see Ex. 1.57. Explain that the score function becomes $T(y) - k'(\theta)$, with Fisher matrix $J(\theta) = k''(\theta)$, that of the second order derivatives of $k(\theta)$.

(f) For the normal $N(\xi, \sigma^2)$ model, show that the score function can be expressed as

$$u(y, \xi, \sigma) = \left(\frac{\frac{1}{\sigma}(y - \xi)/\sigma}{\frac{1}{\sigma}\{(y - \xi)^2/\sigma^2 - 1\}} \right) = \frac{1}{\sigma} \begin{pmatrix} z \\ z^2 - 1 \end{pmatrix},$$

writing $z = (y - \xi)/\sigma$, which is a standard normal when y comes from the model. Demonstrate that the Fisher information matrix becomes

$$J(\xi, \sigma) = \text{Var}_{\xi, \sigma} u(Y, \xi, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

(g) Check with a few more of your favourite parametric models, where you should find the score function and the information function, and then formulae for both $J(\theta)$ and the variance matrix $K(\theta)$ of the score function, verifying that they are the same. Do this for the Poisson, for the geometric, for the $\text{Gam}(a, b)$, and for the binomial with parametrisation $p = \exp(\theta)/(1 + \exp(\theta))$.

(h) If Y has the uniform distribution on $[0, \theta]$, which of the regularity conditions listed above fail?

Ex. 5.7 Cramér–Rao lower bounds for estimators. A certain basic and classic inequality provides a lower bound for the variance of any unbiased estimator of a given parameter. There are various versions and generalisations, some of which we go through here. Inequalities of the type encountered here are sometimes called ‘information inequalities’, as

they may be used to define and analyse how much information there can be in a finite set of data. We also discover clear links to log-likelihoods and ML estimation; the behaviour of ML estimators for growing sample size, described in Ex. 5.10, exactly matches the Cramér–Rao lower bound for variances.

(a) To begin simply, suppose Y is an observation from the density $f(y, \theta)$, assumed smooth in its one-dimensional parameter. Let $u(y, \theta) = \partial \log f(y, \theta) / \partial \theta$ be the score function, with finite variance, by definition equal to the Fisher information, $J(\theta) = \int f(y, \theta) u(y, \theta)^2 dy$. Let $T = T(Y)$ be any estimator unbiased for θ , and assume that $f(y, \theta)$ satisfies the conditions of Ex. B.15(a). From $E_\theta T = \int T(y) f(y, \theta) dy = \theta$, use the conditions on $f(y, \theta)$ to deduce that $(d/d\theta) \int T(y) f(y, \theta) u(y, \theta) dy = 1$. Show that $\text{cov}_\theta(T, u(Y, \theta)) = 1$, and consequently $1 \leq \text{Var}_\theta T \text{Var}_\theta u(Y, \theta)$, from which we get one-dimensional classic Cramér–Rao inequality

$$\text{Var}_\theta T \geq 1/J(\theta).$$

(b) It is easy to generalise the above to the more interesting case of having more observation than one. Suppose $Y_1, \dots, Y_n$ are i.i.d. from the parametric model $f(y, \theta)$. Show that the arguments above still hold, essentially since $Y = (Y_1, \dots, Y_n)$ can be considered a single datum from the model with joint density $f(y_1, \theta) \cdots f(y_n, \theta)$. Show that the score function now becomes $u(y_1, \dots, y_n, \theta) = \sum_{i=1}^n u_0(y_i, \theta)$, writing for emphasis $u_0(y, \theta) = \partial \log f(y, \theta) / \partial \theta$ for the score function for a single observation. Deduce that the combined Fisher information for the full sample is $J_n = \text{Var}_\theta u(Y_1, \dots, Y_n) = nJ_0$, with $J_0 = \text{Var}_\theta u_0(Y_i, \theta)$ the information in a single observation.

(c) Show from this that if $T = T(Y_1, \dots, Y_n)$ is an unbiased estimator for θ , then

Cramér–Rao
lower bound

$$\text{Var}_\theta T \geq \frac{1}{nJ_0(\theta)} = \frac{1/J_0(\theta)}{n}.$$

This says that there is a clear limit to how well one might estimate a parameter in a model, with n observations. If you're not entirely satisfied with $\text{Var}_\theta T = 0.10$, say, and wish for variance 0.05 instead, then shell out more money to get twice as many observations.

(d) Show more generally that if $T = T(Y_1, \dots, Y_n)$ is an estimator for θ , with mean $E_\theta T = \theta + b(\theta)$, i.e. with a certain bias $b(\theta)$, then

$$\text{Var}_\theta T \geq \frac{1}{n} \frac{\{1 + b'(\theta)\}^2}{J_0(\theta)}.$$

In particular, show that there's a lower bound on the mean squared error for *any* estimator (i.e. not merely the unbiased ones):

$$\text{mse}(\theta) = E \{T(Y_1, \dots, Y_n) - \theta\}^2 \geq n^{-1} \{1 + b'(\theta)\}^2 / J_0(\theta) + b(\theta)^2.$$

(e) Go through the following examples, in each case finding the score function, the information $J_0(\theta)$, and the lower bound for any unbiased estimator of the model parameter. (i) y is binomial (n, θ) . (ii) y is Poisson θ . (iii) y is normal (ξ, σ^2) , with σ known. (iv) y is normal (ξ, σ^2) , with ξ known, and σ to be estimated. Comment on the implications of your findings.

Ex. 5.8 *Cramér–Rao bounds for the multidimensional case.* In generalisation of the above situation to the case of multiparameter models, assume first that y is a single

observation from the model $f(y, \theta)$, with $\theta = (\theta_1, \dots, \theta_p)$ of dimension p . Let $u_0(y, \theta) = \partial \log f(y, \theta) / \partial \theta$ be the score function, for such a single y , with the $p \times p$ Fisher information matrix $J_0(\theta) = \text{Var}_\theta u_0(Y, \theta)$ assumed positive definite.

(a) For a symmetric $p \times p$ matrix A we write $A \geq 0$ provided it is nonnegative definite, i.e. that $c^t A c \geq 0$ for all c . Show that the covariance matrix of a random vector is necessarily nonnegative, and that $A \geq 0$ is equivalent to its eigenvalues being nonnegative. Explain that $a_{ii} \geq 0$, for the diagonal elements, but that we may still have $a_{i,j} < 0$ for some off-diagonal elements. With two symmetric matrices, we write $A \geq B$ if $A - B \geq 0$, which is the ordering of variance matrices we use below.

(b) Assume that $T = T(Y)$ is an unbiased estimator of θ , which also means that $E_\theta T_j(Y) = \theta_j$ for each component j . With I_p the identity matrix of size $p \times p$, deduce from $E_\theta T = \int T(y) f(y, \theta) dy$ that

$$(\partial / \partial \theta) E_\theta T = \int T(y) f(y, \theta) u_0(y, \theta)^t dy = I_p.$$

Then work out that

$$\text{Var}_\theta \{T - J_0(\theta)^{-1} u_0(Y, \theta)\} = E_\theta \{T - \theta - J_0(\theta)^{-1} u_0(Y, \theta)\} \{T - \theta - J_0(\theta)^{-1} u_0(Y, \theta)\}^t$$

can be expressed as $\text{Var}_\theta T - J_0(\theta)^{-1}$. We have then demonstrated a multidimensional version of the Cramér–Rao inequality, that $\text{Var}_\theta T \geq J_0(\theta)^{-1}$.

(c) Generalise the above to the case of n i.i.d. observations $Y_1, \dots, Y_n$ from the model. Show that the information matrix for the full data set becomes

$$J_n(\theta) = \text{Var}_\theta \frac{\partial \log \{f(Y_1, \theta) \cdots f(Y_n, \theta)\}}{\partial \theta} = n J_0(\theta),$$

and that for *any* unbiased estimator $T = T(Y_1, \dots, Y_n)$ of θ , we must have

$$\text{Var}_\theta T \geq \{n J_0(\theta)\}^{-1} = n^{-1} J_0(\theta)^{-1}.$$

More generally, if independent observations $Y_1, \dots, Y_n$ come from densities $f_1(y, \theta), \dots, f_n(y, \theta)$, with Fisher information matrices $J_1(\theta), \dots, J_n(\theta)$, show that any unbiased estimator T of θ has $\text{Var}_\theta T \geq \{J_1(\theta) + \cdots + J_n(\theta)\}^{-1}$.

(d) Suppose now that $\mu = \mu(\theta_1, \dots, \theta_p)$ is a one-dimensional parameter in focus, and that $T = T(Y)$ is an unbiased estimator. With $c(\theta) = \partial \mu(\theta) / \partial \theta$ the $p \times 1$ gradient, show that

$$\text{Var}_\theta \{T - c(\theta)^t J(\theta)^{-1} u_0(Y)\} = \text{Var}_\theta T - c(\theta)^t J(\theta)^{-1} c(\theta),$$

and conclude the lower bound $\text{Var}_\theta T \geq \lambda(\theta) = c(\theta)^t J(\theta)^{-1} c(\theta)$. Generalise to the case of $Y_1, \dots, Y_n$ being i.i.d. from $f(y, \theta)$: if $T_n = T(Y_1, \dots, Y_n)$ is unbiased for μ , show that $\text{Var}_\theta T \geq \lambda(\theta)/n$. We learn also a bit more from these arguments. Show that

$$\text{Var}_\theta \left\{ T_n - n^{-1} \sum_{i=1}^n c(\theta)^t J(\theta)^{-1} u_0(Y_i, \theta) \right\} = \text{Var}_\theta T_n - \lambda(\theta)/n,$$

so T_n having a variance coming close to the lower bound means $T_n - \mu$ being equal to or close to the random variable $n^{-1} \sum_{i=1}^n c(\theta)^t J(\theta)^{-1} u_0(Y_i, \theta)$.

(e) Also other estimators for θ deserve to be studied, even when they are not exactly unbiased. We start with a single observation Y from $f(y, \theta)$, with score function $u_0(y, \theta)$ as above, and then generalise to n observations afterwards. Assume therefore that $T = T(Y)$ is such that

$$E_\theta T = \int T(y) f(y, \theta) dy = \theta + b(\theta),$$

with components $\theta_j + b_j(\theta)$, for suitable bias functions $b_1(\theta), \dots, b_p(\theta)$, perhaps not far from zero. Show that

$$(\partial/\partial\theta) E_\theta T = \begin{pmatrix} 1 + \partial b_1(\theta)/\partial\theta \\ \vdots \\ 1 + \partial b_p(\theta)/\partial\theta \end{pmatrix} = \int T(y) f(y, \theta) u_0(y, \theta)^t dy.$$

Then work with $\text{Var}_\theta [T - \{I_p + b'(\theta)\} J_0(\theta)^{-1} u_0(Y, \theta)]$ to demonstrate that

$$\text{Var}_\theta T \geq \{I_p + b'(\theta)\} J_0(\theta)^{-1} \{I + b'(\theta)\}^t.$$

(f) Generalise to the case of n i.i.d. observations, to reach

$$\text{Var}_\theta T \geq n^{-1} \{I_p + b'(\theta)\} J_0(\theta)^{-1} \{I + b'(\theta)\}^t.$$

(g) Suppose that the Y_i are independent, but with different densities, $f_i(y, \theta)$ with the same θ . With $J_i(\theta)$ the Fisher matrix for Y_i alone, show that

$$\text{Var}_\theta T \geq \{J_1(\theta) + \dots + J_n(\theta)\}^{-1},$$

for any unbiased estimator T .

Ex. 5.9 Fisher information and parameter transformations. Consider a parametric model with density $f(y, \theta)$, with θ of length p , say. We have seen the central role the Fisher information matrix $J(\theta)$ plays, with more to come for the likelihood machinery below. Here we look into reparametrisation, going from $\theta = (\theta_1, \dots, \theta_p)$ to say $\phi = (\phi_1, \dots, \phi_p)$, with smooth 1-1 transformations $\theta = h(\phi)$, or $\phi = h^{-1}(\theta)$. It is assumed that the $p \times p$ matrix $G(\phi) = \partial h(\phi)/\partial\phi$, with rows $\partial h_j(\phi)/\partial\phi$, exists and is non-zero throughout the parameter space.

(a) Let $u(y, \theta) = \partial \log f(y, \theta)/\partial\theta$ be the score function with the θ parametrisation. Show that for the ϕ parametrisation, we have

$$u^*(y, \phi) = \frac{\partial \log f(y, h(\phi))}{\partial\phi} = \begin{pmatrix} u_1(y, h(\phi))^t g_1(\phi) \\ \vdots \\ u_p(y, h(\phi))^t g_p(\phi) \end{pmatrix} = G(\phi) u(y, h(\phi)).$$

(b) Show from this that the Fisher information matrix for the ϕ parametrisation becomes $J^*(\phi) = G(\phi) J(h(\theta)) G(\phi)^t$. Deduce also that the important inverse can be written $J^*(\phi)^{-1} = (G(\phi)^t)^{-1} J(h(\theta))^{-1} G(\phi)^{-1}$.

(c) As illustrations, consider the binomial (n, θ) , with reparametrisations (i) $\phi_1 = \theta^{1/2}$; (ii) $\phi_2 = \log(\theta/(1 - \theta))$, so that $\theta = \exp(\phi)/\{1 + \exp(\theta)\}$; (iii) $\phi = 2 \arcsin(\theta^{1/2})$. For each of these, find expressions for the Fisher information $J^*(\phi)$.

(d) Consider the normal (ξ, σ^2) . Find first the Fisher information matrix for the (ξ, σ) parametrisation, then for the (ξ, λ) parametrisation, where $\lambda = \sigma^2$.

Maximum likelihood, profiled likelihood, deviance theorems

Ex. 5.10 *Maximum likelihood estimators: a limit process and main heuristics.* After having motivated and worked through particular instances of the maximum likelihood function estimators, we now return to the general case, aiming to demonstrate both limiting normality and other associated results, finding recipes for large-sample approximations in the process. The aim now is to go through the basic ideas, involving also two main heuristics. Similar arguments will work also for rather more general estimation methods in Ch. 6. To make these heuristics precise we need more conditions and details, to which we return in further exercises below.

least false
parameter value

regularity
conditions

Setting the scene, we have $Y_1, \dots, Y_n$ i.i.d. from some G , and we fit a parametric model $f(y, \theta)$ smooth in its parameters. Let $\theta_0 = \theta_0(G)$ be the least false parameter value, the maximiser of $E_G \log f(Y, \theta)$, and $\hat{\theta}$ is its estimator, the maximiser of $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$. To present the basic ideas and main heuristics we shall start with these regularity conditions: (i) The true $\theta_0 = \theta_0(G)$ is an inner point in its parameter space inside $\mathbb{R}^p$. (ii) The $\log f(y, \theta)$ is smooth in θ , in a neighbourhood around θ_0 , with at least two derivatives, i.e. the score function $u(y, \theta)$ and information function $i(y, \theta)$. (iii) The matrix $J = -E_G i(Y, \theta_0)$ is finite and positive definite. (iv) The score function $u(Y, \theta_0)$ has finite variance matrix K .

(a) A key idea is to work with the following random function. Write

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = U_n^t s - \frac{1}{2} s^t J_n s + r_n(s),$$

with $U_n = (1/\sqrt{n}) \sum_{i=1}^n u(Y_i, \theta_0)$, $J_n = -n^{-1} \sum_{i=1}^n i(Y_i, \theta_0)$. Show that $A_n(s)$ is properly defined for all large enough n , that U_n has mean zero and tends to $U \sim N_p(0, K)$, and that $J_n \rightarrow_{\text{pr}} J$.

(b) It is also clear from the Taylor expansion argument that with moderate further regularity, the remainder term $r_n(s)$ will go to zero in probability, for each s . There is then convergence

$$A_n(s) \rightarrow_d A(s) = U^t s - \frac{1}{2} s^t J s, \quad (5.5)$$

for each s . Heuristic One is then to go from $A_n \rightarrow_d A$ to $\arg\max(A_n) \rightarrow_d \arg\max(A)$. Explain that this then leads to

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1}U, \quad \text{which is a } N_p(0, J^{-1}KJ^{-1}).$$

the sandwich
matrix

We call this limit distribution variance matrix $J^{-1}KJ^{-1}$ the sandwich matrix.

(c) Additionally, Heuristic Two is to go from $A_n \rightarrow_d A$ to $\max(A_n) \rightarrow_d \max(A)$. Argue that this entails

$$W_n = 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \rightarrow_d W = U^t J^{-1}U.$$

For this quadratic form, find formulae $\text{Tr}(J^{-1}K)$ and $2\text{Tr}(J^{-1}KJ^{-1}K)$ for the mean and variance.

(d) For a simple illustration, consider data from a density g on the positive half-line, with true mean and variance ξ and σ^2 . Fitting the data to the exponential $f(y, \theta) = \theta \exp(-\theta y)$, show first that the ML estimator aims at $\theta_0 = 1/\xi$. Then show $J = 1/\theta_0^2$ and $K = \sigma^2$, with sandwich $\Sigma = J^{-1}KJ^{-1} = \sigma^2/\xi^4$. Explain that if the model is correct, then $\sigma^2 = 1/\theta_0^2$, with $\Sigma = 1/\theta_0^2$.

(e) Back to the general narrative, assume the model holds, with some true parameter θ_0 , and Fisher information matrix $J = J(\theta_0)$, as per Ex. 5.6. Derive the celebrated and exceedingly useful results, for maximum likelihood, that

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N_p(0, J^{-1}) \quad \text{and} \quad W_n = 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \rightarrow_d \chi_p^2. \quad (5.6)$$

Explain that the limit distribution variance matrix precisely matches the best one could have hoped for, from the Cramér–Rao information inequalities of Ex. 5.7–5.8.

Ex. 5.11 *Maximum likelihood: from heuristics to proofs.* The general setup is as with the previous Ex. 5.10, with ML estimator $\hat{\theta}$ maximising $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$, with observations coming from G . We shall use the process $A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0)$ from the previous exercise. In addition to regularity conditions (i)–(iv) given there, we postulate (v), that also third derivatives of $\log f(y, \theta)$ exist, and that these have finite means in a neighbourhood around θ_0 . Other sets of sufficient regularity conditions, in various cases simpler and cleaner, are worked with in Ex. 5.32–5.33 below.

(a) Explain first that the maximiser of A_n is $\alpha_n = \sqrt{n}(\hat{\theta} - \theta_0)$, where we shall also study the overall maximum $A_{n,\max} = A_n(\alpha_n)$ below. Let $B_n(s) = U_n^t s - \frac{1}{2} s^t J_n s$ be the quadratic approximation to A_n , with maximiser β_n and overall maximum $B_{n,\max} = \max\{B_n(s) : \text{all } s\}$. Show that

$$\begin{aligned} \beta_n &= J_n^{-1} U_n \rightarrow_d J^{-1} U \sim N_p(0, J^{-1} K J^{-1}), \\ B_{n,\max} &= \frac{1}{2} U_n^t J_n^{-1} U_n \rightarrow_d \frac{1}{2} U^t J^{-1} U. \end{aligned}$$

So things are simple and clear for the quadratic approximation B_n ; we need to show that the same results obtain for the real thing, the A_n .

(b) With the smooth third order derivatives assumptions, show that

$$A_n(s) = B_n(s) + r_n(s) \quad \text{with} \quad r_n(s) = (1/6) \sum_{i,j,k} \frac{\partial^3 \log f(Y_i, \tilde{\theta}_s)}{\partial \theta_i \partial \theta_j \partial \theta_k} \frac{s_i s_j s_k}{n^{3/2}},$$

with $\tilde{\theta}_s$ somewhere between θ_0 and $\theta_0 + s/\sqrt{n}$. From this, explain that $|r_n(s)| \leq D_n \|s\|^3 / \sqrt{n}$, with $D_n = O_{\text{pr}}(1)$. So at least there is pointwise $r_n(s) \rightarrow_{\text{pr}} 0$ for the remainder.

(c) Suppose first there is an increasing $R_n \rightarrow \infty$ for which we can prove (i) both maximisers α_n and β_n lie inside $\|s\| \leq R_n$ with probability tending to one, and (ii) $\max_{\|s\| \leq R_n} |A_n(s) - B_n(s)| \rightarrow_{\text{pr}} 0$. Show that $A_{n,\max} - B_{n,\max} \rightarrow_{\text{pr}} 0$. Then show that conditions (i) and (ii) actually hold here, for $R_n = n^{1/8}$, and conclude, so far, that $2 A_{n,\max} \rightarrow_d U^t J^{-1} U$.

(d) Then consider B_n a certain distance away from the maximum. For given small ε , show that for v with $\|v\| \geq \varepsilon$, we have

$$B_n(\beta_n + v) = B_{n,\max} - \frac{1}{2} v^t J_n v \leq B_{n,\max} - \frac{1}{2} j_n \varepsilon^2,$$

where j_n is the smallest eigenvalue of J_n . Show then that the event Ω_n , where $A_n(\beta_n + v) \geq B_{n,\max} + \frac{1}{4} j_n \varepsilon^2$ for all v with $\|v\| \geq \varepsilon$, must have $\Pr(\Omega_n) \rightarrow 1$. Prove from these established statements that $\alpha_n - \beta_n$ must tend to zero in probability. Conclude that

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &\rightarrow_d J^{-1} U \sim N_p(0, J^{-1} K J^{-1}), \\ 2\{\ell_{n,\max} - \ell_n(\theta_0)\} &\rightarrow_d W = U^t J^{-1} U. \end{aligned} \quad (5.7)$$

These are the statements heuristically arrived at in Ex. 5.10, but now with full proofs.

(e) Via $\alpha_n - \beta_n \rightarrow_{\text{pr}} 0$, show that

$$\hat{\theta} - \theta_0 = J^{-1} n^{-1} \sum_{i=1}^n u(Y_i, \theta_0) + o_{\text{pr}}(n^{-1/2}),$$

the ML is just a zero-mean average away from its target

a powerful linearised representation for the ML.

Ex. 5.12 Profiling quadratic functions. Consider some quadratic function $A(s) = \frac{1}{2} s^t J s$, with J symmetric and positive definite of dimension $p \times p$. It is useful to sort out minima of A under different types of side constraints.

Lagrange
multipliers for
minimisation
with constraints

(a) We start out examining the minimum of $A(s)$ over all s with $c^t s = x$, for some given vector c and level x . The Lagrange multiplier way of solving such a problem is to minimise the function $\frac{1}{2} s^t J s - \lambda(c^t s - x)$, with no constraint on s , and in the process find the λ agreeing with the constraint. Taking derivatives, show that minimum occurs for $s_0 = \lambda J^{-1} c$, leading to $c^t s_0 = \lambda c^t J^{-1} c$, which should then equal x . Explain that this leads to minimiser $s_0 = x/(c^t J^{-1} c)$ and attained minimum $A_{\min} = \frac{1}{2} x^2 / (c^t J^{-1} c)$.

(b) Generalising the above, from a scalar to a vector, consider a $p_0 \times p$ matrix C and an x of dimension p_0 , where it is assumed that $C J^{-1} C^t$ has full rank. The task is to work out the minimiser and minimum of $\frac{1}{2} s^t J s$ over all s with $Cs = x$. Work out that the minimiser is

$$s_0 = J^{-1} C^t \lambda = J^{-1} C^t (C J^{-1} C^t)^{-1} x \quad \text{with} \quad A_{\min} = \frac{1}{2} x^t (C J^{-1} C^t)^{-1} x.$$

(c) For a useful extension, consider now $A(s) = s^t u - \frac{1}{2} s^t J s$, for given u and positive definite J , of dimension p . The task is to maximise this over all s satisfying the constraint $c^t s = t$, for a fixed t . Do first Lagrange, maximising

$$h(s) = s^t u - \frac{1}{2} s^t J s - \lambda(c^t s - t) \quad \text{over all } s.$$

Explain that this leads to $u - Js - \lambda c = 0$, hence $s_0 = J^{-1}(u - \lambda c)$ for the maximiser. Writing $\tau^2 = c^t J^{-1} c$ and $w = c^t J^{-1} u$, show that the constraint implies $w - \lambda \tau^2 = t$, giving $\lambda = (w - t)/\tau^2$. Furthermore, show that the maximum of $A(s)$, under $c^t s = t$, becomes

$$\begin{aligned} s_0^t u - \frac{1}{2} s_0^t J s_0 &= (u - \lambda c)^t J^{-1} u - \frac{1}{2} (u - \lambda c)^t J^{-1} (u - \lambda c) \\ &= \frac{1}{2} u^t J^{-1} u - \frac{1}{2} (w - t)^2 / \tau^2. \end{aligned}$$

Ex. 5.13 Profiling a log-likelihood function, I. For data $Y_1, \dots, Y_n$ from some distribution G , we are using the log-likelihood function $\ell_n(\theta)$ to find its maximiser $\hat{\theta}$, seen as the canonical estimator for the least false parameter θ_0 , the $\text{KL}(g, f_\theta)$ minimiser. For a focus parameter $\mu = \mu(\theta)$, a smooth function of $\theta = (\theta_1, \dots, \theta_p)$, it is very useful to work with the associated profile function

log-likelihood
profile

$$\ell_{n,\text{prof}}(\mu) = \max\{\ell_n(\theta) : \mu(\theta) = \mu\}.$$

(a) As an introductory illustration, simulate $n = 100$ datapoints from some $\text{Gam}(a_0, b_0)$, perhaps with $(a_0, b_0) = (2.22, 3.33)$. From the log-likelihood function $\ell_n(a, b)$, compute first the ML estimators, and then compute and display the profiled $\ell_{n,\text{prof}}(\sigma)$, for the focus parameter $\sigma = a^{1/2}/b$, the standard deviation.

(b) Show that the consequent $\hat{\mu} = \mu(\hat{\theta})$ becomes normal for growing n . With setup and notation as in Ex. 5.10, prove indeed that $\sqrt{n}(\hat{\mu} - \mu_0) \rightarrow_d c^t J^{-1} U$, with $c = \partial\mu(\theta_0)/\partial\theta$, and that the limit is a zero-mean normal with variance $c^t J^{-1} K J^{-1} c$. Show also that this $\hat{\mu}$ is identical to the maximiser of the profile function.

(c) In other words we already know the basic story for any focus parameter estimator $\hat{\mu}$, thanks to the delta method. It is however fruitful to work with representations and approximations stemming from examining the associated profile function. With methods from Ex. 5.11, show that

$$\ell_n(\hat{\theta} + s/\sqrt{n}) - \ell_n(\hat{\theta}) = -\frac{1}{2}s^t J_n s + r_n(s), \quad \text{with } r_n(s) = O_{\text{pr}}(\|s\|^3/\sqrt{n}).$$

For the profiling, therefore, we must maximise this expression over all s such that $\mu(\theta) = \mu(\hat{\theta} + s/\sqrt{n}) = \mu$. With $\mu(\theta) = \hat{\mu} + c_n^t s/\sqrt{n} + O_{\text{pr}}(\|s\|^2/n)$, here writing $c_n = \partial\mu(\hat{\theta})/\partial\theta$, the essence is to minimise $\frac{1}{2}s^t J_n s$ under $c_n^t s = \sqrt{n}(\mu - \hat{\mu}) = x_n$, say. Appealing to Ex. 5.12(a), show that this minimum becomes $\frac{1}{2}x_n^2/c_n^t J_n^{-1} c_n = n(\hat{\mu} - \mu)^2/c_n^t J_n^{-1} c_n$. Fill in more details to prove that with $\mu_0 = \mu(\theta_0)$ the true parameter in question,

the deviance function

$$D_n(\mu_0) = 2\{\ell_{n,\text{prof}}(\hat{\mu}) - \ell_{n,\text{prof}}(\mu_0)\} = \frac{n(\hat{\mu} - \mu_0)^2}{c_n^t J_n^{-1} c_n} + o_{\text{pr}}(1) \rightarrow_d \frac{(c^t J^{-1} U)^2}{c^t J^{-1} c} \sim \kappa \chi_1^2,$$

with $\kappa = c^t J^{-1} K J^{-1} c / c^t J^{-1} c$. The $D_n(\mu)$ is called *the deviance function* for the focus parameter studied.

(d) Verify, under the traditional circumstances that the model is adequate, that

Wilks theorem

$$D_n(\mu_0) \rightarrow_d \chi_1^2 \quad \text{at the correct parameter value } \mu_0 = \mu(\theta_0). \quad (5.8)$$

We call this the focus parameter version of the Wilks theorem; see Ex. 5.14 for a vector generalisation. Explain how this can be used to construct confidence intervals.

and full confidence curves, in Ch. 8

(e) Above we reached $D_n(\mu_0) \rightarrow_d W^2/\tau^2$ via Ex. 5.12(a), writing $W_n = c^t J^{-1} U_n \rightarrow_d W = c^t J^{-1} U$ and $\tau^2 = c^t J^{-1} c$. We learn a bit more by suitably putting also Ex. 5.12(c) to use. For the following suppose the focus parameter is a linear combination, $\mu = c^t \theta$, with null value $\mu_0 = c^t \theta_0$; results can be extended using a linearisation argument. From the basic process $A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0)$, studied in Ex. 5.10–5.11, there approximated with $U_n^t s - \frac{1}{2}s^t J_n s$, work with the profiled $\ell_{n,\text{prof}}(\mu_0 + t/\sqrt{n})$. Using the profiling result pointed to, show that

$$\begin{aligned} \ell_{n,\text{prof}}(\mu_0 + t/\sqrt{n}) &= \max\{A_n(s) : c^t(\theta_0 + s/\sqrt{n}) = \mu_0 + t/\sqrt{n}\} \\ &= \max\{U_n^t s - \frac{1}{2}s^t J_n s : c^t s = t\} + o_{\text{pr}}(1) \\ &= \frac{1}{2}U_n^t J_n^{-1} U_n - \frac{1}{2}(W_n - t)^2/\tau^2 + o_{\text{pr}}(1). \end{aligned}$$

Then use this to demonstrate that

$$\begin{aligned} B_n(t) &= \ell_{n,\text{prof}}(\mu_0 + t/\sqrt{n}) - \ell_{n,\text{prof}}(\mu_0) \\ &\rightarrow_d B(t) = -\frac{1}{2}(W - t)^2/\tau^2 + \frac{1}{2}W^2/\tau^2 = (Wt - \frac{1}{2}t^2)/\tau^2. \end{aligned}$$

Deduce that $\sqrt{n}(\hat{\mu} - \mu_0) \rightarrow_d W$, and, again, that $2\{\ell_{n,\text{max}} - \ell_{n,\text{prof}}(\mu_0)\} \rightarrow_d W^2/\tau^2$.

Ex. 5.14 *Profiling a log-likelihood function, II: Wilks theorems.* We have developed maximum likelihood methods for estimating the least false model parameter vector $\alpha_0 = \alpha_0(G)$, defined as the minimiser of $\text{KL}(g, f_\alpha)$. We now consider situations where some of

the components of the α parameter are specified. Such occur when one tests for lower-dimensional structure, etc. This invites setting up the following framework, with a wide model having $\alpha = (\theta, \gamma)$ of length $p + q$, and the narrow model considered has (θ, γ_0) , with θ unknown but $\gamma = \gamma_0$ fixed. Estimators $(\hat{\theta}, \hat{\gamma})$ in the wide model maximise $\ell_n(\theta, \gamma)$, whereas $\tilde{\theta}$ for the narrow model maximises $\ell_n(\theta, \gamma_0)$.

The theory developed in the previous exercises mentioned holds for the wide and the narrow models, separately, and below we postulate that the regularity conditions (i)–(v) put up in Ex. 5.10–5.11 are in force. Efforts of linear and matrix algebra are required in order to handle these models jointly, however. Define therefore

$$J_{\text{wide}} = -E_G \frac{\partial^2 \log f(Y, \theta_0, \gamma_0)}{\partial \alpha \partial \alpha^t} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{with inverse} \quad J_{\text{wide}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

where $J_{00} = J_{\text{narr}}$ is of size $p \times p$, etc. There is similarly a $(p + q) \times (p + q)$ matrix K_{wide} , with submatrices $K_{00}, K_{01}, K_{10}, K_{11}$, the variance matrix of $U = (U_0^t, U_1^t)^t$, the score function $\partial \log f(Y, \theta_0, \gamma_0) / \partial \alpha$.

(a) The following developments are under the $\gamma = \gamma_0$ constraint, so $\alpha_0 = (\theta_0, \gamma_0)$ is the true parameter, determined by the distribution G . We let J_{wide} and K_{wide} be determined accordingly. The $q \times q$ matrix J^{11} has an important role, here and on later occasions (as with the FIC in Ch. 12). Show that

$$Q = J^{11} = (J_{11} - J_{10} J_{00}^{-1} J_{01})^{-1}. \quad (5.9)$$

Similarly, we have $J^{00} = J_{00}^{-1} + J_{00}^{-1} J_{01} Q J_{10} J_{00}^{-1}$. Show also that $J^{00} - J_{00}^{-1} = J^{01} J_{10} J_{00}^{-1}$, $J^{10} = -Q J_{10} J_{00}^{-1}$.

(b) Argue, under the $\gamma = \gamma_0$ constraint, that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta} - \theta_0) \\ \sqrt{n}(\hat{\gamma} - \gamma_0) \end{pmatrix} \rightarrow_d J_{\text{wide}}^{-1} \begin{pmatrix} U_0 \\ U_1 \end{pmatrix}, \quad \sqrt{n}(\tilde{\theta} - \theta_0) \rightarrow_d J_{00}^{-1} U_0.$$

Show also, again using results reached earlier, that

$$\begin{aligned} \ell_{\text{wide}, \max} - \ell_n(\theta_0, \gamma_0) &\rightarrow_d \frac{1}{2} U^t J_{\text{wide}}^{-1} U, \\ \ell_{\text{narr}, \max} - \ell_n(\theta_0, \gamma_0) &\rightarrow_d \frac{1}{2} U_0^t J_{00}^{-1} U_0, \end{aligned}$$

with $\ell_{\text{wide}} = \ell_n(\hat{\theta}, \hat{\gamma})$ and $\ell_{\text{narr}} = \ell_n(\tilde{\theta}, \gamma_0)$. Deduce that

$$W_n = 2(\ell_{\text{wide}, \max} - \ell_{\text{narr}, \max}) \rightarrow_d W = U^t J_{\text{wide}}^{-1} U - U_0^t J_{00}^{-1} U_0. \quad (5.10)$$

Show that this limit variable has mean $\text{Tr}(J_{\text{wide}}^{-1} K_{\text{wide}}) - \text{Tr}(J_{00}^{-1} K_{00})$.

(c) Under model conditions, we have the Bartlett identity $J = K$, simplifying matters. Show first from the above that $E W = (p + q) - p = q$. Show further that

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \rightarrow_d Z = J^{10} U_0 + J^{11} U_1 \sim N_q(0, Q).$$

Use then linear and matrix algebra to show that W above is in fact equal to $W' = Z^t Q^{-1} Z$. Explain that we have reached the important result that under the narrow model, where $\gamma = \gamma_0$,

$$W_n = 2(\ell_{\text{wide}, \max} - \ell_{\text{narr}, \max}) \rightarrow_d \chi_q^2.$$

This is called *the Wilks theorem*, or, rather, one of several related Wilks theorems. Explain that the (5.8) reached in Ex. 5.13 is the special case of $q = 1$.

the Q matrix

the Wilks theorem for testing a submodel of a bigger model

ML theory and Wilks theorems: examples

Ex. 5.15 *ML estimators in some models.* Here we work through some examples, where the task is to set up the log-likelihood function and if feasible also explicit formulae for the ML estimators. Importantly, the general theory developed above applies nicely, often with easy implementation, also when there is no explicit formula for the ML estimator.

(a) With $Y_1, \dots, Y_n$ from the normal model $N(\xi, \sigma^2)$, write down the log-likelihood function. Find the ML estimator for σ when ξ is a known value, and show also the ML estimators in the case where both parameters are unknown become $\hat{\xi} = \bar{Y}$ and $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$.

(b) Suppose $Y_i \sim \text{Pois}(w_i \theta)$, with known exposure times w_i , and that the observations are independent, for $i = 1, \dots, n$. Write down the log-likelihood function, find the ML estimator, and its mean and variance.

(c) Let $Y_1, \dots, Y_n$ be i.i.d. from the uniform model on $[0, \theta]$, with θ the unknown end-point. Set up the likelihood function and find the ML estimator.

Ex. 5.16 *Maximum likelihood for the Beta and Gamma models.* Consider the Beta and Gamma two-parameter models, with densities

$$\text{be}(y, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1} \quad \text{and} \quad g(y, a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp(-by),$$

for $y \in (0, 1)$ and $y > 0$, respectively. The task is in each case to estimate the parameters based on an i.i.d. sample $Y_1, \dots, Y_n$.

(a) We start with the Beta distribution, see Ex. 1.19, where we in particular have found formulae for the mean and variance in terms of (a, b) . With empirical mean and variance $\bar{y}$ and $\hat{\sigma}^2$, show how (a, b) can be fitted by solving the two equations $\bar{y} = EY$ and $\hat{\sigma}^2 = \text{Var} Y$. With solutions $(\hat{a}_m, \hat{b}_m)$ for these moment estimators, and assuming the Beta model is correct, explain how you can find limit distributions of $\sqrt{n}(\hat{a}_m - a, \hat{b}_m - b)$.

(b) Write down the log-likelihood function, say $\ell_n(a, b)$. Show that the ML estimators $(\hat{a}, \hat{b})$ are the solutions to the two equations

$$n^{-1} \sum_{i=1}^n \log Y_i = \psi(a) - \psi(a+b), \quad n^{-1} \sum_{i=1}^n \log(1 - Y_i) = \psi(b) - \psi(a+b),$$

where $\psi(x) = \partial \log \Gamma(x) / \partial x$ is the so-called digamma function. There are no explicit formulae here, but the two equations may easily be solved numerically. Explain how the limit distribution for $\sqrt{n}(\hat{a} - a, \hat{b} - b)$ may be found, via the two-dimensional central limit theorem. – Here it turns out (i) that the ML estimators are more precise than the moment estimators, and (ii) that finding the limit distribution is rather easier via the general results about ML behaviour, already sorted out in Ex. 5.10. digamma function

(c) Then turn attention to the two-parameter Gamma model, where arguments and results will be similar. Show that the mean and variance are a/b and a/b^2 , and find moment estimators $\hat{a}_m, \hat{b}_m$ based on this. Find the limit distribution of $\sqrt{n}(\hat{a}_m - a, \hat{b}_m - b)$.

(d) Then write down the log-likelihood function, take derivatives, and show that the ML estimators $(\hat{a}, \hat{b})$ are the solutions to the two equations

$$\bar{Y} = a/b, \quad n^{-1} \sum_{i=1}^n \log Y_i = \psi(a) - \log b.$$

Explain from Ex. 5.10 how the limit distribution of $\sqrt{n}(\hat{a} - a, \hat{b} - b)$ may be obtained. Explain why the ML estimators are better than the moment estimators, under model conditions.

Ex. 5.17 *Corner parameter asymptotics.* The ML apparatus has been smoothly developed and used above, with a key assumption being that the true or least false parameter θ_0 is an inner point in the parameter space. Different techniques are called for, with different results, in the case of corner parameters, when the θ_0 sits at a boundary.

(a) Starting with a simple case, suppose data Y_i are $N(\xi, 1)$, and that it is known that $\xi \geq 0$. Show that the log-likelihood function is $\ell_n(\xi) = -\frac{1}{2}n(\xi - \bar{Y})^2$ for $\xi \geq 0$, and that is maximised at $\bar{Y}$ if positive but at zero if negative. Also, explain that $D_n(0) = 2\{\ell_{n,\max} - \ell_n(0)\}$ is $n\bar{Y}^2$ if $\bar{Y}$ is positive, i.e. a clear χ_1^2 , but zero if $\bar{Y} \leq 0$. In particular, at the corner value $\xi = 0$, $D_n(0)$ is zero or a χ_1^2 with probabilities $\frac{1}{2}, \frac{1}{2}$, which we may write as $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

(b) Similar behaviour takes place also in the rather general framework, where $\mu = \mu(\theta)$ is a focus parameter in an otherwise smooth parametric model $f(y, \theta)$, where there is constraint of type $\mu \geq \mu_0$. Use Ex. 5.13(e) to explain that $B_n(t) = \ell_{n,\text{prof}}(\mu_0 + t/\sqrt{n}) - \ell_{n,\text{prof}}(\mu_0)$ can be represented as $(W_n t - \frac{1}{2}t^2)/\tau^2 + o_{\text{pr}}(1)$, now for $t \geq 0$ only, with $W_n = c^t J^{-1} U_n \rightarrow_d W = c^t J^{-1} U$ and $\tau^2 = c^t J^{-1} c$, in notation from that exercise. Explain that with $\hat{\mu}$ the ML estimator, which can be μ_0 with positive probability, we have, at the boundary position, (i) $\sqrt{n}(\hat{\mu} - \mu_0) \rightarrow_d \max(0, W/\tau^2)$ and (ii) $D_n(\mu_0) = 2\{\ell_{n,\max} - \ell_n(\mu_0)\}$ tending to the variable which is 0 or a χ_1^2 with probabilities $\frac{1}{2}, \frac{1}{2}$.

(c) For this limit distribution, say L , show that its c.d.f. is $\Pr(L \leq x) = \frac{1}{2} + \frac{1}{2}\Gamma_1(x)$. For e.g. the 0.95 quantile, explain that we need the 0.90 quantile of the χ_1^2 , i.e. 2.705, not the customary $3.841 = 1.96^2$.

(d) Let Y have the t distribution with degrees of freedom ν . Wishing to understand what happens with large ν , i.e. close to normality, we parametrise to $\gamma = 1/\nu$. Show that the log-density is

$$\log g(y, \gamma) = \log \Gamma(\frac{1}{2} + \frac{1}{2}\gamma) - \log \Gamma(\frac{1}{2}\gamma) - \frac{1}{2} \log(1/\gamma) \log \pi - \frac{1}{2}(1 + 1/\gamma) \log(1 + \gamma y^2).$$

Using $\Gamma(1+x) = x\Gamma(x)$ we can study $\Gamma(x)$ for small values of x . With careful Taylor expansion, show that

$$\log g(y, \gamma) = \log \phi(y) + \gamma A(y) - \frac{1}{2}\gamma^2 B(y) + O(\gamma^3),$$

with $A(y) = \frac{1}{4}y^4 - \frac{1}{2}y^2 - \frac{1}{4}$ and $B(y) = \frac{1}{3}y^6 - \frac{1}{2}y^4$. For Y a standard normal, show that $A(Y)$ has zero mean and variance $\tau^2 = 3.50$, equal also the mean of $B(Y)$. Explain that the log-likelihood function, for an i.i.d. set $Y_1, \dots, Y_n$ from the $t_{1/\gamma}$, can be written $\ell_n(\gamma) = n(\gamma\bar{A} - \frac{1}{2}\gamma^2\bar{B}) + O_{\text{pr}}(n\gamma^3)$, in terms of averages of $A(y_i)$ and $B(y_i)$. Show hence that the ML estimator $\hat{\gamma}$ is essentially $\bar{A}/\bar{B}$ if $\bar{A} > 0$, but zero if $\bar{A} \leq 0$. Write $\sqrt{n}\bar{A} \rightarrow_d U \sim N(0, \tau^2)$. Explain that $\sqrt{n}\hat{\gamma}$ tends to $W = \max(0, U)$, i.e. $W = U$ if $U > 0$ but zero if $U \leq 0$. Show also that

$$2\{\ell_n(\hat{\gamma}) - \ell_n(0)\} \rightarrow_d \max(0, U)^2/\tau^2,$$

which is a χ_1^2 with probability $\frac{1}{2}$, but zero with the complementary probability $\frac{1}{2}$.

Ex. 5.18 *Maximum likelihood for a two-parameter extension of the exponential.* (xx let's see. this is used in Story #65, so we have a choice regarding what is where. xx) Consider

the model for positive w with c.d.f. and density

$$G(w, a, b) = 1 - \{b/(b+w)\}^a \quad \text{and} \quad g(w, a, b) = ab^a/(b+w)^{a+1}.$$

(a) Show that the c.d.f. also can be written $1 - \exp\{-a \log(1 + w/b)\}$, and explain that for b large this is close to the exponential $1 - \exp(-\lambda w)$, with $\lambda = a/b$. To learn more about the log-density and log-likelihood behaviour when a and b become large, with stable a/b , reparametrise to $(a, b) = (\lambda/c, 1/c)$. Show that the log-density becomes

$$\log g = \log \lambda - (\lambda + c)h(w, c),$$

with $h(w, c) = (1/c) \log(1 + cw) = w - \frac{1}{2}cw^2 + \frac{1}{3}c^2w^3 + \dots$. Find the score functions

$$\partial/\partial\lambda = 1/\lambda - h(w, c) \quad \text{and} \quad \partial/\partial c = -h(w, c) - (\lambda + c)h'(w, c),$$

with h' and h'' below signalling derivatives with respect to c , and deduce the exact expressions

$$Eh(W, c) = \frac{1}{\lambda} \quad \text{and} \quad Eh'(W, c) = -\frac{1}{\lambda} \frac{1}{\lambda + c} = -\frac{1}{\lambda^2} \frac{1}{1 + c/\lambda}.$$

For the Fisher information matrix $J(\lambda, c)$, show that $J_{1,1} = 1/\lambda^2$, $J_{1,2} = Eh'(W, c)$. For the lower right element, do $\partial^2/\partial c^2 = -2h'(w, c) - (\lambda + c)h''(w, c)$ and derive from this that

$$J_{2,2} = -\frac{2}{\lambda^2} \frac{1}{1 + c/\lambda} + (\lambda + c)Eh''(W, c) = 2/\lambda^2 + O(c).$$

Explain that this leads to

$$J(\lambda, c) = \frac{1}{\lambda^2} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} + O(c), \quad \text{with} \quad J(\lambda, c)^{-1} = \lambda^2 \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} + O(c),$$

(b) If $c = 0$, i.e. the data are $\text{Expo}(\lambda)$, show using Ex. 5.17 that $\sqrt{n}\hat{c} \rightarrow_d \lambda \max(0, N)$, with N a standard normal, and that the test statistic $W_n = \sqrt{n}\hat{c}/\hat{\lambda}$ tends to $\max(0, N)$. Explain also that $W_n^2 \rightarrow_d \frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2$, equal to zero with probability $\frac{1}{2}$.

(c) With the parametrisation above, show that the log-likelihood for an i.i.d. set of data becomes

$$\ell(\lambda, c) = n\{\log \lambda - (\lambda + c)B_n(c)\}, \quad \text{with} \quad B_n(c) = n^{-1} \sum_{i=1}^n h(w_i, c).$$

Maximising this over all λ , for given c , show that $\hat{\lambda}_c = 1/B_n(c)$, yielding the log-likelihood profile as

$$\ell_{\text{prof}}(c) = n\{-\log B_n(c) - cB_n(c) - 1\}.$$

Explain that for small c , we have $B_n(c) = \bar{w} - \frac{1}{2}cC_n + \frac{1}{3}c^2D_n$, with C_n and D_n the averages of w_i^2 and w_i^3 . Write $C_n = v_n + \bar{w}^2$ with $v_n = n^{-1} \sum_{i=1}^n (w_i - \bar{w})^2$. Show that this again leads to

$$\ell_{\text{prof}}(c) = n[-\log \bar{w} + \frac{1}{2}cC_n/\bar{w} - c\bar{w}] + O_{\text{pr}}(nc^2),$$

with derivative

$$\ell'_{\text{prof}}(0) = n(\frac{1}{2}C_n/\bar{w} - \bar{w}) = n(\frac{1}{2}v_n/\bar{w} + \frac{1}{2}\bar{w} - \bar{w}) = \frac{1}{2}n\bar{w}(v_n/\bar{w}^2 - 1)$$

at zero. If $v_n/\bar{w}^2 > 1$, the ML $\hat{c}$ is positive, otherwise it is zero.

Ex. 5.19 *Log-linear mixing of densities.* How far have we moved, from A to B? Suppose i.i.d. data $Y_1, \dots, Y_n$ come from a density which is perceived of as being ‘between’ given densities f_A and f_B . There are various ways of creating probabilistic bridges from A to B, via which one then can estimate the position of the current density, e.g. to assess whether it is close to A or to B. One such model takes $f(y, \lambda) = f_A(y)^{1-\lambda} f_B(y)^\lambda / R(\lambda)$, with $R(\lambda)$ the normalisation constant $\int f_A^{1-\lambda} f_B^\lambda dy$. This creates such a bridge, with $\lambda \in [0, 1]$, and with A and B corresponding to $\lambda = 0$ and $\lambda = 1$.

(a) Explain that $R(0) = R(1) = 1$, and that the derivative of $R(\lambda)$, at the endpoints 0 and 1, can be written

$$R'(0) = -\text{KL}(f_A, f_B), \quad R'(1) = \text{KL}(f_B, f_A).$$

in terms of the Kullback–Leibler distances discussed in Ex. 5.4 and later on. In particular, the $R(\lambda)$ has negative derivative at the start and positive derivative at the end.

(b) For a concrete illustration, suppose f_A is $N(0, 1)$ and f_B is $N(c, 1)$, with c positive. Show that $R(\lambda) = \exp\{-\frac{1}{2}c^2\lambda(1-\lambda)\}$. From this explain that the score function is $u(y, \lambda) = cy - \lambda c^2$ and that the Fisher information becomes the constant c^2 . Show indeed that $f(y, \lambda)$ is $N(\lambda c, 1)$, that the ML estimator becomes $\hat{\lambda} = \bar{Y}/c$, truncated to $[0, 1]$. From general ML theory, explain that if the real λ is inside $(0, 1)$, then $\sqrt{n}(\hat{\lambda} - \lambda) \rightarrow_d N(0, 1/c^2)$. For c small the variance will be very big, signifying that it is a hard task to estimate the balance parameter when f_A and f_B are close.

(c) Returning to the general setup, write $S(y) = \log f_B(y) - \log f_A(y)$. Show that the log-likelihood function becomes

$$\ell_n(\lambda) = n\{\lambda \bar{S}_n - \log R(\lambda)\}, \quad \text{with } \bar{S}_n = n^{-1} \sum_{i=1}^n S(y_i).$$

Explain that the ML estimator $\hat{\lambda}$ hence is the solution to $\bar{S}_n = R'(\lambda)/R(\lambda)$. For λ well inside $(0, 1)$, give the limiting distribution for $\sqrt{n}(\hat{\lambda} - \lambda)$. Using methods and insights from Ex. 5.17, show that if data really come from f_A , then $\hat{\lambda}$ can be zero with positive probability, and that $\sqrt{n}\hat{\lambda}$ has a limit distribution of the type $\sigma_A \max(0, N)$, with N a standard normal. Similarly, spell out the limit distribution for $\sqrt{n}(1 - \hat{\lambda})$ when data are from f_B . Use these results to set up tests for $\lambda = 0$ and $\lambda = 1$, i.e. the data coming from f_A or f_B .

Ex. 5.20 *Parametric precision when estimating densities.* Assume $X_1, \dots, X_n$ are i.i.d. on the real line. In Ex. 3.42 we briefly investigated the kernel density estimator f_n , and found that under smoothness conditions, the $\text{mise}_n = \int (f_n - f)^2 dx$ is of order $O(n^{-4/5})$ for the best bandwidths. Parametric precision is however better, provided the model can be trusted.

(a) Find useful expressions for the point-wise mean squared error $\text{mse}(x) = E\{\hat{f}(x) - f(x)\}^2$ in the following cases: (i) f the p.d.f. of $N(\xi, \sigma^2)$ with σ known; (ii) f as in (i), but with ξ known; (iii) f as in (i), but with both parameters known; (iv) the exponential $f(x, \theta) = \theta \exp(-\theta x)$.

(b) Then find expressions for the integrated mean squared error $\text{mise} = E \int (\hat{f} - f)^2 dx$, in cases (i)–(iv).

(c) Assume in general terms that data follow a parametric model $f(x, \theta)$, with θ_0 the true parameter, of dimension say p . Writing $u(x, \theta) = \partial \log f(x, \theta) / \partial \theta$ for the score

function, use the general ML theory to show that

$$\begin{aligned}\sqrt{n}\{f(x, \hat{\theta}) - f(x, \theta_0)\} &\rightarrow_d f(x, \theta_0)u(x, \theta_0)^t Z, \\ \sqrt{n}f(x, \theta_0)^{1/2}\{\log f(x, \hat{\theta}) - \log f(x, \theta_0)\} &\rightarrow_d f(x, \theta_0)^{1/2}u(x, \theta_0)^t Z,\end{aligned}$$

where $Z \sim N_p(0, J^{-1})$, with $J = J(\theta_0)$ the Fisher information matrix. Then use this to show that with

$$A_n = \int \frac{\{f(x, \hat{\theta}) - f(x, \theta_0)\}^2}{f(x, \theta_0)} dx \quad \text{and} \quad B_n = \int f(x, \theta_0)\{\log f(x, \hat{\theta}) - \log f(x, \theta_0)\}^2 dx,$$

we have both nA_n and nB_n tending in distribution to $W = Z^t J Z \sim \chi_p^2$. So various versions of weighted MISE go to zero with speed p/n , for parametrically estimated densities, under model conditions, in contrast to the slower $O(n^{4/5})$ speed for nonparametric estimators.

Ex. 5.21 *Completing increasingly simpler tasks.* In a certain game of learning a player needs to complete tasks $1, 2, \dots, n$, which become increasingly simpler with each passing of a new level. Assume that the time needed to complete these tasks are $V_1, \dots, V_n$, with these being independent with $V_i \sim \text{Expo}(i/\theta)$, where θ is an unknown parameter. – For the following questions, you may encounter the partial sums

$$H_n = 1 + 1/2 + 1/3 + \dots + 1/n, \quad b_n = 1 + 1/2^2 + 1/3^2 + \dots + 1/n^2.$$

Here the first is slowly divergent, with $H_n \doteq \log n + 0.5772$, and the second is convergent, with $b_n \rightarrow \pi^2/6$.

Euler 1734; see
Ex. 1.64

(a) Find expressions for the mean and variance of $T_n = V_1 + \dots + V_n$, the time it takes the player to complete all tasks. In particular, show that T_n has mean $H_n\theta$. Put up the unbiased estimator based on T_n , say $\hat{\theta}$. Find its variance, and show that the estimator is consistent.

(b) Then work out a formula for the log-likelihood function, based on having observed not merely the total time T_n , but the individual waiting times $V_1, \dots, V_n$. Find the maximum likelihood estimator, say θ^* . Show that also this estimator θ^* is unbiased, and compare its variance to that of $\hat{\theta}$. Find also the Cramér–Rao lower bound for variances of unbiased estimators for θ , and comment.

(c) Assume the game goes on, up to level $2n$, and consider the time a player needs to pass the last half of these levels, i.e. $T_n^* = T_{2n} - T_n$. Show that T_n^* tends in probability to a certain limit as n grows.

Ex. 5.22 *Wilks theorems for the normal model.* Consider the familiar normal situation with $Y_1, \dots, Y_n$ independent from $N(\xi, \sigma^2)$. The log-likelihood function is $\ell_n(\xi, \sigma)$, as in (5.2), and we have already seen in Ex. 5.15 that the ML estimators are $\hat{\xi} = \bar{Y}$ and $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Here we work with profiled consequences, also to illustrate the Wilks theorem of Ex. 5.13.

(a) For σ in focus, show that

$$\ell_{n,\text{prof}}(\sigma) = -n \log \sigma - \frac{1}{2} Q_0 / \sigma^2 - \frac{1}{2} n \log(2\pi),$$

with $Q_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. From this show that the deviance function may be written

$$D_n(\sigma) = 2\{\ell_{n,\text{max}} - \ell_{n,\text{prof}}(\sigma)\} = nh(\hat{\sigma}^2/\sigma^2),$$

with $h(x) = x - \log x - 1$. The Wilks theorem implies that $D_n(\sigma) \rightarrow_d \chi_1^2$, under model conditions; also verify this via a two-term Taylor expansion for $h(x)$ around $x = 1$.

(b) Then with attention to ξ , show that

$$\ell_{n,\text{prof}}(\xi) = -n \log \hat{\sigma}(\xi) - \frac{1}{2}n\{1 + \log(2\pi)\},$$

with $\hat{\sigma}(\xi)^2 = n^{-1}\{Q_0 + n(\xi - \bar{Y})^2\}$. From this show that

$$D_n(\xi) = 2\{\ell_{n,\text{max}} - \ell_{n,\text{prof}}(\xi)\} = n \log(1 + n^{-1}K_n(\xi)), \quad \text{with } K_n(\xi) = n(\bar{Y} - \xi)^2/\hat{\sigma}^2.$$

Verify directly what the Wilks theorem says here, that $D_n(\xi) \rightarrow_d \chi_1^2$.

(c) The Wilks theorem gives in general terms the recipe $\{\mu: D_n(\mu) \leq 1.96^2\}$ for an approximate 95 percent confidence interval. With some simulated data, with sample sizes say 10, 30, 50, check what this recipe gives for forming intervals for ξ and for σ , compared to the exact ones.

(d) Importantly, the Wilks theorem with consequent confidence recipe can be used for *any* smooth function of the parameters here, i.e. not merely for ξ and for σ . With some simulated data, illustrate this, by computing the log-likelihood profile function and the deviance function, for parameters $\gamma = \xi + 1.96\sigma$ and $p = 1 - \Phi((y_0 - \xi)/\sigma)$, for suitable threshold y_0 .

Likelihood machinery for multinomial data

Ex. 5.23 *log-likelihood and ML for the binomial, trinomial, multinomial.* Working through the likelihood mechanics of the binomial, trinomial, multinomial models provide good illustrations of the developed methodology. We already know the basic large-sample behaviour for the natural estimators in these models, via Ex. 2.51, but here we connect such results to the general likelihood theory.

(a) For $X \sim \text{binom}(n, p)$, a sum of independent Bernoulli variables, show that its log-likelihood function is $\ell_n(p) = X \log p + (n - X) \log(1 - p)$, with maximiser $\hat{p} = X/n$. Show that $nJ = n/\{p(1 - p)\}$, here with J defined relative to a single Bernoulli variable. Explain that we may hence read off the limit distribution of $\sqrt{n}(\hat{p} - p)$ being $N(0, p(1 - p))$, without necessarily even knowing the X/n formula, or about the CLT for binomials. One may even argue the other way, starting with the X/n formula: since we know the limit is $N(0, p(1 - p))$, via the CLT, we must have $J^{-1} = p(1 - p)$, hence J must be $1/\{p(1 - p)\}$.

(b) For further illustration of the binomial likelihood mechanics, draw the log-likelihood function $\ell_n(p)$, computing both its maximiser, its maximum, and its second derivative at the maximum, for cases (i) $n = 20, y = 4$, (ii) $n = 40, y = 8$, (iii) $n = 200, y = 40$. Note how $\hat{J}_{\text{obs}} = -\ell_n''(\hat{p})$ becomes bigger and the implied estimated standard deviance $1/\hat{J}_{\text{obs}}^{1/2}$ becomes smaller.

(c) Let (X, Y) be trinomial (n, p, q) , and use $r = 1 - p - q$ and $Z = n - X - Y$. Show that the log-likelihood function becomes $X \log p + Y \log q + Z \log(1 - p - q)$, and for the Fisher information matrix that

$$J = \begin{pmatrix} 1/p + 1/r & 1/r \\ 1/r & 1/q + 1/r \end{pmatrix} \quad \text{with} \quad J^{-1} = \begin{pmatrix} p(1 - p) & -pq \\ -pq & q(1 - q) \end{pmatrix}.$$

Conclude, even before finding or perhaps without caring that there are clear formulae $\hat{p} = X/n$ and $\hat{q} = Y/n$ for the ML estimators, that $(\sqrt{n}(\hat{p} - p), \sqrt{n}(\hat{q} - q))$ tends to the binormal zero-mean distribution with the J^{-1} covariance matrix. This may of course also be shown, for $(X/n, Y/n)$, without knowing that these are ML estimators.

Ex. 5.24 *Maximum likelihood and Wilks testing for multinomials.* In generalisation of Ex. 5.23, consider a multinomial counts $(N_1, \dots, N_k)$, with sum n and probabilities $p = (p_1, \dots, p_k)$ for the k categories in question.

(a) Before we come to parametrised models, i.e. no constraints are made on the p_j apart from their sum being 1, show that the log-likelihood function may be written $\ell_n(p) = \sum_{j=1}^k N_j \log p_j$, and that this is maximised at $\hat{p}$, with $\hat{p}_j = N_j/n$. Explain that each of these are binomial proportions, where we e.g. know that $\sqrt{n}(\hat{p}_j - p_j) \rightarrow_d X_j \sim N(0, p_j(1 - p_j))$.

(b) We need joint convergence here, to understand other properties of the multinomial counts. Show that the vector of $X_{n,j} = \sqrt{n}(\hat{p}_j - p_j)$ tends to the vector $X = (X_1, \dots, X_k) \sim N_k(0, \Sigma)$ in distribution, where Σ is the $k \times k$ matrix with elements $p_j(\delta_{j,\ell} - p_\ell)$. Show that $\mathbf{1}^t \Sigma \mathbf{1} = 0$, with $\mathbf{1}$ the vector of 1s, so Σ is not invertible. Explain that we from general ML theory know that $(X_{n,1}, \dots, X_{n,k-1}) \rightarrow_d (X_1, \dots, X_{k-1}) \rightarrow_d N_{k-1}(0, \Sigma_0)$, with $\Sigma_0 = J^{-1}$, in terms of the Fisher matrix J for $(p_1, \dots, p_{k-1})$. Argue therefore that we already know the form of J^{-1} , and that J must be equal to

$$J = \Sigma_0^{-1} = \text{diag}(1/p_1, \dots, 1/p_{k-1}) + (1/p_k) \mathbf{1}_0 \mathbf{1}_0^t,$$

writing $\mathbf{1}_0$ for the $(k-1)$ -vector of 1s. Derive indeed the J matrix expression from $-n^{-1} \partial^2 \ell_n(\hat{p}) / \partial p_0 \partial p_0^t$. Note that this generalises results of Ex. 5.23.

(c) Suppose one needs to test $p = p_0$, a fixed probability vector. Show via the log-likelihood ratio machinery that $W_n = 2n \sum_{j=1}^k \hat{p}_j \log(\hat{p}_j / p_{0,j}) \rightarrow_d \chi_{\text{df}}^2$, under the null model condition, with $\text{df} = k - 1$ degrees of freedom. Give a confidence region R_n for the full probability vector p , such that $\Pr_p(p \in R_n) \rightarrow 0.95$.

(d) Often in multinomial models there are parametric theories concerning their probabilities. Suppose $p_j = p_j(\theta)$, for $j = 1, \dots, k$, smooth functions of θ , a parameter of dimension $r < k - 1$. Write $u_j(\theta)$ and $\psi_j(\theta)$ for the 1st and 2nd derivatives of $\log p_j(\theta)$, where we below write p_j, u_j, ψ_j , for short, when these are evaluated at the true parameter θ_0 . Show from $\sum_{j=1}^k p_j(\theta) = 1$ that we have the Bartlett type identities $\sum_{j=1}^k p_j u_j = 0$ and $J = \sum_{j=1}^k p_j u_j u_j^t = -\sum_{j=1}^k p_j \psi_j$. Explain next that the log-likelihood function is $\ell_n(\theta) = n \sum_{j=1}^k \hat{p}_j \log p_j(\theta)$, with

$$\partial \ell_n(\theta) / \partial \theta = n \sum_{j=1}^k \hat{p}_j u_j(\theta), \quad \partial^2 \ell_n(\theta) / \partial \theta \partial \theta = n \sum_{j=1}^k \hat{p}_j \psi_j(\theta).$$

With $\hat{\theta}$ the ML estimator, and following the general results of Ex. 5.10, see (5.5), or arguing directly in the present multinomial setup, show that

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = U_n^t s - \frac{1}{2} s^t J_n s + o_{\text{pr}}(1),$$

in terms of $U_n = \sum_{j=1}^k X_{n,j} u_j \rightarrow_d U \sim N_r(0, J)$ and $J_n = -n^{-1} \sum_{j=1}^k \hat{p}_j \psi_j \rightarrow_{\text{pr}} J$. From this explain that we also have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N_r(0, J^{-1}) \quad \text{and} \quad 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \rightarrow_d \chi_{\text{df}}^2, \quad \text{with } \text{df} = k - 1 - r.$$

(e) For an illustration, suppose (X, Y, Z) is trinomial probabilities $a^2, 2a(1-a), (1-a)^2$, as e.g. with so-called Hardy–Weinberg equilibrium. Find the ML estimator of a and its limit distribution, and construct a 90 percent confidence interval.

Ex. 5.25 *Pearson and minimum chi-squared for multinomials.* Consider again the setup of Ex. 5.24. The Karl Pearson chi-squared, dating all the way back to his far-reaching 1900 paper, is the sum of normalised squares

$$Q_n = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} = n \sum_{j=1}^k \frac{(\hat{p}_j - p_j)^2}{p_j},$$

measuring distance from the vector of p_j to the vector of $\hat{p}_j = N_j/n$ in a different fashion than with the log-likelihood ratio statistics worked with in the earlier exercises.

(a) With notation $X_{n,j} = \sqrt{n}(\hat{p}_j - p_j) \rightarrow_d X_j$, as in the previous exercise, show that $Q_n \rightarrow_d Q = \sum_{j=1}^k X_j^2/p_j$. Show that both Q_n and Q have means $k-1$. The Q is a sum of squared but dependent normals, and $\sum_{j=1}^k X_j = 0$. Show that Q_n may be expressed as $X_{n,0}^t \Sigma_0^{-1} X_{n,0}$, and that its limit must be the quadratic form $X_0^t \Sigma_0^{-1} X_0$. Conclude that $Q_n \rightarrow_d \chi_{k-1}^2$, which is the crowning achievement of [Pearson \(1900\)](#).

(b) Show that

$$Q_n - Q'_n = n \sum_{j=1}^k \frac{(\hat{p}_j - p_j)^2}{p_j} - n \sum_{j=1}^k \frac{(\hat{p}_j - p_j)^2}{\hat{p}_j} \rightarrow_{\text{pr}} 0,$$

so the two have the same large-sample behaviour; both may be used to test some $p = p_0$.

(c) Then consider a parametric model $p_j = p_j(\theta)$, as in Ex. 5.24. An estimation method different from the ML is to minimise

$$Q_n(\theta) = \sum_{j=1}^k \frac{\{N_j - p_j(\theta)\}^2}{N_j} = n \sum_{j=1}^k \frac{\{\hat{p}_j - p_j(\theta)\}^2}{\hat{p}_j} = n \sum_{j=1}^k \frac{p_j(\theta)^2}{\hat{p}_j} - n.$$

Verify this last identity.

(d) It turns out that this alternative estimation method is close to the ML one, with the difference vanishing for growing n . This essentially lies with the first two derivatives of $\ell_n(\theta)$ and $Q_n(\theta)$ being closely related limit, under the model. xx

(e) In addition to computing the full Pearson test statistic, to decide on acceptance vs. rejection of a model, typically along with finding the p-value in question, it is often useful to examine the individual terms in the chi-squared sum. Verify that with

$$\text{pear}_j = \{N_j - np_j(\hat{\theta})\}/(np_j(\hat{\theta}))^{1/2} \quad \text{and} \quad \text{pear}'_j = \{N_j - np_j(\hat{\theta})\}/N_j^{1/2},$$

which we call Pearson residuals, we have $\sum_{j=1}^k \text{pear}_j^2 = Q_{n,\min}$ and $\sum_{j=1}^k (\text{pear}'_j)^2 = Q'_{n,\min}$. If the model is good, the Pearson residuals ought to be approximately standard normal, so we may easily spot the boxes j where there is significant discrepancy between observed N_j and predicted $np_j(\hat{\theta})$.

(f) Suppose Pearson's $Q_n = \sum_{j=1}^k (N_j - np_{j,0})^2 / (np_{j,0})$ is used for testing the hypothesis H_0 that $p = p_0$, for a given relevant $(p_{0,1}, \dots, p_{k,0})$. At the null, $Q_n \rightarrow_d \chi_{k-1}^2$, but what happens outside H_0 ? Suppose the real state of affairs is $p_j = p_{j,0} + \delta_j/\sqrt{n}$, with $\sum_{j=1}^k \delta_j = 0$. With the notation above, show that $X_{n,j} = \sqrt{n}(\hat{p}_j - p_{j,0}) \rightarrow_d X_j + \delta_j$. From this, demonstrate that $Q_n \rightarrow_d Q = \sum_{j=1}^k (X_j + \delta_j)^2/p_j$, and that this has a noncentral $\chi_{k-1}^2(\lambda^2)$ distribution, with excentre parameter $\lambda^2 = \sum_{j=1}^k \delta_j^2/p_{j,0}$. Give an interpretation for λ . What we have found is useful for studying power functions, but also for assessing the distance from the null hypothesis to reality, i.e. the size of λ . Give an estimator for this λ , based on having observed Q_n .

“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”, no less

Pearson residuals

local power

Ex. 5.26 *Contingency tables, and more on Pearson vs. Wilks.* In Exercises 5.24–5.25 we have seen how inference can be carried out in multinomial models, via either Pearson type minimum chi-squared or Wilks type log-likelihood. We learn a bit more here, about their large-sample equivalence, under model conditions.

(a) Consider again the multinomial setting of these two previous exercises, and with a smooth parametric model $p_j = p_j(\theta)$, with ensuing log-likelihood function $\ell_n(\theta) = \sum_{j=1}^k N_j \log p_j(\theta)$. Show in general terms that for p close to $\hat{p}$, we have

$$\hat{p} \log p = \hat{p} \log(\hat{p} + p - \hat{p}) = \hat{p} \log p + (1/\hat{p})(p - \hat{p}) - \frac{1}{2}(1/\hat{p}^2)(p - \hat{p})^2 + O_{\text{pr}}(|\hat{p} - p|^3).$$

Explain that this implies

$$\sum_j \hat{p}_j \log p_j(\theta) = \sum_j \hat{p}_j \log \hat{p}_j - \frac{1}{2} \sum_j \{\hat{p}_j - p_j(\theta)\}^2 / \hat{p}_j + O_{\text{pr}}(\max_j |\hat{p}_j - p_j(\theta)|^3).$$

The implication for the log-likelihood is that

$$\ell_n(\theta) = \ell_{\text{wide,max}} - \frac{1}{2} Q_n(\theta) + \varepsilon_n(\theta), \quad \text{with } Q_n(\theta) = n \sum_j \{\hat{p}_j - p_j(\theta)\}^2 / \hat{p}_j,$$

the Karl Pearson type weighted sum of squares. If the model holds, for some θ_0 , show that the remainder term goes to zero in probability inside neighbourhoods $\|\theta - \theta_0\| \leq c/\sqrt{n}$. Argue from this that ML estimation for large samples is equivalent to minimum chi-squared, with $K_n = Q_n(\hat{\theta}) = \min Q_n(\theta)$. Deduce also for the Wilks statistic, see 5.14, that

$$W_n = 2(\ell_{\text{wide,max}} - \ell_{0,\text{max}}) = K_n + o_{\text{pr}}(1).$$

So from the exercise pointed to we learn that both K_n and W_n have the χ_{df}^2 limit, under model conditions, with $\text{df} = k - 1 - \dim(\theta)$. This extends the classical Karl Pearson 1900 work, where he developed the chi-squared for a fixed set of probabilities.

(b) We now consider a contingency table, with a multinomial $N_{i,j}$, the number of ($A = i, B = j$), for an $r \times s$ table, and probabilities $p_{i,j}$, summing to 1. Let $\Pr(A = i) = p_{i,\cdot} = a_i$, $\Pr(B = j) = p_{\cdot,j} = b_j$, with the ‘ $\cdot$ ’ indicating summation over the index in question. Independence corresponds to $p_{i,j} = a_i b_j$ for all pairs. Letting $\hat{p}_{i,j} = N_{i,j}/n$ be the direct estimates, show that the log-likelihood function for the independence model is

$$\ell(a, b) = n \sum_{i,j} \hat{p}_{i,j} (\log a_i + \log b_j) = n \sum_i \hat{p}_{i,\cdot} \log a_i + n \sum_j \hat{p}_{\cdot,j} \log b_j$$

with ML estimators $\hat{a}_i = \hat{p}_{i,\cdot}$ and $\hat{b}_j = \hat{p}_{\cdot,j}$. Deduce that the Wilks deviance statistic, see Ex. 5.14, is

$$W_n = 2n \left(\sum_{i,j} \hat{p}_{i,j} \log \hat{p}_{i,j} - \sum_i \hat{a}_i \log \hat{a}_i - \sum_j \hat{b}_j \log \hat{b}_j \right).$$

Use results above to learn that under the null model of independence, W_n and the classic

$$K_n = n \sum_{i,j} \frac{(\hat{p}_{i,j} - \hat{a}_i \hat{b}_j)^2}{\hat{p}_{i,j}} = \sum_{i,j} \frac{(N_{i,j} - E_{i,j})^2}{E_{i,j}}$$

have the same limit distribution χ_{df}^2 , with $\text{df} = (r - 1)(s - 1)$, under the null. Here $E_{i,j} = n \hat{a}_i \hat{b}_j$ are the expected numbers in the cells, under the null; these are given to the right of the table of observed numbers. Show also that the same chi-squared limit obtains, under independence, whether one uses $E_{i,j}$ or $N_{i,j}$ in the denominator.

(c) Suppose the independence model does not quite hold up, and that $p_{i,j} = a_i b_j + \delta_{i,j}/\sqrt{n}$, for a set of deviations $\delta_{i,j}$. Show that these then need to sum to zero, also for the marginals, i.e. $\sum_i \delta_{i,j} = 0$, $\sum_j \delta_{i,j} = 0$. Show that $K_n \rightarrow_d \chi_{\text{df}}^2(\lambda)$, with $\lambda = \sum_{i,j} \delta_{i,j}^2/(a_i b_j)$. Explain implications for the local power of the Pearson test against independence.

(d) (xx nice if we could complete the below things. relies on a lemma saying $X \sim \chi_a^2$ and $X + Y \sim \chi_{a+b}^2$ implies Y independent of X and being χ_b^2 . the below would then be a new little proof of chi-squared limit for K_n . at any rate, make clear from the prose that we have different proofs for the chi-square limit. xx) we have found the χ_{df}^2 limit in other ways, with $\text{df} = (r-1)(s-1)$, but it is interesting to understand another path too. we use $\sqrt{n}(\hat{p}_{i,j} - p_{i,j}) \rightarrow_d A_{i,j}$, a big zero-mean multinormal. then work under independence with

$$\sqrt{n}(\hat{p}_{i,j} - \hat{a}_i \hat{b}_j) \rightarrow_d B_{i,j} = A_{i,j} - a_i A_{\cdot,j} - b_j A_{i,\cdot}.$$

So $K_n = n \sum_{i,j} (\hat{p}_{i,j} - \hat{a}_i \hat{b}_j)^2 / \hat{p}_{i,j}$ tends to $K = \sum_{i,j} B_{i,j}^2 / (a_i b_j)$. Multiply out and simplify to get

$$K = \sum_{i,j} \frac{(A_{i,j} - a_i A_{\cdot,j} - b_j A_{i,\cdot})^2}{a_i b_j} = \sum_{i,j} \frac{A_{i,j}^2}{a_i b_j} - \sum_i \frac{A_{i,\cdot}^2}{a_i} - \sum_j \frac{A_{\cdot,j}^2}{b_j}.$$

Intriguingly, the $A_{i,\cdot}$ and $A_{\cdot,j}$ are all independent, under the null, so $K = K_0 - K_a - K_b$, or $K_0 = K_a + K_b + K$, where K_a and K_b are independent and χ_{r-1}^2 , χ_{s-1}^2 , and we also know by KP 1900 story that $K_0 \sim \chi_{rs-1}^2$. but it is not clear how to derive that $K \sim \chi_{(r-1)(s-1)}^2$ from this. might involve some clever algebraic rewriting and Cochran's theorem.

(e) Often enough the two factors have dependencies, of course, and one needs models to assess such. Explain that for any score matrix $h_{i,j}$, the model

$$p_{i,j} = a_i b_j \exp(\gamma h_{i,j}) / \sum_{i',j'} a_{i'} b_{j'} \exp(\gamma h_{i',j'}).$$

is a bona fide 1-parameter extension of independence, How many free parameters are there in such a model?

Ex. 5.27 Comparing two multinomials. Consider two multinomial vectors $M = (M_1, \dots, M_k)$ and $N = (N_1, \dots, N_k)$, with comparable probability vectors $p = (p_1, \dots, p_k)$ and $q = (q_1, \dots, q_k)$. How can we naturally test the hypothesis H_0 that $p = q$?

(a) Writing n_1 and n_2 for the two sample sizes, with sum n , write $\hat{p}_j = M_j/n_1$ and $\hat{q}_j = N_j/n_2$. Explain first that

$$\sqrt{n}(\hat{p} - p) \rightarrow_d (1/c_1^{1/2})X \sim N_k(0, \Sigma_1), \quad \sqrt{n}(\hat{q} - q) \rightarrow_d (1/c_2^{1/2})Y \sim N_k(0, \Sigma_2),$$

where $\Sigma_1 = D_1 - pp^t$, $\Sigma_2 = D_2 - qq^t$, with diagonal matrices D_1 and D_2 having p and q on their diagonals. Here $c_1 = n_1/n$ and $c_2 = n_2/n$, meant to stay stable for growing sample sizes.

(b) Under H_0 , then, show that $\sqrt{n}(\hat{p} - \hat{q}) \rightarrow_d (1/c_1 + 1/c_2)^{1/2} Z$, with $Z \sim N_k(0, D - pp^t)$. Deduce that

$$K_{n_1, n_2} = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^k \frac{(\hat{p}_j - \hat{q}_j)^2}{\hat{r}_j} \rightarrow_d \chi_{k-1}^2,$$

as long as $\hat{r}_j \rightarrow_{\text{pr}} p_j$ for each j , as with e.g. $\hat{r}_j = c_1 \hat{p}_j + c_2 \hat{q}_j = (n_1 X_j + n_2 Y_j)/(n_1 + n_2)$.

check Story
#61 for an
illustration with
fathers and sons
and social
mobility

check Story
#45 for an
application

Maximum likelihood machinery in practice

Ex. 5.28 *Estimating key matrices.* The broadly applicable ML machinery, developed above, relies on having consistent estimators for variances of limiting distributions, via general arguments seen also in connection with Wald tests, see Ex. 4.5: we need not only $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \tau^2)$, say, but also a consistent $\hat{\tau}$. Here we check conditions under which certain matrices can be estimated consistently.

(a) In general terms, the matrix A_n is consistent for A if all elements converge appropriately, i.e. $A_{n,i,j} \rightarrow_{\text{pr}} A_{i,j}$ for all (i, j) . For this statement to be precise a matrix norm is required, and we take $\|A_n - A\| = \max_{i,j} |A_{n,i,j} - A_{i,j}|$. Prove the assertion.

(b) Consider first some function $p(y, \theta)$ with finite mean in a neighbourhood of the true θ_0 , with $\hat{\theta}$ an estimator of θ_0 . Explain first that $p_n = n^{-1} \sum_{i=1}^n p(Y_i, \theta_0)$ tends to $p_0 = E_G p(Y, \theta_0)$. The best we may do for estimating p_0 in practice is $\hat{p}_n = n^{-1} \sum_{i=1}^n p(Y_i, \hat{\theta})$. Show that if $|p(y, \theta_0 + \varepsilon) - p(y, \theta_0)| \leq M(y)\|\varepsilon\|$, for all small $\|\varepsilon\|$, for some function $M(y)$ with finite mean, then indeed $\hat{p}_n \rightarrow_{\text{pr}} p_0$.

(c) For estimating $J = -E_g \partial^2 \log f(Y, \theta_0) / \partial \theta \partial \theta^t$, give (mild) conditions under which $\hat{J}_n = -n^{-1} \partial^2 \ell_n(\hat{\theta}) / \partial \theta \partial \theta^t$ is consistent for J . This matrix is typically computed directly, as a by-product of the numerical maximisation of the log-likelihood. Analogously, for the matrix $K = \text{Var}_g u(Y, \theta_0)$, show that $\hat{K} = n^{-1} \sum_{i=1}^n \hat{u}_i \hat{u}_i^t$ is consistent, under mild conditions, where $\hat{u}_i = u(Y_i, \hat{\theta})$.

(d) For the ML method, we know from the Bartlett identity that the two matrices J and K are equal under model conditions. For various models we would then have two different estimators, with the same aim, and perhaps different precision; this does not matter for the first-order large-sample theory, since consistency is what matter for $\hat{J}$ and $\hat{K}$, and specifically for the estimated sandwich matrix $\hat{\Sigma} = \hat{J}^{-1} \hat{K} \hat{J}^{-1}$. For illustration, consider Y_i from the $\text{Pois}(\theta)$ model. Show that $J = 1/\theta_0$ and $K = \tau_0^2/\theta_0^2$, in terms of the true mean and variance θ_0 and τ_0^2 of the underlying distribution. The recipes above lead to $\hat{J} = 1/\bar{Y}$ and $\hat{K} = V_n/\bar{Y}^2$, in terms of sample mean and sample variance. Under model conditions, both aim for $1/\theta_0$. Show also that the consequent estimated sandwich becomes $\hat{\Sigma} = V_n$. The general recipes hence lead to two somewhat different confidence intervals for θ , namely $\hat{\theta} \pm z_0 \hat{\kappa} / \sqrt{n}$, with either $\hat{\kappa} = \bar{Y}^{1/2}$ or $\hat{\kappa} = V_n^{1/2}$. Argue that both are valid, with both aiming for the same quantity under Poisson conditions, whereas the second option might be preferred if there might be overdispersion.

Ex. 5.29 *ML analysis via optimisers, gradients, and Hessians.* As long as you can programme your log-likelihood function, as with a suitable `logL <- function(theta)`, there are typically well-working optimisation algorithms that with a modicum of extra work will provide you with (i) the ML estimator $\hat{\theta}$ itself, the maximiser; (ii) the $\ell_{\max} = \ell(\hat{\theta})$, the maximum; (iii) the Hessian matrix $\hat{J}_{\text{full}} = -\partial^2 \ell(\hat{\theta}) / \partial \theta \partial \theta^t$, the observed Fisher information; (iv) estimated score function at the data values, $\hat{u}_i = u(y_i, \hat{\theta})$, qua gradients; (v) the gradient vector $\hat{c} = \partial \mu(\hat{\theta}) / \partial \theta$ associated with any focus parameter $\mu(\theta)$, required for delta method analysis; (vi) deviance functions $D(\mu)$ for parameters of interest; (vii) Wilks type differences $\ell_{\max, \text{wide}} - \ell_{\max, \text{narr}}$; and yet other quantities for further uses.

(a) To illustrate the general features, and how the crucial quantities flow from the log-likelihood function, imagine that you have never heard about the normal distribution and that you have not carefully read Chs. 1–4. Get hold of any dataset $y_1, \dots, y_n$ under

counterfactually,
you may even
imagine that
you haven't
been through
all the theorems
and proofs

your fingertips. You are given the formula

$$f(y, \xi, \sigma) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma} \exp\left\{-\frac{1}{2} \frac{(y - \xi)^2}{\sigma^2}\right\},$$

and proceed to programme the log-likelihood function, more or less as below, in your favourite software environment. You throw the `minuslogL` to a minimiser, like `nlm` in R, which then crucially gives you `ML`, the minimiser, and `Jhatfull`, the Hessian matrix. Do this.

```
logL <- function(para)
{ xi = para[1]
  sigma = para[2]
  aux = -log(sigma) - 0.5*(yy - xi)^2/sigma^2 - 0.5*log(2*pi)
  sum(aux) }
then
minuslogL <- function(para)
{ -logL(para) }
```

(b) Read off `sdhat = sqrt(diag(solve(Jhatfull)))`, the square-roots of the diagonal of the inverse Fisher information matrix $\hat{J}_{\text{full}}$. Get 95 percent confidence intervals from `ML-1.96*sdhat` to `ML+1.96*sdhat`.

(c) Under model circumstances the above is fine, but outside model conditions we also need the following. Do the programming required to obtain $\hat{u}_i = \partial \log f(y_i, \hat{\theta})$, without mathematics, so to speak, but with a `grad` operation. From these, compute both 2×2 matrices $\hat{J} = n^{-1} \hat{J}_{\text{full}}$ and $\hat{K} = n^{-1} \sum_{i=1}^n \hat{u}_i \hat{u}_i^t$, and comment regarding their similarity. Do also the sandwich matrix $\hat{\Sigma} = \hat{J}^{-1} \hat{K} \hat{J}^{-1}$. Explain that model robust confidence intervals are as above, but now with `sdhat2 = sqrt(diag(Sigmahat))`.

(d) Consider the probability $p = p(\xi, \sigma) = \Pr(Y \geq y_0) = 1 - \Phi((y_0 - \xi)/\sigma)$. Compute the ML estimate $\hat{p} = p(\hat{\xi}, \hat{\sigma})$. Without doing the math, but just via programming, compute the 2-vector $\hat{c} = \partial p(\hat{\xi}, \hat{\sigma}) / \partial \theta$. Explain that $\hat{c}^t \hat{\Sigma} \hat{c}$ is the estimated variance of $\hat{p}$, and compute a confidence interval.

(e) Via some extra programming, get hold of the profile functions $\ell_{\text{prof}}(\xi)$ and $\ell_{\text{prof}}(\sigma)$. From these compute and display deviance curves $D_1(\xi)$ and $D_2(\sigma)$. Do this also for the focus parameter $p = p(\xi, \sigma)$ above.

(f) Choose a suitable σ_0 value to be interested in. Redo the programming above, now as a function of only ξ , to find the log-likelihood maximum $\ell_{\text{max, narr}}$. Then do Wilks to test the hypothesis $\sigma = \sigma_0$ against $\sigma \neq \sigma_0$.

Ex. 5.30 *We can do things.* The spirit of this exercise is to see that the log-likelihood machinery is useful, versatile, flexible, and not too hard to use in new situations, outside ordinary textbook terrain.

just do it

(a) One records the number of a certain event, say per week, over a time period. Suppose most of these counts are Poisson-like, with some parameter θ , but that a fraction comes from another Poisson with a higher parameter. A model for such data is that Y_i stems from the mixture distribution $(1 - p) \text{Pois}(\theta) + p \text{Pois}(c\theta)$, with $c > 1$. (i) Find the mean and variance for this distribution. (ii) Generate such a dataset, say $y_1, \dots, y_n$ with $n = 250$, $\theta = 10$, $c\theta = 20$, $p = 0.10$. Taking first $c = 2$ known, estimate the parameters (p, θ) , along with confidence intervals. (iii) Using the same data, now with c unknown, estimate all three parameters, with confidence intervals. Briefly investigate how much is earned in precision for estimating (p, θ) when c is known compared to it being unknown.

(b) Suppose data Y_i^0 are generated according to some normal (ξ, σ^2) , but that we are only able to see those falling inside some observation window $[a, b]$. Write down the log-likelihood function, for the observed data Y_i . To see how it works, generate say $n_0 = 1000$ points from the standard normal, keep the n amongst these that fall inside $[1.0, 3.0]$, and estimate the two parameters from these. Find also approximate confidence intervals for the two parameters.

(c) For two distributions with densities f_1, f_2 and c.d.f.s F_1, F_2 , consider

$$f(x, y, a) = f_1(x)f_2(y)[1 + a\{F_1(x) - \tfrac{1}{2}\}\{F_2(y) - \tfrac{1}{2}\}].$$

What is the range of the dependence parameter a ? Show that the marginal distributions are f_1, f_2 . Generate first binormal data with $N(0, 1)$ marginals, and estimate a , with confidence interval for a . Then generate another binormal dataset, and estimate the five parameters $(\xi_1, \xi_2, \sigma_1, \sigma_2, a)$, along with confidence intervals.

(d) Tired of normality? Then generalise it. With $\text{Be}_{a,b}(x)$ the c.d.f. for the Beta(a, b) distribution, consider the four-parameter model with c.d.f. $F(y) = \text{Be}_{a,b}(\Phi((y - \xi)/\sigma))$. Write up an expression for the log-likelihood function. Generate first $n = 500$ datapoints from the standard normal, and check if the ML method succeeds in coming close to the true parameters $(0, 1, 1, 1)$. Argue that you here have a test for normality, and spell out the ingredients. Then simulate another dataset, from $(a, b) = (1.23, 2.34)$, compute ML estimates and their estimated standard deviations.

Ex. 5.31 *Independence-likelihood might be ok with dependent data.* Consider a setup with data $Y_1, \dots, Y_n$, with each data point drawn from the same distribution, but perhaps with dependencies. In Story #62, for example, we deal with $m = 90$ p-values p_t computed for different subtables, from a common bigger table. These may still be worked with using ‘the wrong likelihood’, ignoring dependence.

(a) Suppose the Y_i above are each from a density g , but perhaps with dependencies. With $h(y)$ any bounded function, show that $\bar{h}_n = n^{-1} \sum_{i=1}^n h(Y_i)$ has mean $h_0 = \int hg \, dy$, and that $\bar{h}_n \rightarrow_{\text{pr}} h_0$ as long as $\text{Var } \bar{h}_n \rightarrow 0$. The point is that this often holds even with dependence. Explain that $G_n(b) - G_n(a)$, with the empirical c.d.f., tends to cell probability $G(b) - G(a)$ under such conditions. Hence a histogram of the data is perfectly informative about the underlying distribution, even if variances and covariances may be more complicated than under independence.

so a LLN with dependence

(b) For the sake of the argument, assume there really are dependencies for the data. With a parametric model $f(y, \theta)$, explain that $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i, \theta)$ is not the correct log-likelihood; the correct one would need specification of the dependency and specifically a joint model for all the data. Nevertheless, explain that $n^{-1}\ell_n(\theta) \rightarrow_{\text{pr}} A(\theta) = \int g \log f_\theta \, dy$, under mild conditions, and that the maximiser $\hat{\theta}$, though formally not the ML (but rather the maximiser of a wrong likelihood), will aim at the usual least false $\text{KL}(g, f_\theta)$ minimiser. Explain also, however, that the ML theory developed above, with variances, limit distributions, etc., does not apply as such.

Convex criteria, log-concave likelihoods, differentiability in q.m.

Ex. 5.32 *Minimisers of convex processes, I.* We have seen useful constructions, methods, and results for ML estimators in Ex. 5.10, 5.11, 5.13. There are issues worth refining and generalising, however. The regularity conditions required for the Taylor expansion based

arguments to go fully through are a bit cumbersome, and there are important constructions where the $E_G \log f(Y, \theta)$ is not smooth. Here we give the basics for how matters simplify, with weaker conditions, if the log-likelihood is concave, or more generally when the estimator is defined as the minimiser of a convex criterion function.

(a) From pointwise to uniform: Suppose $A_n(s)$ is a sequence of convex random functions defined on an open convex set $\mathcal{S}$ of $\mathbb{R}^p$, which converges in probability to some $A(s)$, for each $s \in \mathcal{S}$. Show that the convergence is automatically uniform; $\max_{s \in \mathcal{S}} |A_n(s) - A(s)| \rightarrow_{\text{pr}} 0$.

(b) Nearness of argmins: Suppose $A_n(s)$ is convex and is approximated by $B_n(s)$. Let α_n and β_n be the argmins of A_n and B_n . Then there is a probabilistic bound on how far these minimisers can be from each other: show that

$$\Pr(\|\alpha_n - \beta_n\| \geq \delta) \leq \Pr\{\Delta_n(\delta) \geq \frac{1}{2}h_n(\delta)\},$$

in which

$$\Delta_n(\delta) = \sup_{\|s - \beta_n\| \leq \delta} |A_n(s) - B_n(s)| \quad \text{and} \quad h_n(\delta) = \inf_{\|s - \beta_n\| = \delta} B_n(s) - B_n(\beta_n).$$

(c) Basic corollary: Suppose $A_n(s)$ is convex and can be represented as $\frac{1}{2}s^t J s + U_n^t s + C_n + r_n(s)$, where J is symmetric and positive definite, U_n is stochastically bounded, C_n is arbitrary, and $r_n(s) \rightarrow_{\text{pr}} 0$ for each s . For the approximation $B_n(s) = \frac{1}{2}s^t J s + U_n^t s + C_n$, show that $\beta_n = -J^{-1}U_n$ is its argmin. Then demonstrate that their minimisers as well as their minima are close. Specifically, show (i) that $\alpha_n - \beta_n \rightarrow_{\text{pr}} 0$; and (ii) that $A_{n,\min} - B_{n,\min} \rightarrow_{\text{pr}} 0$.

(d) Show that if in addition $U_n \rightarrow_d U$, then $\alpha_n \rightarrow_d -J^{-1}U$, and that $B_{n,\min} - C_n$ as well as $A_{n,\min} - C_n$ tend to $-\frac{1}{2}U^t J^{-1}U$. These two statements are what we worked hard for in Ex. 5.10–5.11, now obtained in a simpler fashion and with weaker smoothness assumptions, though bought with the extra convexity condition.

(e) Prove the following modest but useful generalisation of the above: the statements continue to hold if a random matrix J_n replaces V , provided $J_n \rightarrow_{\text{pr}} J$.

Ex. 5.33 *Minimisers of convex processes, II.* The framework worked with now is as in Ex. 5.10–5.11, with $Y_1, \dots, Y_n$ being i.i.d. from some G , a possibly multidimensional parameter θ_0 defined as the minimiser of $H(\theta) = \int h(y, \theta) dG(y)$, with estimator $\hat{\theta}$ the minimiser of the distance function $H_n(\theta) = \int h(y, \theta) dG_n(y) = n^{-1} \sum_{i=1}^n h(Y_i, \theta)$. Here we put in one more condition, however, that $h(y, \theta)$ is convex in θ . The point is that this both simplifies various technical arguments, via methods of Ex. 5.32, and allows for nonsmooth distance functions.

(a) With $h(y, \mu) = |y - \mu|$, show that $\mu_0 = \text{med}(G)$, the median, with $\hat{\mu} = M_n$, the sample median. More generally, for some $q \in (0, 1)$, consider

$$h_q(y, \mu) = q(y - \mu)_+ + (1 - q)(\mu - y)_+ = \begin{cases} q(y - \mu) & \text{if } y \geq \mu, \\ (1 - q)(\mu - y) & \text{if } y \leq \mu. \end{cases}$$

Show that $\mu_0 = G^{-1}(q)$, the q quantile.

(b) For an $\alpha \geq 1$, consider the parameter θ_0 being the minimiser of $E_G |Y - \theta|^\alpha$. Show that the distance function indeed is convex, and that the special cases $\alpha = 1$ and $\alpha = 2$ correspond to the median and the mean, respectively.

(c) We now work through regularity conditions ensuring control over the behaviour of such estimators. Part of the point is that we avoid needing smooth derivatives in θ . Suppose that

$$h(y, \theta_0 + \varepsilon) - h(y, \theta_0) = D(y)^t \varepsilon + R(y, \varepsilon),$$

for a $D(y)$ with mean zero under G , and that

$$\mathbb{E} \{h(Y, \theta_0 + \varepsilon) - h(Y, \theta_0)\} = \mathbb{E} R(Y, \varepsilon) = \frac{1}{2} \varepsilon^t J \varepsilon + o(\|\varepsilon\|^2) \quad \text{as } \varepsilon \rightarrow 0$$

for a positive definite J . Assume furthermore that the variance matrix $K = \text{Var}_G D(Y)$ is finite and that $\text{Var} R(Y, \varepsilon) = o(\|\varepsilon\|^2)$. Show that $\sqrt{n}(\hat{\theta} - \theta_0) = -J^{-1}(1/\sqrt{n}) \sum_{i=1}^n D(Y_i) + o_{\text{pr}}(1)$. In particular, it tends to $N_p(0, J^{-1} K J^{-1})$. Show also that

$$W_n(\theta_0) = 2n\{H_n(\theta_0) - H_n(\hat{\theta})\} \rightarrow_d U^t J^{-1} U,$$

and explain how this may be used to find a confidence region for θ_0 .

(d) The median: Suppose $Y_1, \dots, Y_n$ are i.i.d. from a distribution G with a density g positive at the median μ . For the median distance function $|y - \mu|$, show that

$$|y - (\mu + \varepsilon)| - |y - \mu| = D(y)\varepsilon + R(y, \varepsilon),$$

with $D(y) = I(y \leq \mu) - I(y > \mu)$ and

$$R(y, \varepsilon) = \begin{cases} 2\{\varepsilon - (y - \mu)\} I(\mu \leq y \leq \mu + \varepsilon) & \text{if } \varepsilon > 0, \\ 2\{(y - \mu) - \varepsilon\} I(\mu + \varepsilon \leq y \leq \mu) & \text{if } \varepsilon < 0, \end{cases}$$

with $R(y, 0) = 0$. Verify from this that $\mathbb{E} R(Y, \varepsilon) = g(\mu)\varepsilon^2 + o(\varepsilon^2)$ and $\mathbb{E} R(Y, \varepsilon)^2 = (4/3)g(\mu)|\varepsilon|^3 + o(|\varepsilon|^3)$. Deduce that $\sqrt{n}(M_n - \mu) \rightarrow_d N(0, 1/\{4g(\mu)^2\})$.

(e) Generalise to the case of $\mu_q = G^{-1}(q)$, with $Q_{n,q} = G_n^{-1}(q)$ the empirical q quantile. Show in fact that

$$Z_n(q) = \sqrt{n}(Q_{n,q} - \mu_q) = -g(\mu_q)^{-1} \sqrt{n}\{G_n(\mu_q) - q\} + \varepsilon_n(q),$$

where $\varepsilon_n(q) \rightarrow_{\text{pr}} 0$ for each q . Derive from this that the limit is a $N(0, q(1-q)/g(\mu_q)^2)$. We saw this in Ex. 3.21, but here the technique is distinctly different, and we find an interesting representation in terms of the empirical $G_n(\mu_q)$.

(f) For $\alpha \geq 1$, let ξ_α be the minimiser of $\mathbb{E}|Y_i - \xi|^\alpha$, with estimator $M_{n,\alpha}$ minimising $\sum_{i=1}^n |Y_i - \xi|^\alpha$. Show that

$$\sqrt{n}(M_{n,\alpha} - \xi_\alpha) \rightarrow_d N(0, \tau_\alpha^2) \quad \text{with } \tau_\alpha^2 = \frac{\mathbb{E}|Y - \xi_\alpha|^{2(\alpha-1)}}{\{(\alpha-1)\mathbb{E}|Y - \xi_\alpha|^{\alpha-2}\}^2}.$$

Explain that if the distribution is symmetric, then the ξ_α is the same for each α . Compute and display τ_α , for $\alpha \in [1, 2]$, i.e. the range from median to mean, from Laplace to Gauß.

Ex. 5.34 *Maximum likelihood asymptotics with log-concave likelihood.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from a density g , modelled as $f(y, \theta)$, where $\log f(y, \theta)$ is concave in θ for each y ; the $\theta = (\theta_1, \dots, \theta_p)$ is allowed to be multidimensional. Let $\ell_n(\theta)$ be the log-likelihood, with ML estimator $\hat{\theta}$. From efforts of Ex. 5.10 we know $\hat{\theta}$ is consistent for the least false θ_0 , the KL minimiser, assumed here to be an inner point in the parameter space. log-concave density

(a) Show first that ℓ_n is concave. To start with mild conditions, assume merely that

$$\log f(y, \theta_0 + h) - \log f(y, \theta_0) = h^t D(y) + R(y, h), \quad (5.11)$$

for a $D(y)$ with mean zero, and that $E R(Y, h) = \frac{1}{2} h^t J h + o(\|h\|^2)$ as $h \rightarrow 0$ for some positive definite J , along with $\text{Var } R(Y, \varepsilon) = o(\|h\|^2)$. Show via the convexity driven methods of Ex. 5.32 and 5.33 that $\sqrt{n}(\hat{\theta} - \theta_0) = J^{-1} U_n + o_{\text{pr}}(1)$, where $U_n = (1/\sqrt{n}) \sum_{i=1}^n u(Y_i, \theta_0)$. Here $U_n \rightarrow_d U \sim N_p(0, K)$, with $K = \text{Var}_g u(Y, \theta_0)$ assumed finite. Deduce the two fundamental results

$$\begin{aligned} Z_n &= \sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N_p(0, J^{-1} K J^{-1}), \\ W_n &= 2\{\ell_{n, \max} - \ell_n(\theta_0)\} \rightarrow_d U^t J^{-1} U, \end{aligned} \quad (5.12)$$

with simplifications under model conditions. Show also that consistency of the ML estimator for θ_0 is a simple consequence of the first statement. We have seen such results before, as in (5.6); the point is that they have now been derived more simply and under weaker conditions, as long as there is log-density concavity.

(b) Work through the details for the case of the Laplace distribution with $f(y, \theta) = \frac{1}{2} \exp(-|y - \theta|)$. The point is that with log-concavity, we reach the required results of (5.12) even without needing full smoothness in the parameter; here the score function is not defined for all values, etc. Usually, though, the $D(y)$ is the score function $u(y, \theta_0) = \partial \log f(y, \theta_0) / \partial \theta$ and $J = J(\theta_0)$ is the variance matrix of this score function, i.e. the Fisher information matrix, evaluated at the true position in the parameter space. Explain how you can deal also with the density $f(y, \theta) = (1/k) \exp\{-|y - \theta|^{1.5}\}$.

(c) Consider

$$f(y, \theta) = (1/k) \exp\{-(y - \theta) \arctan(y - \theta)\} \{1 + (y - \theta)^2\}^{1/2} \quad \text{for } y \in \mathbb{R}.$$

Show that it has finite integral, and give an expression for the integration constant k making it a density. Show that it is log-concave, and that the ML estimator $\hat{\theta}$, based on an i.i.d. sample $y_1, \dots, y_n$, is the unique solution to the equation $\sum_{i=1}^n \arctan(y_i - \theta) = 0$. Find the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta)$.

Ex. 5.35 Quadratic expansions of the log-likelihood. Classical conditions. In several of the above exercises the key to proving asymptotic normality of the maximum likelihood estimator has been the quadratic expansion

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = s^t U_n - \frac{1}{2} s^t J s + r_n(s), \quad (5.13)$$

where $r_n(s) \rightarrow_{\text{pr}} 0$ for each s . See, for example, Ex. 5.10. In this exercise we look into various sets of conditions ensuring that such an expansion of the log-likelihood ratio holds, both inside and outside of model conditions. The setting is an i.i.d. sample $Y_1, \dots, Y_n$ from some distribution G , that we model parametrically by $f(y, \theta)$. As above, $u(y, \theta) = \partial \log f(y, \theta) / \partial \theta$ and $i(y, \theta) = \partial^2 \log f(y, \theta) / \partial \theta \partial \theta^t$, while $J = -E_G i(Y, \theta_0)$ and $K = \text{Var}_G u(Y, \theta_0)$, where θ_0 denotes the least-false parameter value, assumed throughout to be an inner point of its parameter space.

(a) Assume that $\theta \mapsto \log f(y, \theta)$ is two times continuously differentiable for every y ; that J is finite; that every element of $i(y, \theta)$ is bounded by an integrable function $k(y)$, at least in a neighbourhood of θ_0 ; and that

$$\sup_{\theta \in B_\varepsilon(\theta_0)} \left| n^{-1} \sum_{j=1}^n i(Y_j, \theta) - E_G i(Y, \theta) \right| \rightarrow_{\text{pr}} 0,$$

where $B_\varepsilon(\theta_0)$ is an ε -ball around the least false parameter value θ_0 . Show that, under these conditions, (5.13) holds.

(b) Hold onto the conditions you need from (a), and assume, in addition, that every element of $u(y, \theta)$ is dominated by an integrable function $k_1(y)$, at least in a neighbourhood of θ_0 . Show that $E_G u(Y, \theta_0) = 0$.

Ex. 5.36 *Differentiability in quadratic mean.* The quadratic expansion in (5.13) can be derived under conditions that only involve a certain first derivative of $f(y, \theta)$, called differentiability in quadratic mean, abbreviated DQM. Let $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$ be a family of distributions, all with densities $f(y, \theta)$ with respect to some dominating measure μ , and let $Y_1, \dots, Y_n$ be i.i.d. from P_{θ_0} , with θ_0 in the interior of its parameter space $\Theta \subset \mathbb{R}^p$. We are thus working under model conditions in this exercise. The model $\mathcal{P}$ is differentiable in quadratic mean at θ_0 if there exists a (vector-valued) function $u(y, \theta)$ such that

$$\int (\sqrt{f(y, \theta_0 + h)} - \sqrt{f(y, \theta_0)} - \frac{1}{2} h^t u(y, \theta_0) \sqrt{f(y, \theta_0)})^2 d\mu(y) = o(\|h\|^2), \quad (5.14)$$

differentiability
in quadratic
mean

as $h \rightarrow 0$. In the following exercises we will see that the differentiability in quadratic mean condition entails that the log-likelihood ratio $A_n(h) = \ell_n(\theta_0 + h/\sqrt{n}) - \ell_n(\theta_0)$, worked with in several of the exercises above, has a quadratic expansion of the form given in (5.13). Since it is $A_n(h)$ we want to expand, it will be convenient to introduce the abbreviations $f(y) = f(y, \theta_0)$, $f_n(y) = f(y, \theta_0 + h/\sqrt{n})$, and $u(y) = u(y, \theta_0)$. The DQM condition above then reads $\int (\sqrt{f_n} - \sqrt{f} - \frac{1}{2} n^{-1/2} h^t u \sqrt{f})^2 d\mu = o(1/n)$ as $n \rightarrow \infty$.

(a) Argue that the DQM condition entails that the $\sqrt{n}(\sqrt{f_n} - \sqrt{f})$ converges to $\frac{1}{2} h^t u \sqrt{f}$ in $L_2(\mu)$ as n tends to infinity, and use this to show that $E u(Y, \theta_0) = 0$. Show also that $J = E u(Y, \theta_0) u(Y, \theta_0)^t$ is finite.

(b) We start by manipulating the log-likelihood ratio process $A_n(h)$. Define the random variables

$$W_{n,i} = 2(\sqrt{f_n(Y_i)/f(Y_i)} - 1), \quad i = 1, \dots, n, \quad n \geq 1,$$

these being independent replicates of $W_n = 2(\sqrt{f_n(Y)}/\sqrt{f(Y)} - 1)$. Verify that we may write $A_n(h) = 2 \sum_{i=1}^n \log(1 + \frac{1}{2} W_{n,i})$, and, by a Taylor expansion with the remainder on mean value form, establish that

$$A_n(h) = \sum_{i=1}^n W_{n,i} - \frac{1}{4} \sum_{i=1}^n W_{n,i}^2 + \frac{1}{4} \sum_{i=1}^n W_{n,i}^2 R(\frac{1}{2} W_{n,i}),$$

where $R(x) = \frac{1}{2} \{1 - (1 + \tilde{x})^{-2}\}$ for some $\tilde{x}$ between 0 and x .

(c) Let's establish a few useful facts. Show that $n E W_n = -\frac{1}{4} h^t J h$; establish the identity $E W_n^2 = -4 E W_n$; and show that $E (W_n - n^{-1/2} h^t u(Y, \theta_0))^2 \leq 4 D(h/\sqrt{n})$, where $D(h/\sqrt{n}) = \int (\sqrt{f_n} - \sqrt{f} - \frac{1}{2} n^{-1/2} h^t u \sqrt{f})^2 d\mu$. Deduce that

$$E |W_n^2 - n^{-1} (h^t u(Y, \theta_0))^2| \leq n^{-1/2} 4 \{D(h/\sqrt{n}) h^t J h\}^{1/2}.$$

(d) Define the mean zero random variables $a_{n,i} = W_{n,i} - n^{-1/2} h^t u(Y_i, \theta_0) - E W_{n,i}$, and verify, using the results from (c), that $\sum_{i=1}^n W_{n,i} = h^t U_n - \frac{1}{4} h^t K h + \sum_{i=1}^n a_{n,i}$, and that, indeed, $\sum_{i=1}^n a_{n,i} = o_{\text{pr}}(1)$.

(e) Define the random variables $b_{n,i} = W_{n,i}^2 - \{n^{-1/2} h^t u(Y_i, \theta_0)\}^2$, and also the mean zero random variables $c_{n,i} = h^t K h - (h^t u(Y_i, \theta_0))^2$. Use the results from (c), along with

the law of large numbers, to show that

$$\sum_{i=1}^n W_{n,i}^2 = h^t Jh - \sum_{i=1}^n b_{n,i} + n^{-1} \sum_{i=1}^n c_{n,i} = h^t Jh + o_{\text{pr}}(1).$$

Verify that at this point we have $A_n(h) = h^t U_n - \frac{1}{2} h^t Jh + \sum_{i=1}^n W_{n,i}^2 R(\frac{1}{2} W_{n,i}) + o_{\text{pr}}(1)$. It thus remains to show that $\sum_{i=1}^n W_{n,i}^2 R(\frac{1}{2} W_{n,i}) = o_{\text{pr}}(1)$.

(f) Add and subtract $n^{-1}(h^t u(Y, \theta_0))^2$ from W_n^2 to show that

$$\Pr(W_n^2 \geq 2\varepsilon^2) \leq \frac{\|h\|}{n\varepsilon^2} \mathbb{E} \|u(Y, \theta_0)\|^2 I\{\|u(Y, \theta_0)\| \geq \frac{\sqrt{n}\varepsilon}{\|h\|}\} + \frac{4\{D(h/\sqrt{n})h^t Jh\}^{1/2}}{\varepsilon^2 \sqrt{n}},$$

for each $\varepsilon > 0$; and deduce from this expression that $\max_{1 \leq i \leq n} |W_{n,i}| \rightarrow_{\text{pr}} 0$ as $n \rightarrow \infty$. Finally, show that $\max_{1 \leq i \leq n} |W_{n,i}| = o_{\text{pr}}(1)$ implies that $\sum_{i=1}^n W_{n,i}^2 R(\frac{1}{2} W_{n,i}) = o_{\text{pr}}(1)$.

(g) [xx an exercise with conditions ensuring that a family of distributions is dqm xx]

Ex. 5.37 *Quadratic expansions outside of model conditions.* The DQM condition in (5.14) works well in deriving the expansion (5.13) when we are working under model conditions. Outside of model conditions, however, we need something else. The classical conditions explored in Ex. 5.35 will do, but we would like something weaker. Let $\text{KL}(g, f_\theta) = \int \log(g/f_\theta) g d\mu$ be the Kullback–Leibler divergence, and let θ_0 be the least false parameter value, i.e., the minimiser of $\text{KL}(g, f_\theta)$ viewed as a function of θ . The setting is an i.i.d. sample $Y_1, \dots, Y_n$ from a distribution G with density g that we modelled by the parametric density f_θ with $\theta \in \Theta \subset \mathbb{R}^p$. Consider the following conditions: (i) θ_0 resides in the interior of its parameter space; (ii) $\theta \mapsto \text{KL}(g, f_\theta)$ admits a second-order Taylor expansion around θ_0 ; and (iii) there is a vector valued function u_θ such that

$$\int (\log f_{\theta_0+h} - \log f_{\theta_0} - h^t u_{\theta_0})^2 g d\mu = o(\|h\|^2), \quad \text{as } h \rightarrow 0.$$

Write KL' and KL'' for the first- and second derivative of $\theta \mapsto \text{KL}(g, f_\theta)$, evaluated in the least false parameter value θ_0 .

(a) Deduce from Condition (iii) that $\lim_{t \rightarrow 0} t^{-1} \mathbb{E}_G \log f_{\theta_0+th}/f_{\theta_0} = h^t \mathbb{E}_G u_{\theta_0}$; while from Conditions (i)&(ii) we have $\lim_{t \rightarrow 0} t^{-1} \mathbb{E}_G \log f_{\theta_0+th}/f_{\theta_0} = -h^t \text{KL}' = 0$, together implying that u_{θ_0} is mean zero random variable.

(b) Use the result from (a) along with the three conditions to establish that as $n \rightarrow \infty$ we have both

$$\mathbb{E}_G \sum_{i=1}^n \left(\log \frac{f_{\theta_0+h/\sqrt{n}}}{f_{\theta_0}} - \frac{h^t}{\sqrt{n}} u_{\theta_0} \right) = -\frac{1}{2} \text{KL}'' h + o(\|h\|^2)$$

and $\text{Var}_G \sum_{i=1}^n (\log f_{\theta_0+h/\sqrt{n}}/f_{\theta_0} - n^{-1/2} h^t u_{\theta_0}) = o(1)$. Deduce from this that

$$A_n(h) = \frac{h^t}{\sqrt{n}} \sum_{i=1}^n u_{\theta_0}(Y_i) - \frac{1}{2} \text{KL}'' h + o(1).$$

(c) Recall that $i(y, \theta)$ is the second partial derivative of $\log f_\theta$. Show that under suitable conditions (i.e., those allowing use to pass a second derivative under the integral sign, see Ex. 5.35), we get that $\text{KL}'' = -\mathbb{E}_G i(Y, \theta_0)$, and the expression for A_n above coincides with that given in (5.13).

(d) The Laplace model with density $f_\theta = \frac{1}{2} \exp(-|x - \theta|)$ is perhaps the canonical example of a model that does not satisfy the classical conditions of Ex. 5.35. Verify that the classical conditions fail for the Laplace model, but that it does, however, satisfy the three conditions of the present exercise.

(e) Let $Y_1, \dots, Y_n$ be i.i.d. from a density g on the real line, that is such that $g(\theta_0) > 0$, and suppose that these are modelled by the Laplace model. Combine the theory of the present exercise with the concavity arguments explored in Ex. 5.34 to show that with $\hat{\theta}_n$ the maximum likelihood estimator, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, (4g(\theta_0)^{-1})$.

Ex. 5.38 *The exponential family class and ML.* The ML and log-likelihood machinery works particularly well for the exponential family class of models, see Ex. 1.57, 4.21. This is due the log-linear structure for parameters and sufficient statistics, and also to the consequent log-concavity of densities. Consider i.i.d. data $Y_1, \dots, Y_n$ from a density g , modelled with the generic parametric $f(y, \theta) = \exp\{\theta^t T(y) - k(\theta)\} h(y)$, with $\theta^t T(y) = \theta_1 T_1(y) + \dots + \theta_p T_p(y)$ involving basic data functions $T_1, \dots, T_p$.

(a) Show that the score function is $u(y, \theta) = T(y) - \xi(\theta)$, with $\xi(\theta) = \partial k(\theta) / \partial \theta$. Also, the information function becomes $i(y, \theta) = -J(\theta)$, with $J(\theta) = \partial^2 k(\theta) / \partial \theta \partial \theta^t$.

(b) Under model conditions, show that $E_\theta T(Y) = \xi(\theta)$ and $\text{Var}_\theta T(Y) = J(\theta)$. Outside model conditions, with g not belonging to the f_θ , assume merely that $T(Y)$ has true mean ξ_0 and true variance K . Show that the least false parameter θ_0 minimising $\text{KL}(g, f(\cdot, \theta))$ is characterised by $\xi_0 = \xi(\theta_0)$. This may also be written $\theta_0 = M(\xi_0)$, with M the inverse map. Write also $J = J(\theta_0)$, the Fisher information matrix computed at the least false parameter position.

(c) Show that the log-likelihood function becomes $\ell_n(\theta) = n\{\theta^t \bar{T} - k(\theta)\}$, in terms of sample averages $\bar{T} = (\bar{T}_1, \dots, \bar{T}_p)^t$, and that it is concave. Deduce that the ML estimator has $\xi(\hat{\theta}) = \bar{T}$, of $\hat{\theta} = M(\bar{T})$.

(d) We have $U_n = \sqrt{n}(\bar{T} - \xi_0) \rightarrow_d U \sim N_p(0, K)$, by the CLT. For the basic log-likelihood function process, show that

$$\begin{aligned} A_n(s) &= \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = \sqrt{n} \bar{T}^t s - n\{k(\theta_0 + s/\sqrt{n}) - k(\theta_0)\} \\ &= U_n^t s - \frac{1}{2} s^t J(\theta_0) s + o(\|s\|^3 / \sqrt{n}) \rightarrow_d A(s) = U^t s - \frac{1}{2} s^t J s. \end{aligned}$$

This is as with Ex. 5.10-5.11, but now established rather simply and directly, with a minimum of regularity conditions. Deduce also that

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N_p(0, J^{-1} K J^{-1}), \quad 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \rightarrow_d W = U^t J^{-1} U.$$

Under model conditions, the sandwich matrix is $J(\theta_0)^{-1}$ and $W \sim \chi_p^2$. Also, the Wilks theorems of Ex. 5.10 hold, for focus parameters and for testing submodels.

(e) From $\hat{\theta}_j = M_j(\bar{T})$, for components $M_1, \dots, M_p$ of the mapping $\theta_0 = M(\xi_0)$, write

$$\hat{\theta}_j = M_j(\xi_0) + M'_j(\xi_0)^t (\bar{T} - \xi_0) + \frac{1}{2} (\bar{T} - \xi_0)^t M''_j(\xi_0) (\bar{T} - \xi_0) + O_{\text{pr}}(1/n^{3/2}).$$

Show that this implies $E \hat{\theta}_j = \theta_{0,j} + \frac{1}{2} c_j / n + o(1/n)$, with $c_j = \text{Tr}(M''_j(\xi_0) K)$.

(f) (xx a simple example here. xx)

Ex. 5.39 *ML asymptotics under model conditions: applications.* (xx cleaning required, and more of the directly useful up front; we estimate parameters and have confidence, almost automatically, via normality and delta method. xx) Results reached in Ex. 5.10

are central in applied statistics. The versatile ML machinery allows the statistician to construct good estimators for even complicated functions of parameters in new models, and to supplement these estimators with confidence intervals, tests, etc. We will also see that the general ML asymptotics results may be used to verify what we already knew, so to speak, regarding estimators in the more familiar models. (xx check all this to make sure that we don't become repetitive. xx)

(a) Let Y be binomial (n, p) . Even before you find a formula for the ML estimator for p , show that $\sqrt{n}(\hat{p} - p) \rightarrow N(0, p(1 - p))$. By all means, show also that $\hat{p} = Y/n$.

(b) In a similar vein, study the classic case of $Y_1, \dots, Y_n$ being i.i.d. from the normal (ξ, σ^2) . Using the Fisher information matrix found in Ex. 5.6, show that the ML estimators $\hat{\xi}$ and $\hat{\sigma}$ must be independent, in the limit, with approximate distributions $N(\xi, \sigma^2/n)$ and $N(\sigma, \frac{1}{2}\sigma^2)$. Remarkably, these results follow from the ML apparatus even without or before knowing any formulae for the estimators, and without or before knowing any finite-sample theory for these. As we know there is *exact* independence here, see Ex. 1.49, and the distribution for $\hat{\xi}$ is exactly correct, for each n .

(c) (xx something with $\text{Gam}(a, b)$, and approximate distribution for $\hat{\mu}$, estimator for the median $\mu = \mu(a, b)$. illustrate also Wilks. which can be used without explicit formulae. xx)

(d) Consider the Weibull with c.d.f. $F(t) = 1 - \exp\{-(t/a)^b\}$. From the log-density, find the score functions, here also in terms of $v = (t/a)^b$:

$$\begin{aligned}\partial/\partial a &= -b(t/a)^b/at/a^{b+1} - b/a = (b/a)(v - 1), \\ \partial/\partial b &= -(t/a)^b \log(t/a) + 1/b + \log(t/a) = (1/b)(1 + \log v - v \log v).\end{aligned}$$

Fisher matrix
for Weibull

Explain that leads to the Fisher matrix and inverse,

$$J = \begin{pmatrix} b^2/a^2 & c/a \\ c/a & d/b^2 \end{pmatrix} \quad \text{with } J^{-1} = \frac{a^2}{d - c^2} \begin{pmatrix} d/b^2 & -c/a \\ -c/a & b^2/a^2 \end{pmatrix}.$$

The constants $c = \text{cov}(v - 1, 1 + \log v - v \log v) = -0.428$ and $d = \text{Var}(\log v - v \log v) = 1.824$ are found easily enough via numerical integration, but may also be found mathematically, via derivatives of the gamma function: $c = -(1 - \gamma_e)$, $d = \psi'(2) + \psi(2)^2 + 1 = \psi'(1) + (1 - \gamma_e)^2$; also, $d - c^2$ is Euler's miraculous $\pi^2/6$. Write down the limit distributions for $\sqrt{n}(\hat{a} - a)$ and $\sqrt{n}(\hat{b} - b)$, for the ML estimators, and also that of $\sqrt{n}(\hat{m} - m)$, with m the median.

(e) Consider random i.i.d. pairs (X_i, Y_i) from the standardised binormal distribution with zero means, unit variances, and correlation ρ . Set up the log-likelihood function $\ell_n(\rho)$, show that the Fisher information becomes $J(\rho) = (1 + \rho^2)/(1 - \rho^2)^2$, and find the limiting distribution of the ML estimator. How much better is the ML estimator compared to the usual empirical correlation coefficient R_n ? Show that the ratio of variances, for the 1-parameter divided by that of the 5-parameter model, converges to $1/(1 + \rho^2)$.

Ex. 5.40 *Examples of agnostic ML operations.* It is useful to go through a list of special cases, to see how the agnostic ML theory pans out in practice. Note that convergence to the normal $N_p(0, J^{-1}KJ^{-1})$ takes place in general, model after model after model (including those you might invent next week), without any need for working with explicit formulae for the ML estimators etc.

(a) For the exponential model $\theta \exp(-\theta y)$, show that the score function is $u(y, \theta) = 1/\theta - y$, that its least false parameter value is $\theta_0 = 1/\xi_0$, in terms of the true mean $\xi_0 = EY$. Show that $\sqrt{n}(\hat{\theta} - \theta_0)$ has limit distribution $N(0, \sigma_0^2 \theta_0^4)$, where σ_0^2 is the true variance. Show that this generalises the ‘usual result’ derived under model conditions.

(b) Then do the normal: assume data follow some density g , and the normal $N(\xi, \sigma^2)$ model is used. We already know that the least false parameters are ξ_0 and σ_0 , the true mean and standard deviation (i.e. even if g is far from the normal). Assume that the fourth moment is finite, so that skew = $E Z^3$ and kurt = $E Z^4 - 3$ are finite, with $Z = (Y - EY)/\text{sd}(Y) = (Y - \xi_0)/\sigma_0$. Working with the score function, and the second order derivatives, show that

$$J = \frac{1}{\sigma_0^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad K = \frac{1}{\sigma_0^2} \begin{pmatrix} 1 & \gamma_3 \\ \gamma_3 & 2 + \gamma_4 \end{pmatrix}.$$

(c) For the ML estimators $\hat{\xi}$ and $\hat{\sigma}$, show from this that

$$\begin{pmatrix} \sqrt{n}(\hat{\xi} - \xi) \\ \sqrt{n}(\hat{\sigma} - \sigma) \end{pmatrix} \rightarrow_d N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \frac{1}{2}\gamma_3 \\ \frac{1}{2}\gamma_3 & \frac{1}{2} + \frac{1}{4}\gamma_4 \end{pmatrix} \right).$$

Note that this is a ‘rediscovery’ of what we found in 2.54, but here we managed to find the limit distribution fully without knowing (or caring) about the exact expressions for the ML estimators.

(d) Study the model for data on the unit interval with c.d.f. y^θ . With i.i.d. data $Y_1, \dots, Y_n$, find a formula for the ML estimator, and describe its limit distribution.

Ex. 5.41 *An average power optimality property.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from a smooth parametric model $f(y, \theta)$, where we need to test $\theta = \theta_0$, a given value, against $\theta \neq \theta_0$. In general there is no uniformly most powerful test. We have seen in Ex. 4.17, however, that there is a well-defined test maximising the weighted average power $\bar{\pi}_n = \int \pi_n(\theta) dw(\theta)$, with $\pi_n(\theta)$ the power at position θ and $dw(\theta)$ a given probability measure on the alternative region, here $\theta \neq \theta_0$. This optimal strategy is to use the Neyman–Pearson Lemma for the marginal density $\bar{f}(y_1, \dots, y_n) = \int \prod_{i=1}^n f(y_i, \theta) dw(\theta)$.

(a) Let $\ell_n(\theta)$ be the log-likelihood function, with $\hat{\theta}$ the ML estimator, and write also f_0 for the model at the null value θ_0 . Show that the Neyman–Pearson ratio can be expressed as

$$\begin{aligned} R_n &= \frac{\bar{f}(y_1, \dots, y_n)}{f_0(y_1, \dots, y_n)} = \int \exp\{\ell_n(\theta) - \ell_n(\theta_0)\} dw(\theta) \\ &= \exp\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} \int \exp\{\ell_n(\theta) - \ell_n(\hat{\theta})\} dw(\theta). \end{aligned}$$

(b) For θ close to $\hat{\theta}$, use Taylor expansion to get

$$\ell_n(\theta) - \ell_n(\hat{\theta}) \doteq -\frac{1}{2}n(\theta - \hat{\theta})^t J_n(\theta - \hat{\theta}),$$

with $J_n = -n^{-1} \partial^2 \ell_n(\hat{\theta}) / \partial \theta \partial \theta^t$ the normalised Hessian matrix at the max point.

(c) The optimal test consists in rejecting when R_n is above its null distribution threshold. Show that the above leads to

$$\begin{aligned} R_n &\doteq \exp(\tfrac{1}{2} D_n) (2\pi)^{p/2} |n J_n|^{1/2} w(\hat{\theta}), \\ 2 \log R_n &\doteq D_n - p \log n + \log |J_n| + p \log(2\pi) + 2 \log w(\hat{\theta}). \end{aligned}$$

Here $D_n = 2\{\ell_{n,\max} - \ell_n(\theta_0)\}$ is the Wilks or log-likelihood-ratio test statistic, with its χ_p^2 limiting null distribution.

(d) Conclude that the D_n test is an approximation to the maximum averaged power test, almost regardless of the weighting measure.

Likelihoods in regression models

Ex. 5.42 *Extending theory and methods to regression setups, I.* Above we have dealt with likelihood methods, involving ML estimation, limit distributions under and outside model conditions, the Wilks theorem for profiled log-likelihoods, broadly valid for all smooth parametric models, etc. – but after all under simple i.i.d. conditions. Crucially, most of these concepts, methods, and results extend to classes of general regression models. Here we go through the various steps to see how the scene broadens and to learn the appropriate extensions for concepts, techniques, and results.

Consider in general terms regression data of the form (x_i, Y_i) , with x_i a covariate vector, of length say p , thought to influence the main outcome Y_i . We assume here that the Y_i are independent given the covariates. Let $f(y_i | x_i, \theta)$ be a suitable density for y_i given x_i , with score function $u(y_i | x_i, \theta) = \partial \log f(y_i | x_i, \theta) / \partial \theta$ and information function $i(y_i | x_i, \theta) = \partial^2 \log f(y_i | x_i, \theta) / \partial \theta \partial \theta^t$. The θ could comprise both regression coefficients and parameters describing the shape of the distributions. In this exercise we assume that the model holds, with θ_0 denoting the true parameter, an inner point in the parameter space. We also postulate *the ergodic condition*, that averages over covariates stabilise, with increasing sample size; formally, for each bounded $h(x)$, there is a well-defined limit $h_0 = \int h(x) dR(x)$ for $n^{-1} \sum_{i=1}^n h(x_i)$, for an appropriate distribution R on the covariate space. For theory and applications, we do not need to model this R , or take it explicitly into account, beyond postulating its existence.

ergodic
conditions

(a) First of all, there is a log-likelihood function, also in these regression setups, $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i | x_i, \theta)$. The ML estimator $\hat{\theta}$ is its maximiser, satisfying also $U_n(\hat{\theta}) = 0$, with $U_n(\theta) = \sum_{i=1}^n u(y_i | x_i, \theta)$. Secondly, to extend theory and results for the i.i.d. case, see Ex. 5.10, we need to understand $U_n = n^{-1/2} U_n(\theta_0) = n^{-1/2} \sum_{i=1}^n u(Y_i | x_i, \theta_0)$. Show that it has mean zero and variance matrix $J_n = n^{-1} \sum_{i=1}^n J(x_i)$, where $J(x_i) = \text{Var}_{\theta_0} u(Y_i | x_i, \theta_0)$. Under ergodic assumptions, there is convergence $J_n \rightarrow J$, say. Give Lindeberg type conditions under which $U_n \rightarrow_d U \sim N_p(0, J)$.

(b) Extend techniques from Ex. 5.10 to deduce the natural parallel to $A_n \rightarrow_d A$ convergence in this general regression setting, that

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = U_n^t s - \frac{1}{2} J_n s + r_n(s) \rightarrow_d A(s) = U^t s - \frac{1}{2} s^t J s,$$

Set up clear but mild regularity conditions, as in Ex. 6.9, to secure the required $r_n(s) \rightarrow_{\text{pr}} 0$ for each s . Explain as in Ex. 5.10 that this leads to the two fundamental results

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N_p(0, J^{-1}) \quad \text{and} \quad 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \rightarrow_d \chi_p^2,$$

seen before in (5.6), but now under the more general Lindeberg type conditions. Show also that the observed Fisher information matrix

observed Fisher
information
matrix

$$\hat{J}_{n,\text{full}} = -\partial^2 \ell_n(\hat{\theta}) / \partial \theta \partial \theta^t, \quad (5.15)$$

i.e. the Hessian matrix associated with the maximisation of the log-likelihood, satisfies $\hat{J}_n = n^{-1} \hat{J}_{n,\text{full}} \rightarrow_{\text{pr}} J$. Deduce from this that $\hat{\theta} \approx_d N_p(\theta_0, \hat{J}_{n,\text{full}}^{-1})$. (xx need to point to Ch2 thing with weak LLN for averages of non-i.i.d. xx)

(c) (xx rounding this off. xx) also log-likelihood profiling and deviance and χ_1^2 limit and CDs. the point is that i.i.d. results all extend to the broad regression cases. point also to exercise below with outside the model. ML representation from Ex. 5.11 goes through: $\hat{\theta} - \theta_0 = J_n^{-1} n^{-1} \sum_{i=1}^n u(Y_i | x_i, \theta_0) + o_{\text{pr}}(n^{-1/2})$.

Ex. 5.43 *Linear regression revisited.* (xx edit and clean. xx) Consider the linear regression model of Ex. 3.33, with $Y_i | x_i \sim N(x_i^t \beta, \sigma^2)$, for which exact finite-sample theory has been well developed. We now take another look at this classical model, with the general likelihood tools.

(a) For the log-likelihood, show that $\ell_n(\beta, \sigma) = -n \log \sigma - \frac{1}{2} Q(\beta)/\sigma^2 - \frac{1}{2} n \log(2\pi)$, with $Q(\beta) = \sum_{i=1}^n (y_i - x_i^t \beta)^2$. Show that the ML estimator for β is the least squares estimator $\hat{\beta} = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i Y_i$, with $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t$, see the exercise mentioned, and that $\hat{\sigma} = (Q_0/n)^{1/2}$, with $Q_0 = Q(\hat{\beta})$ the minimum of $Q(\beta)$.

(b) Show that the score function becomes

$$u(y_i | x_i, \beta, \sigma) = \begin{pmatrix} (1/\sigma^2)(y_i - x_i^t \beta)x_i \\ -1/\sigma + (1/\sigma^3)(y_i - x_i^t \beta)^2 \end{pmatrix} = \begin{pmatrix} (1/\sigma)\varepsilon_i x_i \\ (1/\sigma)(\varepsilon_i^2 - 1) \end{pmatrix}$$

in terms of $\varepsilon_i = (y_i - x_i^t \beta)/\sigma$, which are independent standard normals under the model. With (β_0, σ_0) the true parameters, deduce that the $(p+1) \times (p+1)$ Fisher information matrix becomes

$$J_n = n^{-1} \sum_{i=1}^n \text{Var}_{\beta_0, \sigma_0} u(Y_i | x_i, \beta_0, \sigma_0) = (1/\sigma_0^2) \text{diag}(\Sigma_n, 2),$$

with $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t = n^{-1} X^t X$. Show also that the observed Fisher information matrix becomes $\hat{J}_{n, \text{full}} = (n/\hat{\sigma}^2) \text{diag}(\Sigma_n, 2)$.

(c) (xx then on to what likelihood theory implies for $\hat{\beta}$ and $\hat{\sigma}$. the J_n and $\hat{J}_n$. a t_{n-p} vs. approximate normality. we reproduce the $\hat{\beta}$ distribution, and come close for $\hat{\sigma}^2$. xx)

Ex. 5.44 *Logistic regression.* Consider binary outcome data, where the values 0-1 for Y_i are influenced by a covariate vector x_i , of dimension say p . The logistic regression model takes the probabilities to be

logistic
regression

$$p_i = \Pr(Y_i = 1 | x_i) = H(x_i^t \beta) = \frac{\exp(x_i^t \beta)}{1 + \exp(x_i^t \beta)} \quad \text{for } i = 1, \dots, n, \quad (5.16)$$

with $H(u) = \exp(u)/\{1 + \exp(u)\}$ the logistic transform, studied in Ex. 1.65. One may interpret the model via underlying or latent i.i.d. variables Z_i , having the logistic distribution, with the individuals having different thresholds; $p_i = \Pr(Z_i \leq x_i^t \beta)$ says that outcomes are '1' for those individuals whose Z_i fall to the left of the threshold.

(a) Show that $H(u) = p$ means $u = H^{-1}(p) = \log\{p/(1-p)\}$, so that the model can be represented as $\log\{p_i/(1-p_i)\} = x_i^t \beta$.

(b) Show that $\Pr(Y_i = y | x_i) = p_i^y (1-p_i)^{1-y}$, for the two outcomes, and deduce that the log-likelihood function can be written

$$\ell_n(\beta) = \sum_{i=1}^n \{y_i \log p_i + (1-y_i) \log(1-p_i)\} = \sum_{i=1}^n [y_i x_i^t \beta - \log\{1 + \exp(x_i^t \beta)\}].$$

Show from this that the estimation equation, giving rise to the ML estimator $\hat{\beta}$, is

$\sum_{i=1}^n (y_i - p_i)x_i = 0$, and that

$$J_{n,\text{full}}(\beta) = -\frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^t} = \sum_{i=1}^n p_i(1-p_i)x_i x_i^t = \sum_{i=1}^n H(x_i^t \beta) \{1 - H(x_i^t \beta)\} x_i x_i^t.$$

This matrix is assumed here to be positive definite, which in particular requires $n \geq p$. Explain that the log-likelihood function is concave, with $\hat{\beta}$ the unique maximiser.

(c) Show that, under model conditions, $\hat{\beta} \approx_d N_p(\beta, \hat{J}_{n,\text{full}}^{-1})$, where $\hat{J}_{n,\text{full}} = J_{n,\text{full}}(\hat{\beta})$ is the observed Fisher information matrix (the Hessian matrix of minus the normalised log-likelihood function, at the ML position).

(d) Consider an individual, perhaps outside the dataset, with covariate vector x_0 . Show that $x_0^t \hat{\beta}$ is approximately a normal $(x_0^t \beta, x_0^t \hat{J}_{n,\text{full}}^{-1} x_0)$, and use this to construct a confidence interval for $p(x_0) = \Pr(Y_0 = 1 | x_0)$.

(e) (xx repair a bit here. xx) Consider the important special case of a single x_i recorded for Y_i , where we write the model equation as $p_i = H(a + bx_i)$, corresponding to 2-size vectors $(1, x_i)^t$ in the more general notation used above. Show that $(\hat{a}, \hat{b})$ are the solutions to $\sum_{i=1}^n (y_i - p_i) = 0$ and $\sum_{i=1}^n (y_i - p_i)x_i = 0$, and that

$$J_n(a, b) = \sum_{i=1}^n \frac{\exp(a + bx_i)}{\{1 + \exp(a + bx_i)\}^2} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}.$$

(f) The logistic regression model uses the H transform, which is symmetric around zero. A more flexible model takes $p_i = H(x_i^t \beta)^\kappa$, allowing skewness in the underlying distribution for the latent variables. Write down the log-likelihood function, explain that standard theory applies for the ML estimator $(\hat{\beta}, \hat{\kappa})$. (xx point perhaps to illustration, Polish girls story, where κ is significantly larger than 1. xx)

Ex. 5.45 Probit regression. Consider again data (x_i, Y_i) , with covariates x_i and 0-1 outcomes Y_i . For the logistic regression of Ex. 5.44 we modelled the 1-probabilities as $p_i = \Pr(Z_i \leq x_i^t \beta) = H(x_i^t \beta)$, with H the logistic c.d.f. Clearly there are alternatives, and the more famous one among these is *probit regression*, which takes the Z_i to be standard normal, i.e. $p_i = \Phi(x_i^t \beta)$.

probit
regression, from
probability of
unit

(a) Explain that the log-likelihood becomes $\ell_n(\beta) = \sum_{i=1}^n [y_i \log p_i(\beta) + (1 - y_i) \log \{1 - p_i(\beta)\}]$. The ML estimator $\hat{\beta}$ is its maximiser.

(b) Explain also that standard regression likelihood theory applies, with $\hat{\beta} \approx_d N_p(\beta, \hat{J}_n^{-1})$, and give a recipe for computing the variance matrix.

(c) To compare analyses from logistic and probit regressions, we may scale the normal above to have the same variance $\tau^2 = \pi^2/3$ as the logistic. Explain that this leads to an equivalent model, though with scaled parameters, namely $p_i = \Pr(Z'_i \leq x_i^t \beta) = \Phi(x_i^t \beta / \tau)$.

(d) Simulate a simple dataset of size $n = 1000$ pairs, with x_i taken $N(0, \kappa^2)$ with $\kappa = 2$ and true probabilities $p_i = H(a_0 + b_0 x_i)$ with $a_0 = -0.33$ and $b_0 = 0.77$. The little point with $\kappa = 2$ is to secure a fair range for the x_i , in view of the probabilities. (i) Estimate (a, b) in the logistic regression model $p_i = H(a + bx_i)$; and then (ii) estimate (a, b) in the probit model $p_i = \Phi((a + bx_i)/\tau)$. Draw the two estimated regression curves, say $H(\hat{a} + \hat{b}x)$ and $\Phi((\hat{a} + \hat{b}x)/\tau)$. Comment on what you find.

Ex. 5.46 Poisson regression. Consider independent count data $y_1, \dots, y_n$, influenced by covariate vectors $x_1, \dots, x_n$. The Poisson regression model, in its standard form, takes $Y_i \sim \text{Pois}(\mu_i)$, with $\mu_i = \exp(x_i^t \beta)$.

(a) Show that the log-likelihood function becomes

$$\ell_n(\beta) = \sum_{i=1}^n \{-\mu_i + y_i \log(\mu_i)\} = \sum_{i=1}^n \{y_i x_i^t \beta - \exp(x_i^t \beta)\},$$

and that the equations $\sum_{i=1}^n \{y_i - \mu_i(\beta)\} x_i = 0$ define the ML estimators.

(b) Show that $-\partial^2 \ell_n(\beta) / \partial \beta \partial \beta^t = \sum_{i=1}^n \mu_i x_i x_i^t$, leading to the observed Fisher information matrix $\hat{J}_{n,\text{full}} = \sum_{i=1}^n \hat{\mu}_i x_i x_i^t$, where $\hat{\mu}_i = \exp(x_i^t \hat{\beta})$.

(c) For a case with covariate vector x_0 , estimate the associated expected count $\mu_0 = \exp(x_0^t \beta)$, and construct a confidence interval.

Ex. 5.47 Poisson with overdispersion. Sometimes there is overdispersion, compared to how the counts observations y_i should behave under Poisson conditions. Such overdispersion could e.g. reflect ‘hidden covariates’ not taken on board in the model.

(a) A method handling overdispersion is to take $Y_i | \mu_i \sim \text{Pois}(\mu_i)$, but modelling potential extra randomness via $\mu_i \sim \text{Gam}(\exp(x_i^t \beta) / \tau, 1/\tau)$. This is an application of Poisson-gamma mixtures from Ex. 1.27. Show that μ_i has mean $\exp(x_i^t \beta)$ and variance $\tau \exp(x_i^t \beta)$; τ small means getting back to ordinary Poisson regression. Show also that Y_i has mean $\exp(x_i^t \beta)$ variance $(1 + \tau) \exp(x_i^t \beta)$. Furthermore, work out that an expression for the log-likelihood function $\ell_n(\beta, \tau)$ is

variances are bigger with the factor $1 + \tau$

$$\sum_{i=1}^n \{(a_i/\tau) \log(1/\tau) - \log \Gamma(a_i/\tau) + \log \Gamma(a_i/\tau + y_i) - (a_i/\tau + y_i) \log(1/\tau + 1)\},$$

in which $a_i = \exp(x_i^t \beta)$. In applications it is often fruitful to profile out the β , and studying $\ell_{n,\text{prof}}(\tau) = \max_{\beta} \ell_n(\beta, \tau)$. (xx pointer to Story #28.

(b) As long as the τ sought for is safely inside its parameter range, explain that the ML machinery works, for growing n , giving estimators $(\hat{\beta}, \hat{\tau})$ along with inference, with testing and confidence. Matters are different when the overdispersion parameter τ is zero or close to zero, however; in particular, the log-likelihood profile $\ell_{n,\text{prof}}(\tau)$ can have its maximum at zero. To understand what happens at this boundary, where data are actually Poisson distributed, we now study the simplest case, senza covariates, parameters (θ, τ) only. Explain that the log-density is

$$\log g(y, \theta, \tau) = \log \Gamma(\theta/\tau + y) - \log \Gamma(\theta/\tau) + (\theta/\tau) \log(1/\tau) - (\theta/\tau + y) \log(1 + 1/\tau).$$

The task is now to approximate this function carefully, expanded up to order $O(\tau^2)$.

(c) Below we use the careful approximation to the log-gamma function given by

$$\log \Gamma(z) \doteq h(z) = (z - \frac{1}{2}) \log z - z + (1/12)(1/z) + O(1/z^3)$$

for z large. Show that the derivatives are

$$\begin{aligned} h'(z) &= \log z - 1 - \frac{1}{2}z^{-1} - (1/12)z^{-2} + O(1/z^4), \\ h''(z) &= 1/z + \frac{1}{2}z^{-2} - (1/6)z^{-3} + O(1/z^4), \\ h'''(z) &= -1/z^2 - 1/z^3 + O(1/z^4), \end{aligned}$$

if curious, check Abramowitz and Stegun (1965, Section 6.1)

and explain hence that $\log \Gamma(z + a) - \log \Gamma(z)$ can be well approximated with

$$\{\log z - 1 - \frac{1}{2}z^{-1} - (1/12)z^{-2}\}a + \frac{1}{2}\{1/z + \frac{1}{2}z^{-2}\}a^2 - (1/6)(1/z^2)a^3.$$

This leads to $\log \Gamma(\theta/\tau + y) - \log \Gamma(\theta/\tau)$ well approximated by

$$I = \{\log(\theta/\tau) - 1 - \frac{1}{2}\tau/\theta - (1/12)\tau^2/\theta^2\}y + \frac{1}{2}\{\tau/\theta + \frac{1}{2}\tau^2/\theta^2\}y^2 - (1/6)(\tau^2/\theta^2)y^3.$$

Now couple this with

$$\begin{aligned} II &= (\theta/\tau)[- \log \tau - \log(1 + \tau) + \log \tau] - y \log(1 + \tau) + y \log \tau \\ &= -\theta(1 - \frac{1}{2}\tau + \frac{1}{3}\tau^2) - y(\tau - \frac{1}{2}\tau^2) + O(\tau^3) \end{aligned}$$

to get $I + II$ equal to

$$\begin{aligned} \log g(y, \theta, \tau) &= y \log \theta - \theta + \tau A(y) - \frac{1}{2}\tau^2 B(y) + O(\tau^3), \text{ with} \\ A(y) &= \frac{1}{2}y^2/\theta - y(1 + \frac{1}{2}/\theta) + \frac{1}{2}\theta, \\ B(y) &= (1/6)y/\theta^2 - (1/2)y^2/\theta^2 + (1/3)y^3/\theta^2 + (2/3)\theta - y. \end{aligned}$$

(d) When the Y is actually a $\text{Pois}(\theta)$, show that $A(Y)$ miraculously has mean zero and constant variance $\frac{1}{2}$, and that this variance is also equal to the mean of $B(Y)$. Show also that the Fisher matrix $J(\theta, \tau)$ tends to $J(\theta, 0)$, diagonal, with elements $1/\theta$ and $\frac{1}{2}$.

(e) For an i.i.d. set of size n , show using these approximations that the ML $\hat{\tau}$ is close to $\bar{A}_n/\bar{B}_n$, of this is positive, and zero, if $\bar{A}_n$ is negative. Using methods of Ex. 5.17, show that $\sqrt{n}\hat{\tau} \rightarrow_d \sqrt{2}\max(0, N)$, with N a standard normal. Also, explain that the deviance $D_n = 2\{\ell_{n,\max} - \ell_{n,\text{prof}}(0)\}$ has the half chi-squared limit, $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. A 0.05 level test for the hypothesis of Poisson-ness is to reject if (xx complete this xx). (xx compare to the simpler $S_n^2/\bar{Y}_n$ test, see Story #56. but there is no easy parallel to regression case. $\sum_{i=1}^n (y_i - \exp(x_i^t \hat{\beta}))^2$ compared to $\sum_{i=1}^n \exp(x_i^t \hat{\beta})$. could be equivalent to ML. xx)

(f) Now you may go back to (a), with $p+1$ parameters (β, τ) for covariate vectors of length p . Generalise the above, to show that as $\tau \rightarrow 0$, the Fisher matrix tends to $\text{diag}(\Sigma_n, 2)$, with $\Sigma_n = n^{-1} \sum_{i=1}^n \exp(x_i^t \beta) x_i x_i^t$. Give the limiting distribution for $\sqrt{n}\hat{\tau}$ under Poisson conditions, and give a test for $\tau = 0$.

Ex. 5.48 Exponential regression. Consider regression data (x_i, y_i) for $i = 1, \dots, n$, for nonnegative outcomes y_i , which we now model as $Y_i \sim \text{Expo}(\theta_i)$ with $\theta_i = \exp(x_i^t \beta)$.

(a) Show that the log-likelihood may be written $\ell_n(\beta) = \sum_{i=1}^n \{x_i^t \beta - \exp(x_i^t \beta) y_i\}$, and the the ML estimator $\hat{\beta}$ is the solution to $\sum_{i=1}^n x_i \{1 - \exp(x_i^t \beta) y_i\} = 0$.

(b) Show next that the normalised expected Fisher information matrix takes the simple form $J_n = n^{-1} \sum_{i=1}^n x_i x_i^t$. Explain from this that $\hat{\beta} \approx_d N_p(\beta, n^{-1} J_n^{-1})$, under model conditions.

(c) Conservative analysis, i.e. not necessarily trusting model conditions, leads to a more general approximation for the variance matrix, with $n^{-1} J_n^{-1} K_n J_n^{-1}$. Identify the K_n and give a consistent estimator.

Ex. 5.49 A heteroscedastic linear regression model. (xx edit and clean. xx) In various linear regression type applications for (x_i, y_i) data the linear mean assumption can be reasonable, whereas the variance might not be taken constant across covariates. Consider therefore the model with independent $Y_i | (x_i, w_i) \sim N(x_i^t \beta, \sigma_i^2)$, for $i = 1, \dots, n$, with covariate vectors x_i of length p and variance related covariates w_i of length q , influencing $\sigma_i = \sigma \exp(\gamma^t w_i)$. These w_i could be a subset of the x_i or functions thereof. It is convenient to normalise these such that $\bar{w} = n^{-1} \sum_{i=1}^n w_i = 0$, which also means that σ is the standard deviation for an average individual, with w_i equal to $\bar{w}$.

(a) Show that the log-likelihood function can be written

$$\ell_n(\beta, \gamma) = -n \log \sigma - \frac{1}{2}(1/\sigma^2)Q(\beta, \gamma), \quad \text{with } Q(\beta, \gamma) = \sum_{i=1}^n \frac{(y_i - x_i^t \beta)^2}{\exp(2\gamma^t w_i)}.$$

Show that minimising $Q(\beta, \gamma)$ over β , for fixed γ , takes place for

$$\hat{\beta}(\gamma) = \left\{ \sum_{i=1}^n \frac{x_i x_i^t}{\exp(2\gamma^t w_i)} \right\}^{-1} \sum_{i=1}^n \frac{x_i y_i}{\exp(2\gamma^t w_i)}.$$

Demonstrate that this leads to the profiled log-likelihood $\ell_{n,\text{prof}}(\gamma) = -n \log \hat{\sigma}(\gamma) - \frac{1}{2}n$, where $\hat{\sigma}(\gamma)^2 = Q_0(\gamma)/n$, with $Q_0(\gamma) = Q(\hat{\beta}(\gamma), \gamma)$ the minimum sum of squares. Deduce from this that a recipe for finding the ML estimators consists in (i) minimising $Q_0(\gamma)$ over γ , yielding $\hat{\gamma}$; (ii) reading off $\hat{\beta} = \hat{\beta}(\hat{\gamma})$ and $\hat{\sigma} = \hat{\sigma}(\hat{\gamma})$.

(b) (xx calibrate this with Wilks things. xx) Is it worthwhile, turning from classic linear regression, to include the extra layer of variance heterogeneity sophistication? Show that the log-likelihood-ratio test becomes that of comparing $D_n = 2n \log\{\hat{\sigma}(0)/\hat{\sigma}(\hat{\gamma})\}$ to the χ_q^2 , in which $\hat{\sigma}(0)^2 = Q_0(0)/n$ is the standard estimator for σ^2 under variance constancy.

(c) For the $p + q + 1$ -parameter model, with parameters β, γ, σ , show that the score function becomes

$$u(y_i | x_i) = \begin{pmatrix} (1/\sigma^2)(y_i - x_i^t \beta)x_i / \exp(2\gamma^t w_i) \\ -w_i + (1/\sigma^2)(y_i - x_i^t \beta)^2 w_i / \exp(2\gamma^t w_i) \\ -1/\sigma + (1/\sigma^3)(y_i - x_i^t \beta)^2 / \exp(2\gamma^t w_i) \end{pmatrix} = \begin{pmatrix} (1/\sigma)\varepsilon_i x_i / \exp(\gamma^t w_i) \\ (1/\sigma)(\varepsilon_i^2 - 1)w_i \\ (1/\sigma)(\varepsilon_i^2 - 1) \end{pmatrix},$$

in terms of $\varepsilon_i = (y_i - x_i^t \beta) / \{\sigma \exp(\gamma^t w_i)\}$. Show from this that the normalised Fisher information matrix becomes $J_n = (1/\sigma_0^2) \text{diag}(\Sigma_n(\gamma_0), M_n, 2)$, at the true parameters $(\beta_0, \gamma_0, \sigma_0)$, in terms of $\Sigma_n(\gamma) = n^{-1} \sum_{i=1}^n x_i x_i^t / \exp(2\gamma^t w_i)$ and $M_n = n^{-1} \sum_{i=1}^n w_i w_i^t$.

(d) (xx round off. confidence for $\mu(x_0, w_0)$, $x_0^t \hat{\beta} \pm 1.96 \hat{\sigma} \exp(\hat{\gamma}^t w_0)$. pointer to Story #27. xx)

Ex. 5.50 Nonlinear regression. (xx calibrate this with both classic linear regression and what we've said with general regression models. xx) Consider in general terms the model with independent $y_i \sim N(m_i(\beta), \sigma^2)$ for $i = 1, \dots, n$, where the means $m_i(\beta)$ are perhaps nonlinear functions of an appropriate vector parameter β , involving also covariates.

(a) Show that the log-likelihood function becomes $-n \log \sigma - \frac{1}{2}Q_n(\beta)/\sigma^2$, with $Q_n(\beta) = \sum_{i=1}^n \{y_i - m_i(\beta)\}^2$. Show that the ML estimator for β is the minimiser of Q_n , and that $\hat{\sigma}^2 = Q_n(\hat{\beta})/n$.

(b) With $m_i^*(\beta) = \partial m_i(\beta) / \partial \beta$, show that the normalised Fisher matrix becomes

$$J_n(\beta, \sigma) = \frac{1}{\sigma^2} \begin{pmatrix} \Sigma_n & 0 \\ 0 & 2 \end{pmatrix}, \quad \text{with } \Sigma_n = n^{-1} \sum_{i=1}^n m_i^*(\beta) m_i^*(\beta)^t.$$

(c) Deduce that $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2 \Sigma^{-1})$, under Lindeberg type conditions, where Σ is the limit of Σ_n .

Counting the uncounted

Ex. 5.51 How many unseen are there? Suppose a multinomial experiment has been carried out, involving N individuals, sorted in boxes $0, 1, \dots, k$, with probabilities $p_0, p_1, \dots, p_k$

summing to one. The full set of observations ought to have been $(N_0, N_1, \dots, N_k)$, counting the numbers in these boxes, but here we have in mind situations where the zero-box count N_0 has not been observed. The task is to estimate N_0 , and hence $N = N_0 + R$, with $R = N_1 + \dots + N_k$ the total observed count in boxes $1, \dots, k$. The setup of Ex. 5.53, involving boxes 00, 01, 10, 11, is a special case. Methods developed below are being used in Stories #63 and #64 to estimate how many persons were killed in Srebrenica 1995 and in Guatemala, 1978–1996, based on a Venn diagram constructed from respectively two and three lists of murdered people.

(a) Assume a smooth parametric model is used for the probabilities, with $p_j = p_j(\theta)$ for $j = 0, 1, \dots, k$, with θ of dimension $r \leq k - 1$, admitting two continuous derivatives. Show that the log-likelihood may be written

$$\ell(N, \theta) = \log(N!) - \log((N - R)!) + (N - R) \log p_0(\theta) + \sum_{j=1}^k N_j \log p_j(\theta).$$

Explain that this also leads to a profiled log-likelihood in N ,

$$\ell_{\text{prof}}(N) = \max\{\ell(N, \theta) : \text{all } \theta\} = \ell(N, \hat{\theta}_N).$$

Maximising this function gives the ML estimators $(\hat{N}, \hat{\theta})$.

(b) We do assume that the parametric model holds, with true parameter θ_0 an inner point in the parameter space; write also N_0 for the true N . Large-sample results will be reached below in the framework of growing N_0 . It will be seen that $\hat{N}$ and $\hat{\theta}$ will converge at different rates, so it is helpful to write $N = N_0 \rho$ and $\hat{\rho} = \hat{N}/N_0$. Explain that this leads to the approximate log-likelihood $\ell(\rho, \theta)$ equal to

$$N_0 \rho \log(N_0 \rho) - (N_0 \rho - R) \log(N_0 \rho - R) + (N_0 \rho - R) \log p_0(\theta) + \sum_{j=1}^k N_j \log p_j(\theta).$$

(c) From basic multinormal limits for multinomials, as with Ex. 2.51, show that there is joint convergence $N_0^{1/2} \{N_j/N - p_j(\theta_0)\} \rightarrow_d A_j$, a zero-mean multinormal vector with $\text{cov}(A_j, A_\ell) = p_j(\theta_0) \{\delta_{j,\ell} - p_\ell(\theta_0)\}$; note that $A_0 + \dots + A_k = 0$. With $\tilde{p}_0 = 1 - R/N_0$, show that $N_0^{1/2} \{\tilde{p}_0 - p_0(\theta_0)\} \rightarrow_d A_0$.

(d) Letting $u_j(\theta) = \partial \log p_j(\theta) / \partial \theta$, explain that $\sum_{j=0}^k p_j(\theta) u_j(\theta) = 0$. Below, write for crispness p_j and u_j for $p_j(\theta_0)$ and $u_j(\theta_0)$. Put up formulae for the first-order partial derivatives of the log-likelihood $\ell(\rho, \theta)$. Show that when evaluated at (N_0, θ_0) , these become

$$\begin{aligned} U_0 &= N_0(-\log \tilde{p}_0 + \log p_0), \\ V_0 &= N_0 \left\{ (\tilde{p}_0 - p_0) u_0 + \sum_{j=1}^k (N_j/N_0 - p_j) u_j \right\}. \end{aligned}$$

(e) A key quantity is $M = \sum_{j=0}^k p_j u_j u_j^t$. Show that there is convergence in distribution

$$\frac{1}{N_0^{1/2}} \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \rightarrow_d \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} -(1/p_0) A_0 \\ \sum_{j=0}^k A_j u_j \end{pmatrix} \sim N_{r+1}(0, J), \text{ with } J = \begin{pmatrix} (1 - p_0)/p_0 & -u_0^t \\ -u_0 & M \end{pmatrix}.$$

This is the Fisher information matrix for this reparametrised model, with parameters (ρ, θ) , evaluated at $(1, \theta_0)$. Verify via derivatives and limits that J is also the limit in

probability of $-N_0^{-1}H_0$, where H_0 is the $(r+1) \times (r+1)$ Hessian matrix of second order partial derivatives of $\ell(\rho, \theta)$, evaluated at $(1, \theta_0)$.

(f) Show from this that $N_0^{1/2}(\hat{N}/N_0 - 1) = (\hat{N} - N_0)/N_0^{1/2}$ tends in distribution to $N(0, \tau^2)$, with

$$\tau^2 = \frac{1}{(1-p_0)/p_0 - u_0^t M^{-1} u_0} = \frac{p_0}{1-p_0-p_0\delta}, \quad \text{where } \delta = u_0^t M^{-1} u_0.$$

Though the main concern might be estimating N_0 and $N = N_0 + R$, show also that $N_0^{1/2}(\hat{\theta} - \theta_0) \rightarrow_d N_r(0, K)$, with

$$K = (M - (1-p_0)u_0 u_0^t / p_0)^{-1} = M^{-1} + dM^{-1}u_0 u_0^t M^{-1},$$

where we write $d = r_0/(1-r_0\delta) = p_0/(1-p_0-p_0\delta)$ and $r_0 = p_0/(1-p_0)$. Explain that had N_0 been observed too, the result would have been $N_r(0, M^{-1})$, so $dM^{-1}u_0 u_0^t M^{-1}$ is the extra amount of uncertainty when estimating θ , caused by N_0 not being counted.

(g) Show via Wilks theory for profiled log-likelihood functions, as in Ex. 5.13, that

$$D(N_0) = 2\{\ell_{\text{prof}, \max} - \ell_{\text{prof}}(N_0)\} \rightarrow_d \chi_1^2.$$

This leads to confidence intervals, e.g. $\{N_0 : D(N_0) \leq 1.96^2\}$, and a full confidence curve $\text{cc}(N) = \Gamma_1(D(N_0))$, as per Ch. 8, which we can read off directly from the profiled log-likelihood, without needing the perhaps cumbersome estimation of the J matrix.

(h) For the above results we are assuming that the parametric model $p_j(\theta)$ holds, also for the null box, which cannot really be verified or tested from $N_1, \dots, N_k$ data alone, without other assumptions again, or perhaps follow-up data. Show in general terms that the likelihood can be written

$$\begin{aligned} L(N, \theta) &= \frac{N!}{N_0! N_1! \dots N_k!} p_0^{N_0} p_1^{N_1} \dots p_k^{N_k} \\ &= \frac{N!}{R! (N-R)!} p_0^{N-R} (1-p_0)^R \frac{R!}{N_1! \dots N_k!} q_1^{N_1} \dots q_k^{N_k}, \end{aligned}$$

splitting into a binomial $R \sim \text{binom}(N, 1-p_0)$ and a multinomial $(R, q_1, \dots, q_k)$ for what is actually observed, with $q_j = p_j/(1-p_0)$. We may infer the q_j from the p_j but not p_0 from $q_1, \dots, q_k$. Show also the ML estimator $\hat{\theta}$ can be found from the multinomial $(N_1, \dots, N_k)$ given R , followed by $\hat{N} = R/\{1-p_0(\hat{\theta})\}$.

Ex. 5.52 Hardy-Weinberg when seeing only two of the three categories. We visit Hardy-Weinberg island, in clear equilibrium, where individuals may be sorted into categories 0, 1, 2, with probabilities $(1-\theta)^2, 2\theta(1-\theta), \theta^2$. Suppose however that when accountants arrive, all the zero box people go into hiding, so from the trinomial (N_0, N_1, N_2) only N_1, N_2 are counted. How can we estimate N_0 , and the size of the population $N = N_0 + N_1 + N_2$ investigated?

(a) Show that given $R = N_1 + N_2$, N_2 is a binomial (R, q_2) , with $q_2 = p_1/(p_1 + p_2) = \theta/(2-\theta)$. Show that the ML estimators are

$$\hat{\theta} = \frac{2N_2}{N_1 + 2N_2}, \quad \hat{N} = \frac{R}{1 - (1-\hat{\theta})^2} = \frac{R}{1 - \{N_1/(N_1 + 2N_2)\}^2}.$$

(b) We may find confidence intervals via the methods of Ex. 5.51. Explain that the log-likelihood is

$$\begin{aligned}\ell(N, \theta) &= \log \frac{(N!)}{(N-R)!} + 2(N-R) \log(1-\theta) + N_1 \{\log \theta + \log(1-\theta)\} + 2N_2 \log \theta \\ &= \log(N!) - \log((N-R)!) + (N_1 + 2N_2) \log \theta + (2(N-R) + N_1) \log(1-\theta),\end{aligned}$$

and that this leads to the profiled log-likelihood

$$\ell_{\text{prof}}(N) = \log(N!) - \log((N-R)!) + (N_1 + 2N_2) \log \hat{\theta}_N + (2(N-R) + N_1) \log(1 - \hat{\theta}_N),$$

in which $\hat{\theta}_N = (N_1 + 2N_2)/N$.

(c) To find the limiting variance τ^2 of $(\hat{N} - N)/N^{1/2}$, we need $u_0 = -2/(1-\theta)$, $u_1 = 1/\theta - 1/(1-\theta)$, $u_2 = 2/\theta$. Verify that $\sum_{0,1,2} p_j u_j = 0$, and show that

$$M = \sum_{0,1,2} p_j u_j^2 = 8 + 8 \frac{(\theta - \frac{1}{2})^2}{\theta(1-\theta)} = \frac{2}{\theta(1-\theta)}.$$

Derive from this that $\delta = u_0^2/M = 2\theta/(1-\theta)$, and that the formula for τ^2 of Ex. 5.51 simplifies to $(1-\theta)^2/\theta^2$. It is hence considerably harder to estimate N with good precision if θ is small, with a high proportion N_0/N in hiding.

Ex. 5.53 Capture-recapture. The following setup is a special separate interest case of the general hidden-box framework of Ex. 5.51. Consider a 2×2 table setup with counts $N_{0,0}, N_{0,1}, N_{1,0}, N_{1,1}$, corresponding to $N_{i,j}$ counting the cases of $(X = i, Y = j)$, for $i, j = 0, 1$, for two factors X and Y . We take the four counts to be a multinomial vector with probabilities $p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1}$. Assume now that the 00 box is hidden, hence also the total number $N = N_{0,0} + R$, with $R = N_{0,1} + N_{1,0} + N_{1,1}$ the actually observed items for boxes $(0, 1), (1, 0), (1, 1)$. How can one estimate the hidden $N_{0,0}$, and then in its turn N ? Such questions, with different estimation methods, go back to Petersen (1896), who needed to estimate fish populations in the 1890ies. Versions of methods developed and illustrated below are being used in Stories #63 and #64 for estimating the number of killed persons in two utterly horrible cases, Srebrenica 1995 and Guatemala 1978–1995.

(a) Assume in this exercise that factors X and Y are independent, with $\Pr(X = 1) = p = p_{1,\cdot}$ and $\Pr(Y = 1) = q = p_{\cdot,1}$; we use ‘ $\cdot$ ’ notation to indicate that the index in question is being summed over. Show that

$$p_{0,0} = (1-p)(1-q), \quad p_{0,1} = (1-p)q, \quad p_{1,0} = p(1-q), \quad p_{1,1} = pq.$$

Argue then that $N_{1,\cdot}N_{\cdot,1}/N^2$ and $N_{1,1}/N$ are both valid estimates of $p_{1,1}$. Discuss conditions under which $N^* = N_{1,\cdot}N_{\cdot,1}/N_{1,1}$ is a reasonable estimator of N .

(b) The N is unknown, but we may still study the usual ratios $\hat{p}_{i,j} = N_{i,j}/N$. Show that there is joint convergence in distribution, say $N^{1/2}(\hat{p}_{i,j} - p_{i,j}) \rightarrow_d A_{i,j}$, as N increases, with the $A_{i,j}$ forming a four-dimensional mean zero normal. Give its variance matrix. Under independence, show that $(N^* - N)/N^{1/2} = N^{1/2}(N^*/N - 1)$ has limit distribution

$$\begin{aligned}U &= (1/p)(A_{1,0} + A_{1,1}) + (1/q)(A_{0,1} + A_{1,1}) - \{1/(pq)\}A_{1,1} \\ &= \{pA_{0,1} + qA_{1,0} + (p+q-1)A_{1,1}\}/(pq).\end{aligned}$$

This is a normal $(0, \tau^2)$; show that indeed $\tau^2 = (1-p)(1-q)/(pq)$. How can this be used to form a confidence interval for N ?

see also Hjort
and Petersen
(1905)

(c) Show that the N^* leads to the natural estimator $\hat{p} = N_{1,1}/N_{1\cdot}$, for p . Find its approximate distribution, and assess how much is lost in precision by not knowing N .

(d) Show that the log-likelihood function may be written

$$\begin{aligned}\ell(N, p, q) &= \log(N!) - \log((N - R)!) + (N - R) \log p_{0,0} \\ &\quad + N_{0,1} \log p_{0,1} + N_{1,0} \log p_{1,0} + N_{1,1} \log p_{1,1}.\end{aligned}$$

Under the independence assumption above, simplify this expression, and show that the profiled log-likelihood becomes

$$\ell_{\text{prof}}(N) = \ell(N, \hat{p}_N, \hat{q}_N) = \log(N!) - \log((N - R)!) + N\{H(\hat{p}_N) + H(\hat{q}_N)\},$$

with $H(x) = x \log x + (1 - x) \log(1 - x)$, and where $\hat{p}_N = N_{1\cdot}/N$ and $\hat{q} = N_{\cdot,1}/N$. With $\hat{N}$ the maximiser, i.e. the ML estimator, we have from Ex. 5.51 that $D(N_0) = 2\{\ell_{\text{prof}}(\hat{N}_0) - \ell_{\text{prof}}(N_0)\} \rightarrow_d \chi_1^2$, at the true N_0 position. Explain that $\text{cc}(N) = \Gamma_1(D(N))$ becomes an approximate confidence curve for N .

(e) The theory of Ex. 5.51 applies, both for reading off confidence intervals and for assessing the variance of $\hat{N}$. To this end, using notation from that exercise, find the four log-probability derivatives $u_{i,j}(p, q)$, verify that $\sum_{i,j} p_{i,j} u_{i,j} = 0$, and show that $M = \sum_{i,j} p_{i,j} u_{i,j} u_{i,j}^t$ is the diagonal matrix with elements $1/(p(1 - p)), 1/(q(1 - q))$. From this find also

$$\delta = u_{0,0}^t M^{-1} u_{0,0} = p/(1 - p) + q/(1 - q),$$

and show that the limit variance for $N^{1/2}(\hat{N}/N - 1)$ is $\tau^2 = (1 - p)(1 - q)/(pq)$. This is indeed equal to the limit variance found for $N^{1/2}(N^*/N - 1)$ in Story #63 for the Petersen estimator N^* . This is no coincidence, show that N^* and $\hat{N}$ is at most 1 apart. Note that these estimates and confidence intervals are found not only without knowing $N_{0,0}$, but also not knowing the probabilities p and q of fish being caught in the two rounds.

(f) We now go on to three rounds of capture-mark-release, with probabilities p, q, r . In the multinomial setup, with $2^3 = 8$ counts $N_{i,j,k}$, show that $p_{0,0,0} = (1 - p)(1 - q)(1 - r)$, and so on, up to $p_{1,1,1} = pqr$ the probability of a fish being caught in each of the three rounds. Again, the theory of Ex. 5.51 applies, so it is a matter of working out the details for this setup. Find an explicit formula for the profiled log-likelihood function $\ell_{\text{prof}}(N) = \ell(N, \hat{p}_N, \hat{q}_N, \hat{r}_N)$, amenable for implementation and for reading off confidence intervals for N via the Wilks theorem. Set up formulae for the eight $u_{i,j,\ell}$; verify that $\sum_{i,j,\ell} p_{i,j,\ell} u_{i,j,\ell} = 0$; and show that $M = \sum_{i,j,\ell} p_{i,j,\ell} u_{i,j,\ell} u_{i,j,\ell}^t$ is diagonal, with elements $\{p(1 - p)\}^{-1}, \{q(1 - q)\}^{-1}, \{r(1 - r)\}^{-1}$. Explain that also leads to

$$\delta = u_{0,0,0}^t M^{-1} u_{0,0,0} = p/(1 - p) + q/(1 - q) + r/(1 - r).$$

Conclude that $(\hat{N} - N)/N^{1/2} \rightarrow_d N(0, \tau^2)$, with

$$\tau^2 = \frac{(1 - p)(1 - q)(1 - r)}{1 - (1 - p)(1 - q)(1 - r)(1 + \delta)} = \frac{(1 - p)(1 - q)(1 - r)}{pq + pr + qr - 2pqr}.$$

Discuss how the variance of $\hat{N}$ is influenced by say small, moderate, and higher values of p, q, r . In various applications, from fishery sciences to estimating the number of bank account cheaters, these probabilities might tend to be small.

Further material and leftovers (as of 16 August 2026)

with vocabulary
from Schweder
and Hjort
(2016, Ch. 7)

Ex. 5.54 *GLLM regression.* Consider positive data y_i , modelled in terms of two covariates x_i and w_i , both taken centred, for ease of interpretation. The model studied here takes $Y_i \sim \text{Gam}(a_i, b_i)$, with $a_i = a \exp(\beta x_i)$ and $b_i = \exp(\gamma w_i)$. It is a generalised linear-linear model, modelled log-linearly in both parameters, similar in spirit to the heteroscedastic regression of Ex. 5.49. (xx can we point to a dataset, or do this as part of a story? xx)

(a) Show that the log-likelihood function can be written

$$\ell_n(a, b, \beta, \gamma) = \sum_{i=1}^n \{a_i \log b_i - \log \Gamma(a_i) + (a_i - 1) \log y_i - b_i y_i\}.$$

Explain how this can be maximised, and indicate how confidence intervals can be set for the four parameters. How can we test $\beta = 0$ and $\gamma = 0$?

(b)

Ex. 5.55 *Extending theory and methods to regression setups, II.* (xx smooth and polish. regression outside model conditions. careful with notion, point to previous for u, i . then $\theta_{0,n}$, J_n , K_n , sandwich. xx) For i.i.d. setups we have seen how the behaviour of ML estimators can be accurately described also when the data-generating density g is not inside the parametric model f_θ , via a clearly defined least false parameter θ_0 and the two matrices J and K , see Ex. 5.10. Under model conditions, the two are equal. Here we extend these notions and results to regression models. The setting is partly as in Ex. 5.42, with a parametric model $f(y_i | x_i, \theta)$ for $Y_i | x_i$, but notions and results need to be lifted to the agnostic state of affairs when the real density $g(y_i | x_i)$ is not necessarily inside the model.

(a) For given x , there is a KL distance

$$\text{KL}_x(g(\cdot | x), f(\cdot | x, \theta)) = \int g(y | x) \log \frac{g(y | x)}{f(y | x, \theta)} dy.$$

Consider the overall KL distance, weighted over the covariate vectors in the sample,

$$\text{KL} = n^{-1} \sum_{i=1}^n \text{KL}_{x_i} = \int \int g(y | x) \log \frac{g(y | x)}{f(y | x, \theta)} dy dR_n(x),$$

with R_n the empirical distribution over these covariate vectors. Show that this is a divergence, i.e. nonnegative and equal to zero only if $g(y | x) = f(y | x, \theta)$ for all y and all $x_1, \dots, x_n$. Let $\theta_{0,n}$ be the least false parameter, minimising this KL, for the given covariate vectors, and show that this is the same as the maximiser of

$$M_n(\theta) = n^{-1} \sum_{i=1}^n \int g(y | x_i) \log f(y | x_i, \theta) dy.$$

Explain also that $M_n(\theta) \rightarrow M(\theta)$, under ergodic conditions, with this limit involving the distribution R over covariate space, using $\int h dR_n \rightarrow \int h dR$. Letting θ_0 be the maximiser of this limit $M(\theta)$, show also that $\theta_{0,n} \rightarrow \theta_0$, under weak conditions.

(b) The following arguments support the notion that the ML estimator $\hat{\theta}$ aims at this least false $\theta_{0,n}$. With $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i | x_i, \theta)$ the log-likelihood function, explain that $n^{-1} \ell_n(\theta)$ has mean $M_n(\theta)$. the mean of $n^{-1} \ell_n(\theta)$ is $M_n(\theta)$, and that its variance goes to zero. Now show that $\hat{\theta} - \theta_{0,n} \rightarrow_{\text{pr}} 0$.

(c) As in the earlier and simpler case of (5.5) it is fruitful to work with the log-likelihood function process. Show that

$$A_n(s) = \ell_n(\theta_{0,n} + s/\sqrt{n}) - \ell_n(\theta_{0,n}) = U_n^t s - \frac{1}{2} s^t J_n s + r_n(s),$$

now with $U_n = n^{-1/2} \sum_{i=1}^n u(Y_i | x_i, \theta_{0,n})$ and $J_n = -n^{-1} \sum_{i=1}^n i(Y_i | x_i, \theta_{0,n})$. Let $K(x_i) = \text{Var } u(Y_i | x_i, \theta_{0,n})$. Under Lindeberg type conditions, show that $U_n \rightarrow_d U \sim N_p(0, K)$, with K the limit of $n^{-1} \sum_{i=1}^n K(x_i)$. Explain that $A_n(s) \rightarrow_d A(s) = U^t s - \frac{1}{2} s^t J s$.

(d) (xx then read off limits. proof of pudding lies in eating, i.e. applications, below. estimates $\hat{J}$ and $\hat{K}$. xx) $\hat{J} = -n^{-1} \partial^2 \ell_n(\hat{\theta}) / \partial \theta \partial \theta^t$, for estimating K , use the estimated scores $\hat{u}_i = u(Y_i | x_i, \hat{\theta})$, noting that these sum to zero. then use $\hat{K} = n^{-1} \sum_{i=1}^n \hat{u}_i \hat{u}_i^t$. sandwich $\hat{\Sigma} = \hat{J}^{-1} \hat{K} \hat{J}^{-1}$.

Ex. 5.56 *Linear regression: agnostic analysis.* When a regression model is used, without being fully correct, the general theory of Ex. 5.55 explains (i) how the ML estimators implicitly aim for the best parametric approximation, defined via Kullback–Leibler divergences weighted over the available covariate vectors, and (ii) how the approximate distributions are affected. Here we consider such consequences for the classic linear normal regression model. For a dataset of pairs (x_i, Y_i) , assume merely that $E Y_i = m(x_i)$, for some smooth $m(x)$ function, with i.i.d. errors ε_i with mean zero and variance σ_0^2 . The linear model approximates this via the standard $Y_i | x_i \sim N(x_i^t \beta, \sigma^2)$, for which we have found the ML estimators $\hat{\beta}$ and $\hat{\sigma}$ in Ex. 5.43.

(a) Explain that minimising the KL distance, as per Ex. 5.55, is the same as maximising

$$n^{-1} \sum_{i=1}^n \left\{ -\log \sigma - \frac{1}{2} \frac{1}{\sigma^2} E(Y_i - x_i^t \beta)^2 \right\} = -\log \sigma - \frac{1}{2} \frac{1}{\sigma^2} n^{-1} \sum_{i=1}^n [\{m(x_i) - x_i^t \beta\}^2 + \sigma_0^2].$$

Argue from this that the least false parameter $\beta_{0,n}$ is the one minimising $Q_n(\beta) = n^{-1} \sum_{i=1}^n \{m(x_i) - x_i^t \beta\}^2$, and show that this means $\beta_{0,n} = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n m(x_i) x_i$, where $\Sigma_n = n^{-1} \sum_{i=1}^n x_i x_i^t$. If the model is perfect, there is a true β_{true} for which $E(Y_i | x_i) = x_i^t \beta_{\text{true}}$, in which case $\beta_{0,n} = \beta_{\text{true}}$. Show next that the least false $\sigma_{0,n}$ is determined by $\sigma_{0,n}^2 = Q_{0,n} + \sigma_0^2$, where $Q_{0,n}$ is the minimum of $Q_n(\beta)$. If the linear mean is a good model, the $Q_{0,n}$ is small and the $\sigma_{0,n}$ aimed at by the ML estimator $\hat{\sigma}$ is not much bigger than σ_0 . If the linear mean is not a good approximation to the real $m(x)$, then the $\hat{\sigma}$ implicitly picks up both data variability and the difference between $x_i^t \beta$ and the real $m(x_i)$.

(b) For the two matrices J and K determining the behaviour of the ML estimators we have found in Ex. 5.43 that $\hat{J} = (1/\hat{\sigma}^2) \text{diag}(\Sigma_n, 2)$, and that the score vectors, computed at the least false values, can be written $u(y_i | x_i, \theta_0) = (1/\sigma_{0,n})(x_i z_i, z_i^2 - 1)^t$, in which $z_i = (y_i - x_i^t \beta_{0,n})/\sigma_{0,n}$. Note that we for the estimated standardised residuals $\hat{z}_i = (y_i - x_i^t \hat{\beta})/\hat{\sigma}$ have $\sum_{i=1}^n \hat{z}_i x_i = 0$ and $n^{-1} \sum_{i=1}^n \hat{z}_i^2 = 1$. For the K matrix, and with notation from Ex. 5.55, show that

$$n^{-1} \sum_{i=1}^n K(x_i) = n^{-1} \sum_{i=1}^n \frac{1}{\sigma_{0,n}^2} \text{Var} \begin{pmatrix} z_i x_i \\ z_i^2 - 1 \end{pmatrix}.$$

(xx round this off. two substories, depending on what is assumed for z_i . with $y_i = m(x_i) + \sigma_0 \varepsilon_i$, the z_i are not i.i.d., but we may cope with that too, via the right $\hat{K}$.

simplification if we postulate $m(x_i) = x_i^t \beta_0$, then involving only $\hat{\gamma}_3$ and $\hat{\gamma}_4$ for the $\hat{z}_i$. xx) a somewhat messy model-robust approach would use

$$\hat{K} = \frac{1}{\hat{\sigma}^2} n^{-1} \sum_{i=1}^n \begin{pmatrix} \hat{z}_i^2 x_i x_i^t & \hat{z}_i^3 x_i \\ \hat{z}_i^3 x_i^t & \hat{z}_i^4 - 1 \end{pmatrix}.$$

Ex. 5.57 *Logistic regression, Poisson regression: agnostic analysis.* When the regression model used does not necessarily hold up, ML estimators aim for the relevant least false parameters, the general theory of Ex. 5.55 shows that $\hat{\theta} \approx_d N(\theta_{0,n}, \hat{\Sigma}/n)$, for the relevant least false parameter and with estimated sandwich matrix $\hat{\Sigma}_n = \hat{J}^{-1} \hat{K} \hat{J}^{-1}$. Under model conditions, matters simplify, the underlying J_n and K_n matrices are equal, and we have the standard result $\hat{\theta} \approx_d N(\theta_0, \hat{J}^{-1}/n)$, used extensively in statistical software packages for a range of regression models. Here we check what the model agnostic setup leads to, for securing model robust inference, for standard logistic and Poisson regression models; see also the agnostic analysis for the linear regression model in Ex. 5.56.

(a) Consider logistic regression, as in Ex. 5.44, with $p_i = H(x_i^t \beta)$: show that $u(y_i | x_i, \beta) = (y_i - p_i)x_i$, and that

$$\hat{J}_n = n^{-1} \sum_{i=1}^n \hat{p}_i(1 - \hat{p}_i)x_i x_i^t, \quad \hat{K}_n = n^{-1} \sum_{i=1}^n (y_i - \hat{p}_i)^2 x_i x_i^t.$$

(b) Then look at Poisson regression, see Ex. 5.46, with $\mu_i = \exp(x_i^t \beta)$: show that $u(y_i | x_i, \beta) = (y_i - \mu_i)x_i$, with

$$\hat{J}_n = n^{-1} \sum_{i=1}^n \hat{\mu}_i x_i x_i^t, \quad \hat{K}_n = n^{-1} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i x_i^t.$$

If there is overdispersion, with $(y_i - \hat{\mu}_i)^2$ tending to be bigger than $\hat{\mu}_i$, this is picked up here, with sandwich matrix bigger than $\hat{J}^{-1}$. The potential error of applying straightforward Poisson regression, perhaps via standard packages, is that variability is underestimated, with confidence intervals becoming too narrow. For an illustration of this, see Story #28.

(c) With the Poisson-gamma overdispersion regression model studied in Ex. 5.46, where $Y_i | x_i$ has mean $\exp(x_i^t \beta)$ and variance $(1 + \tau) \exp(x_i^t \beta)$, show that $\hat{K}_n \rightarrow_{\text{pr}} (1 + \tau)J$, in terms of the limit J of $\hat{J}_n$. Deduce that if $\hat{\beta}$ is the ML estimator computed for the standard Poisson regression model, then $\sqrt{n}(\hat{\beta} - \beta_{0,n}) \rightarrow_d N_p(0, (1 + \tau)J^{-1})$.

Notes and pointers

(xx some pointers to more, including Empirical Likelihood, Hjort et al. (2009, 2018). mention Edwards (1974) on the history of likelihood, and Edwards (2025) for defense of Fisher. xx)

point to key references for GLM: Efron (2023), McCullagh and Nelder (1989), Johansen, point to Ex. 8.16 for Beta correction for Wilks.

point to Story #84, Wilks often better than symmetric normal approximation; Schweder and Hjort (2016, Ch. 7), Brazzale et al. (2007), Brazzale and Davison (2008).

t-ness etc., in Ex. 5.17: see Hjort (1994a) for more.

capture-recapture: point to Petersen (1896), Hjort and Petersen (1905), uses and much more, mention Lincoln, mention CloseKin and marine abundance assessment.

(xx Bhattacharyya, 1946, lower bounds more general than CR. xx)
 but beware Hodges, seemingly clashing with the $N(0, J(\theta)^{-1})$.
 (xx make sure we have something on board with BHHJ for regression models, placed
 after ML for regression. xx)
 [xx CR bound: In its simplest form, the inequality goes back to [Cramér \(1946\)](#) and
[Rao \(1945\)](#). xx]
 (xx M-estimators, Z-estimators, Huber, minimum divergence, point to [Basu et al. \(2011\)](#),
 more on minimum divergence. two-stage estimators, WalkerHjort24. xx)
 [xx least false: a term invented by Hjort, Hjort believes, see [Hjort \(1986b, 1992\)](#),
 and now used somewhat frequently in the literature. xx]
 Read more about risk functions in [DeGroot \(1970\)](#).
 For Ex. 6.6, point to [Hjort and Jones \(1996\)](#), [Schweder and Hjort \(2016\)](#), and Walk-
 erHjort24.
 [xx check and calibrate what's here and what's in Ch. 8, regarding CD things. xx]
 (xx we see that ML matches CR bounds, under model conditions. point to Hajek
 convolution theorem, and other characterisations. perhaps Hodges superefficient thing
 too. xx)
 (xx point to [Hjort \(2008\)](#), re ML and least false etc. xx)

I.6

Minimum divergence

The previous Ch. 5 has developed a versatile machinery for likelihood analysis in parametric models, with estimators, approximate distributions, deviances, Wilks theorems, etc. We have also seen the key role played by the Kullback–Leibler divergence $KL(g, f_\theta)$. In the present chapter the scope is widened to allow for other natural divergences, leading then to classes of minimum divergence estimators and inference recipes. These might have favourable properties not shared by the ML ones, regarding e.g. robustness, both with respect to potential outliers and datapoints being outside the normal scope of the model, and to model specification itself. We discuss in particular the Basu-Harris-Hjort-Jones estimation scheme, associated with a particular BHHJ divergence. The apparatus works for i.i.d. data, regression models, and beyond. Also included below is the useful tool of influence functions for estimators and functionals.

Key words: BHHJ estimators, Huber estimator, influence functions, minimum divergence, regression models, robustness

Our start setup for the chapter is that of i.i.d. data $Y_1, \dots, Y_n$ from a data-generating distribution with c.d.f. G and density g , and that a parametric model $f_\theta(y) = f(y, \theta)$ is put forward as an approximation to the g . Suppose in general terms that $h(y, \theta)$ is such that $H(\theta) = E_G h(Y, \theta) = \int g(y)h(y, \theta) dy$ has a unique minimum θ_0 , to be thought of as providing the best parametric approximation to g . Then the *minimum divergence estimator* $\hat{\theta}$ is the minimiser of the *divergence criterion function*

$$H_n(\theta) = n^{-1} \sum_{i=1}^n h(y_i, \theta). \quad (6.1)$$

The chief special case is that of $h(y, \theta) = -\log f(y, \theta)$, for which $H_n(\theta)$ is minus the normalised log-likelihood function $\ell_n(\theta)$, and $\hat{\theta}$ hence being the ML estimator. Thus the maximum likelihood machinery dealt with in Ch. 5 is a special case. The ML estimator, seen here to minimise minus the log-likelihood, aims for θ_0 , the maximiser of $\int g \log f_\theta dy$, i.e. the minimiser of the KL divergence $KL(g, f_\theta)$.

We might of course pedantically distinguish between ‘minimum divergence’ and ‘minimum criterion function’ estimators, but it is common practice to use the former for classes of minimisers of functions of the type above, even when the $h(y, \theta)$ criterion function is not immediately or necessarily related to a divergence. In many cases one may start with a bona fide divergence, say $d(g, f_\theta)$, measuring the distance from g to its approximation f_θ , and then discover that a suitable version of (6.1) matches that. The ML is again a

case in point, as the KL divergence may be written $\text{KL}(g, f_\theta) = \int g \log g \, dy - \int g \log f_\theta \, dy$. The first term does not involve the parameter, so maximising an empirical version of $\int \log f_\theta \, dG(y)$ does the minimum divergence job. This is indeed the ML procedure.

The object of the present follow-up chapter is to widen the scope, exhibiting other natural divergences. As long as minimising such a divergence, perhaps with some extra efforts, can be translated to a minimisation of a criterion function, as above, then the general machinery developed below will handle the central inference questions, regarding estimation, confidence, testing, limit distributions, etc. We shall also discover that what we learned in the likelihood chapter can be lifted, largely without laborious efforts, to these more general cases. Regression setups will also be worked through, with parameters constructed to minimise suitable $\sum_{i=1}^n h(y_i | x_i, \theta)$ criteria functions. In particular, there will be robust alternatives to linear regression and other regression models, where large or probabilistically outlying datapoints will be properly downplayed. mutatis
mutandis

Divergences and minimising them

Ex. 6.1 *Minimising L_2 distance.* Suppose i.i.d. data $Y_1, \dots, Y_n$ come from some data generating density g which we wish to approximate with some parametrically modelled f_θ . The L_2 distance between g and f_θ is

$$D(g, f_\theta) = \int (g - f_\theta)^2 \, dy = \int f_\theta^2 \, dy - 2 \int g f_\theta \, dy + a(g),$$

where $a(g)$ does not depend on θ .

(a) Use this to motivate what we may call the minimum L_2 estimator $\hat{\theta}$, the minimiser of $D_n(\theta) = \int f_\theta^2 \, dy - 2n^{-1} \sum_{i=1}^n f(Y_i, \theta)$. Operationally, this is easiest when there is a closed-form formula for $\int f_\theta^2 \, dy$, but numerical minimisation might be carried out even without this. Simulate 100 datapoints from a $\text{Gam}(a, b)$, where you choose (a, b) as you wish, and estimate these using this method.

(b) Carry out a similar simple experiment with 100 datapoints drawn from a normal, i.e. estimate the mean and standard deviation using the minimum L_2 method. Then do the following variation: push one of the 100 datapoints far away from the others, which will cause the traditional estimates of μ and σ to be off; show however that the minimum L_2 estimates are far less affected. – The minimum L_2 method is a special case of the BHHJ method, see Ex. 6.4.

Ex. 6.2 *Estimating means via minimum divergence.* Suppose again i.i.d. data $Y_1, \dots, Y_n$ stem from some data generating density g . The point to convey here is that key parameters can be seen as minimisers of natural distance functions, or divergences, leading to recipes for estimating them.

(a) For any distribution G , let $\xi(G)$ be the minimiser of $H(\xi) = E_G(Y - \xi)^2$, with estimator the minimiser $\hat{\xi}$ of the empirical version $H_n(\xi) = n^{-1} \sum_{i=1}^n (Y_i - \xi)^2$. Show, for this simple illustration of minimum divergence function methods, that $\xi(G)$ is the mean $E_G Y$ and that $\hat{\xi} = \bar{Y}$, the sample mean.

(b) For characterising and then estimating both mean and standard deviation, consider $h(y, \xi, \sigma) = (y - \xi)^2 + \{y^2 - (\xi^2 + \sigma^2)\}^2$, and let (ξ, σ) be the minimiser of the distance function $H(\xi, \sigma) = E_G h(Y, \xi, \sigma)$. Show that $\xi(G)$, $\sigma(G)$ are the mean and standard deviation, and work out expressions for $\hat{\xi}$, $\hat{\sigma}$, the minimisers of the empirical distance function $H_n(\xi, \sigma) = n^{-1} \sum_{i=1}^n h(Y_i, \xi, \sigma)$.

(c) For positive Y_i data, to be fitted to the $\text{Gam}(a, b)$, consider the function $h(y, a, b) = (y - a/b)^2 + (y^2 - (a/b^2 + a^2/b^2))^2$, constructed in view of a/b and a/b^2 being the formulae for the mean and variance for a Gamma. Explain that the empirical version of the distance function $H(a, b) = E_G h(Y, a, b)$ becomes $H_n(a, b) = n^{-1} \sum_{i=1}^n \{(Y_i - a/b)^2 + (Y_i^2 - (a/b^2 + a^2/b^2))^2\}$. Show that the minimum divergence function estimators $(\hat{a}, \hat{b})$ are equivalent to the moment estimators, worked with in Ex. 3.26, i.e. the solution to the two equations $\bar{Y} = a/b$ and $\hat{\sigma}^2 = a/b^2$.

(d) Consider $h_0(x) = x \arctan x - \frac{1}{2} \log(1 + x^2)$, and define θ_0 as the minimiser of $E_G h_0(Y - \theta)$. Show that $\hat{\theta}$, the minimiser of $H_n(\theta) = n^{-1} \sum_{i=1}^n h_0(Y_i - \theta)$, is also the unique solution to $\sum_{i=1}^n \arctan(Y_i - \theta) = 0$.

Ex. 6.3 *Minimum divergence function estimators: general setup.* We learn from Ex. 6.1–6.2 that classes of estimators may be formed via minimisation of suitable empirical distance functions; these estimate the corresponding minimisers of distance functions operating on the underlying distributions. We call these *minimum divergence function estimators* or *minimum distance estimators*. In general terms, for observations $Y_1, \dots, Y_n$ from some distribution G , consider a parameter $\theta_0 = \theta(G)$ defined as the minimiser of the function $H(\theta) = E_G h(Y, \theta)$, for a suitable $h(y, \theta)$. It is assumed that θ_0 thus defined, which may also be multidimensional, is the unique minimiser. The empirical version of $H(\theta)$ is $H_n(\theta) = n^{-1} \sum_{i=1}^n h(Y_i, \theta)$, so a natural estimator for θ_0 is $\hat{\theta} = \text{argmin}(H_n)$. In fact many important estimators are of this or related types, perhaps minimising somewhat more complicated random functions, as we shall see in this chapter. In exercises below we shall develop clear results for how the minimum divergence function estimators behave, under sets of natural assumptions, but the present exercise is meant to illustrate the basic construction via different types of examples.

(a) Explain that $H_n(\theta)$ can be written $\int h(y, \theta) dG_n(y)$, with G_n the empirical distribution, having mass $1/n$ at each datapoint; see Ex. 3.9. Explain why $H_n(\theta) \rightarrow_{\text{pr}} H(\theta)$ for each θ , and find the limit distribution of $\sqrt{n}\{H_n(\theta) - H(\theta)\}$. What we need, tended to in several exercises to follow, are conditions under which $\hat{\theta} = \text{argmin}(H_n)$ tends to $\theta_0 = \text{argmin}(H)$, along with a limit distribution.

(b) For one-dimensional Y_i , work through the details of $h(y, \theta) = (y - \theta)^2$. Then consider $h(y, \theta) = [\exp\{c(y - \theta)\} - 1 - c(y - \theta)]/c^2$, with c a balance parameter. Draw 100 datapoints from a normal $N(\theta, 1)$, with θ of your choice, and estimate θ in this minimum H_n fashion, for a few values of the balance parameter c . Show that c close to zero corresponds to the mean.

(c) Let generally $h(y, \theta) = \{r(y) - \theta\}^t V \{r(y) - \theta\}$, for some $r(y) = (r_1(y), \dots, r_p(y))$ and a symmetric positive definite matrix V . Show that $\theta_0 = E_G r(Y)$ and that $\hat{\theta} = n^{-1} \sum_{i=1}^n r(Y_i)$.

(d) Consider $h(y, \xi, \tau) = p(\tau) + \frac{1}{2}(y - \xi)^2/\tau^2$, where $p(\tau)$ is a smooth increasing function of $\tau > 0$. Find a recipe for computing the estimates $(\hat{\xi}, \hat{\tau})$ associated with the distance function $n^{-1} \sum_{i=1}^n h(Y_i, \xi, \tau)$. Check in particular the case of $p(\tau) = \log \tau$.

(e) There are connections here to the moment matching estimation method, as worked with in Ex. 3.25. For a one-parameter model first, with $EY = m(\theta)$, consider $h(y, \theta) = \{y - m(\theta)\}^2$. Show that the implied best parameter is for $m(\theta_0) = EY$, or $\theta_0 = m^{-1}(EY)$, and that the minimum divergence function method leads to solving $m(\theta) = \bar{Y}$. Generalise to the case of there being two parameters in the model, and then to the general vector parameter case.

connection to
M-estimators

minimum
divergence
function
estimators

BHHJ, maximum weighted likelihood, minimum quantile difference

Ex. 6.4 *The BHHJ density power divergence method.* Here we set up the basics for the so-called *density power divergence method*. It involves raising the density function $f(y, \theta)$ to some power a , as one of its ingredients. In the literature it is sometimes called the *BHHJ divergence method*. We shall see in later exercises that it amounts to a robust modification of the ML method.

from its
inventors Basu,
Harris, Hjort,
Jones

(a) For a density g , in what follows to be seen as the true underlying data-generating model, consider measuring the distance to an approximate $f_\theta(y) = f(y, \theta)$ density as

$$d_a(g, f_\theta) = \int \left\{ f_\theta^{1+a} - \left(1 + \frac{1}{a}\right) g f_\theta^a + \frac{1}{a} g^{1+a} \right\} dy, \quad (6.2)$$

with a a positive tuning parameter. Show that $d_a(g, f_\theta) \geq 0$, and that the distance is zero only when $g = f_\theta$ a.e. Note that this for $a = 1$ is the same as the L_2 distance of Ex. 6.1.

(b) Use Taylor expansion for f_θ^a and g^a for small a , to demonstrate that the integrand in (6.2) may be written

$$f_\theta - g + g \log(g/f_\theta) - a(g - f_\theta) \log f_\theta + \frac{1}{2} a g \{(\log g)^2 - (\log f_\theta)^2\} + O(a^2).$$

Hence show that for a small, we have $d_a(g, f_\theta) = \text{KL}(g, f_\theta) + O(a)$, with the Kullback–Leibler distance $\int g \log(g/f_\theta) dy$, assuming that the functions $g \log f_\theta$, $f_\theta \log f_\theta$, $g(\log f_\theta)^2$, $g(\log g)^2$ have finite integrals.

(c) Suppose $Y_1, \dots, Y_n$ are i.i.d. from some unknown g , and that we wish to estimate θ by making the distance $d_a(g, f_\theta)$ small. The point is now that the third term of d_a does not depend on θ , and that we may accurately estimate the two first, using

$$H_{n,a}(\theta) = \int f(y, \theta)^{1+a} dy - (1 + 1/a) n^{-1} \sum_{i=1}^n f(y_i, \theta)^a. \quad (6.3)$$

Show that this is also coming out of the general minimum divergence function apparatus of Ex. 6.3, with $h(y, \theta) = \int f_\theta^{1+a} dy - (1 + 1/a) f(y, \theta)^a$. Show indeed that $H_n(\theta)$ is an unbiased estimator of the two first terms of $d_a(g, f_\theta)$, and give an expression for its variance. We call the minimiser $\hat{\theta}$ of $H_n(\theta)$ the BHHJ estimator.

the BHHJ
estimator

(d) Assuming regularity, with derivatives allowed to be taken under the integrals, etc., show that the BHHJ estimator is the solution to

$$\int f_\theta^{1+a} u_\theta dy = n^{-1} \sum_{i=1}^n f(y_i, \theta)^a u(y_i, \theta) = 0,$$

with $u_\theta(y) = u(y, \theta)$ the score function. For small a , explain that this is close to the score equations determining the ML estimator.

(e) Return to the ML estimator illustration of Ex. 5.3, with $n = 100$ points drawn from the uniform, fitted to the $\theta y^{\theta-1}$ density. For some values of a , compute and plot the $H_{n,a}$ function, finding also the BHHJ estimate $\hat{\theta}_a$. After having computed this for a grid of a , plot the resulting $\hat{\theta}_a$ as a function of a , and comment.

Ex. 6.5 *Integrals for BHHJ estimation.* Using the BHHJ method of Ex. 6.4 for estimating the parameters of models $f(y, \theta)$, we are very much helped, algorithmically and numerically, by having formulae for the term $A(a) = \int f_\theta^{1+a} dy$.

(a) For the normal $N(\xi, \sigma^2)$, show that

$$\int f(y, \xi, \sigma)^{1+a} dy = (2\pi)^{-a/2} (1+a)^{-1/2} \sigma^{-a}.$$

Generalise to the multinormal case of $N_p(\xi, \Sigma)$, with

$$\int f(y, \xi, \Sigma)^{1+a} dy = (2\pi)^{-ap/2} (1+a)^{-p/2} |\Sigma|^{-a/2}.$$

(b) For the gamma model with density proportional to $y^{\alpha-1} \exp\{-\beta(y-y_0)\}$ for $y \geq y_0$, we find

$$\int_{y_0}^{\infty} f(y, \alpha, \beta)^{1+a} dy = \frac{\Gamma(\alpha + a\alpha - a)}{\Gamma(\alpha)^{1+a}} \frac{\beta^a}{(1+a)^{\alpha+a\alpha-a}}.$$

(c) For the log-normal model, where Y is such that $\log Y \sim N(\xi, \sigma^2)$: Show that

$$\int_0^{\infty} f(y, \xi, \sigma)^{1+a} dy = (2\pi)^{-a/2} \sigma^{-a} (1+a)^{-1/2} \exp\{-a\xi + \frac{1}{2}a^2\sigma^2/(1+a)\}.$$

(d) For the Weibull, with c.d.f. $1 - \exp\{-(y/\alpha)^\beta\}$ for $y \geq 0$, show that

$$\int_0^{\infty} f(y, \alpha, \beta)^{1+a} dy = \left(\frac{\beta}{\alpha}\right)^a \frac{\Gamma(1+a-a/\beta)}{(1+a)^{1+a-a/\beta}}.$$

(e) For robust estimation of the three-parameter t distribution, consider the density $f(y, \xi, \sigma, \nu) = g_\nu((y-\xi)/\sigma)(1/\sigma)$ for $Y = \xi + \sigma t_\nu$, with g_ν the t density with ν degrees of freedom. Here find that $\int f^{1+a} dy$ can be written

$$\frac{1}{\sigma^a} \int g_\nu^{1+a} dx = \frac{1}{\sigma^a} \frac{\Gamma((1+a)\nu+a/2)}{\Gamma((1+a)(\nu+1)/2)} \left\{ \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \right\}^{1+a} \frac{1}{(\nu\pi)^{a/2}}.$$

(f) For the two-parameter extreme value distribution worked with in Story #71, the c.d.f. is $G(y, a, \sigma) = 1 - (1 - ay/\sigma)^{1/a}$ and density $g(y, a, \sigma) = (1/\sigma)(1 - ay/\sigma)^{1/a-1}$, for a positive, and range $y \in [0, \sigma/a]$. Show that

$$A(b) = \int g(y, a, \sigma)^{1+b} dy = \frac{1}{\sigma^b} \frac{1}{1+b-ab}.$$

Ex. 6.6 *Maximum weighted likelihood estimation.* Another fruitful generalisation of the ML method, in addition to the BHHJ method of Ex. 6.4, starts from adding a weight function to the Kullback–Leibler $KL(g, f)$ divergence.

(a) Show that $KL(g, f)$ may be written $\int \{g \log(g/f) - (g-f)\} dy$, that the integrand is nonnegative, and is zero only if $g = f$ a.e. With any nonnegative weight function $w(y)$ on the sample space, deduce that

$$KL_w(g, f) = \int w(y) \left[g(y) \left\{ \log \frac{g(y)}{f(y)} - \{g(y) - f(y)\} \right\} \right] dy$$

is a divergence, i.e. nonnegative, and is zero only if $g = f$ on the support set of w . In particular, for g fixed and parametric $f(\cdot, \theta)$, show that $KL_w(g, f_\theta)$ can be expressed as $\int w(-g \log f_\theta + f_\theta)$ plus other terms not depending on θ .

(b) For data $Y_1, \dots, Y_n$ from a generating density g , to be approximated with a parametric $f(y, \theta)$, explain that the general minimum divergence function method leads to maximisation of

$$\ell_{n,w}(\theta) = \sum_{i=1}^n w(y_i) \log f(Y_i, \theta) - nB(\theta),$$

with $B(\theta) = \int w f_\theta dy$. We call the maximiser $\hat{\theta}_w$ the maximum weighted likelihood estimator, associated with weight function $w(y)$. The classic ML estimator then corresponds to constant weight function $w(y) = 1$. Properties of $\hat{\theta}_w$ are derived in Ex. 6.12.

(c) For an illustration, access the birthweight dataset for Oslo children, reported on in Story #22. For the boys and the girls, fit normal distributions, using maximum weighted likelihood, with weight function $w(y)$ being one on $[2.5, 4.5]$ kg and zero outside. Plot the estimated densities in a diagram. This requires programming a function for optimisation. Compare with curves estimated via full ML.

Ex. 6.7 *Quantile matching estimators.* Suppose i.i.d. $Y_1, \dots, Y_n$ are observed from some smooth density g , and that a parametric model f_θ is to be fitted. One may then compare the sample quantiles, say $Q_{n,1}, \dots, Q_{n,k}$ at levels $q_1 < \dots < q_k$, to those derived from the model, i.e. $F^{-1}(q_j, \theta)$.

(a) Simulated a normal dataset in your computer, say of size $n = 333$. For the nine deciles $1/10, \dots, 9/10$, compute the sample deciles $Q_{n,j}$, and explain why minimising criterion functions of the type $\sum_{j=1}^k w_j \{Q_{n,j} - F^{-1}(q_j, \theta)\}^2$ will be sensible estimation procedure. The w_j are positive weights, which can be taken equal or not, to reflect higher importance. Fit the normal distribution to your data in this way, using $F^{-1}(q_j, \xi, \sigma) = \xi + \sigma \Phi^{-1}(q_j)$. Experiment a bit with other quantiles and weights.

(b) Simulate another dataset, this time from the exponential distribution. Find the quantile function $F^{-1}(q, \theta)$. Estimate the θ by minimising $\sum_{j=1}^k w_j \{Q_{n,j} - F^{-1}(q_j, \theta)\}^2$, this time with $w_j = 1 + q_j$, i.e. weights becoming more important for growing quantiles. Play with your code a bit and comment on your findings.

The minimum divergence process and its limit

Ex. 6.8 *Minimum divergence function estimators: a limit process and main heuristics.* After having motivated and worked through particular instances of the minimum divergence function estimators, we now return to the general case, aiming to demonstrate both limiting normality and other associated results, finding recipes for large-sample approximations in the process. The aim of the present exercise is to go through the basic ideas, involving also two main heuristics. These then apply to large classes of estimators and associated minimum divergence function minima. To make these heuristics precise we need more conditions and details, to which we return in further exercises below. In particular, results derived here and in a few of the following exercises will apply immediately to ML estimators, BHHJ estimators, MWL estimators, via their definitions in Ex. 5.3, 6.4, 6.6.

So $\theta_0 = \theta_0(G)$ is the minimiser of $H(\theta) = E_G h(y, \theta)$, and $\hat{\theta}$ is its estimator, the minimiser of $H_n(\theta) = \int h(y, \theta) dG_n(y)$, with G_n the empirical distribution for an observed i.i.d. sample $Y_1, \dots, Y_n$ from G . To present the basic ideas and main heuristics we shall start with these regularity conditions: (i) The true $\theta_0 = \theta_0(G)$ is an inner point in its parameter space inside $\mathbb{R}^p$. (ii) The $h(y, \theta)$ is smooth in θ , in a neighbourhood around

check corresponding themes for the special case of ML in Ex. 5.10 and 5.11

θ_0 , with at least two derivatives, say $h'(y, \theta)$ (a vector, with components h'_a), $h''(y, \theta)$ (a matrix, with components $h''_{a,b}$). (iii) The matrix $J = E_G h''(Y, \theta_0)$ is finite and positive definite. (iv) The first derivative $h'(Y, \theta_0)$ has finite variance matrix K .

generalising the A_n process of Ex. 5.10

(a) A key idea is to work with the following random function. Write

$$\begin{aligned} A_n(s) &= n\{H_n(\theta_0 + s/\sqrt{n}) - H_n(\theta_0)\} \\ &= \sum_{i=1}^n \{h(Y_i, \theta_0 + s/\sqrt{n}) - h(Y_i, \theta_0)\} = U_n^t s + \frac{1}{2} s^t J_n s + r_n(s), \end{aligned}$$

with $U_n = (1/\sqrt{n}) \sum_{i=1}^n h'(Y_i, \theta_0)$, $J_n = n^{-1} \sum_{i=1}^n h''(Y_i, \theta_0)$. Show that $A_n(s)$ is properly defined for all large enough n , that U_n has mean zero and tends to $U \sim N_p(0, K)$, and that $J_n \rightarrow_{\text{pr}} J$.

the sandwich matrix

(b) It is also clear from the Taylor expansion argument that with moderate further regularity, the remainder term $r_n(s)$ will go to zero in probability, for each s . There is then convergence $A_n(s) \rightarrow_d A(s) = U^t s + \frac{1}{2} s^t J s$, for each s . Heuristic One is then to go from $A_n \rightarrow_d A$ to $\text{argmin}(A_n) \rightarrow_d \text{argmin}(A)$. Explain that this then leads to $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d -J^{-1}U$, which is a $N_p(0, J^{-1}KJ^{-1})$. We call this limit distribution variance matrix $J^{-1}KJ^{-1}$ the sandwich matrix.

(c) Similarly, Heuristic Two is to go from $A_n \rightarrow_d A$ to $\min(A_n) \rightarrow_d \min(A)$. Argue that this entails $W_n = 2n\{H_n(\theta_0) - H_n(\hat{\theta})\} \rightarrow_d W = U^t J^{-1}U$. Show for this quadratic form that its mean and variance are $\text{Tr}(J^{-1}K)$ and $2\text{Tr}(J^{-1}KJ^{-1}K)$.

(d) Consider the minimum L_2 method of Ex. 6.1, used first to estimate only the mean in the $N(\xi, 1)$ model, and then for estimating both parameters in the $N(\xi, \sigma^2)$. Set up expressions for the J and K .

Ex. 6.9 *Minimum divergence function estimators: from heuristics to proofs.* The general setup is as with the previous Ex. 6.8, with minimum divergence function estimator $\hat{\theta}$ minimising $H_n(\theta) = n^{-1} \sum_{i=1}^n h(Y_i, \theta)$, with observations coming from G . In addition to regularity conditions (i)–(iv) given there, we postulate (v), that also third derivatives of $h(y, \theta)$ exist, say $h'''_{a,b,c}(y, \theta)$, and that these have finite means in a neighbourhood around θ_0 . We use the process $A_n(s) = n\{H_n(\theta_0 + s/\sqrt{n}) - H_n(\theta_0)\}$ from the previous exercise.

(a) Explain first that the minimiser of A_n is $\alpha_n = \sqrt{n}(\hat{\theta} - \theta_0)$, where we shall also study the overall minimum $A_{n,\min} = A_n(\alpha_n)$ below. Let $B_n(s) = U_n^t s + \frac{1}{2} s^t J_n s$ be the quadratic approximation to A_n , with minimiser β_n and overall minimum $B_{n,\min} = \min\{B_n(s) : \text{all } s\}$. Show that

$$\begin{aligned} \beta_n &= -J_n^{-1}U_n \rightarrow_d -J^{-1}U \sim N_p(0, J^{-1}KJ^{-1}), \\ B_{n,\min} &= -\frac{1}{2}U_n^t J_n^{-1}U_n \rightarrow_d -\frac{1}{2}U^t J^{-1}U. \end{aligned}$$

So things are simple and clear for the quadratic approximation B_n ; we need to show that the same results obtain for the real thing, the A_n .

mutatis mutandis

(b) Mimick arguments from Ex. 5.10, there for ML estimation, to verify that $A_{n,\min} - B_{n,\min} \rightarrow_{\text{pr}} 0$ and that $\alpha_n - \beta_n \rightarrow_{\text{pr}} 0$. Conclude that

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &\rightarrow_d -J^{-1}U \sim N_p(0, J^{-1}KJ^{-1}), \\ 2n\{H_n(\theta_0) - H_n(\hat{\theta})\} &\rightarrow_d W = U^t J^{-1}U. \end{aligned} \tag{6.4}$$

Ex. 6.10 *The J , K , and sandwich matrices for the BHHJ method.* When fitting a parametric family $f(y, \theta)$ to observations, we have seen in Ex. 6.8-6.9 that the behaviour

of minimum divergence function estimators is characterised by (i) the implied least false parameter $\theta_0 = \theta_0(G)$ and then (ii) the crucial matrices J and K , leading to the sandwich matrix $\Sigma = J^{-1}KJ^{-1}$. As above G , with density g , is the data generating mechanism. Here we consider the BHHJ estimation method Ex. 6.4, with the estimator $\hat{\theta}$ minimising $H_{n,a}(\theta)$ of (6.3). Again, a is a positive finetuning parameter, and for small a the method is close to ML estimation. For the following points, we assume regularity conditions (i)-(v) are in place, from Ex. 6.8-6.9, and in particular that there is a unique minimiser $\theta_0 = \theta_{0,a} = \theta_{0,a}(G)$ of the $d_a(g, f_\theta)$ of (6.2).

(a) Explain that the theory developed in these earlier exercises implies that $\hat{\theta} \rightarrow_{\text{pr}} \theta_0$. In the following points, write for simplicity $f_0(y) = f(y, \theta_0)$, $u_0(y) = u(y, \theta_0)$, $i_0(y) = i(y, \theta_0)$. Show also that θ_0 is characterised as the solution to $\int f_\theta^{1+a} u_\theta dy = \int g f_\theta^a u_\theta dy$. For this vector, evaluated at θ_0 , we write

$$\xi_a = \int g f_0^a u_0 dy = \int f_0^{1+a} u_0 dy.$$

(b) As seen in Ex. 6.4, the BHHJ method corresponds to using $h(y, \theta) = \int f_\theta^{1+a} dy - (1 + 1/a)f(y, \theta)^a$ in the general minimum divergence function setup. Find the derivative $h'(y, \theta)$ and deduce that

$$K_a = (1 + a)^2 \left\{ \int g f_0^{2a} u_0 u_0^t - \xi_a \xi_a^t \right\},$$

Find then an expression for the second derivative $h''(y, \theta)$, and derive a formula for the J matrix:

$$J_a = (1 + a) \int \{f_0^{1+a} u_0 u_0^t + (f_0^{1+a} - g f_0^a)(i_0 + a u_0 u_0^t)\} dy.$$

Note that for estimating this J_a in practice we typically get it for free via the minimisation procedure, as the Hessian matrix, i.e. we then do not need the formula per se.

Ex. 6.11 BHHJ precision under model conditions. The BHHJ method is designed to do well when the parametric model used is not perfect. We may however also study its precision when the model actually holds. For matrices studied in Ex. 6.10 there are simplified expressions under model conditions, i.e. when $g(y) = f(y, \theta_0)$ for all y .

(a) Under such model conditions, and with f_0, u_0 notation from Ex. 6.10, show that

$$J_a = (1 + a) \int f_0^{1+a} u_0 u_0^t dy, \quad K_a = (1 + a)^2 \left(\int f_0^{1+2a} u_0 u_0^t dy - \xi_a \xi_a^t \right),$$

with a consequent simplified sandwich matrix $\Sigma_a = J_a^{-1} K_a J_a^{-1}$. Show that for $a \rightarrow 0$, $\Sigma_a \rightarrow J_0^{-1}$, with $J_0 = J(\theta_0)$ the Fisher information matrix for the model.

(b) We may now check the loss of efficiency of the BHHJ method, with tuning parameter a , compared to the ML method, under model conditions, by comparing $J_a^{-1} K_a J_a^{-1}$ to the inverse Fisher information matrix. Carry out the relevant computations for estimation θ in the exponential $\theta \exp(-\theta y)$ model. For the case of the $N(\xi, \sigma^2)$ model, check with Story #95. For small a , there is significant gain in robustness at a low cost in efficiency.

(c) For the following, assume

$$A = \int f_0 \log f_0 u_0 u_0^t dy, \quad B = \int f_0 (\log f_0)^2 u_0 u_0^t dy, \quad c = \int f_0 \log f_0 u_0 dy$$

are finite; A and B are $p \times p$ matrices and c a $p \times 1$ vector. Below we carry out Taylor expansions to order a^2 , and use $\doteq$ to indicate the consequent approximations for small a . Show that to this order

$$J_a \doteq (1+a)(J_0 + aA + \tfrac{1}{2}a^2B), \quad K_a \doteq (1+a)^2\{J_0 + 2aA + a^2(2B - cc^t)\}.$$

As a useful interlude, consider $p \times p$ symmetric matrices M and ε , where M is positive definite and ε is smaller, in the technical sense that the eigenvalues of $M^{-1}\varepsilon$ are smaller than 1 in absolute value. Show that

$$(M + \varepsilon)^{-1} = M^{-1} - M^{-1}\varepsilon M^{-1} + M^{-1}\varepsilon M^{-1}\varepsilon M^{-1} - \dots$$

Use this to reach

$$(J_0 + aA + \tfrac{1}{2}a^2B)^{-1} \doteq J_0^{-1} - aJ_0^{-1}AJ_0^{-1} + a^2(J_0^{-1}AJ_0^{-1}AJ_0^{-1} - \tfrac{1}{2}J_0^{-1}BJ_0^{-1}).$$

It is then a matter of bureaucratic algebra to reach an informative approximation to the sandwich matrix

$$\Sigma_a \doteq (J_0 + aA + \tfrac{1}{2}a^2B)^{-1}\{J_0 + 2aA + a^2(2B - cc^t)\}(J_0 + aA + \tfrac{1}{2}a^2B)^{-1}.$$

Show that this leads to $\Sigma_a \doteq J_0^{-1} + a^2D$, with $D = J_0^{-1}(B - cc^t - AJ_0^{-1}A)J_0^{-1}$. This indicates indeed that for small a , the efficiency loss is small.

Ex. 6.12 *The J , K , and sandwich matrices for the maximum weighted likelihood method.* Just as we for the BHHJ estimation method worked out the basics for J_a , K_a , $J_a^{-1}K_aJ_a^{-1}$ in Ex. 6.10, we here tend to their parallels for the maximum weighted likelihood method of Ex. 6.6. For a parametric model $f(y, \theta)$, and with a given weight function $w(y)$, it consists in maximising the weighted log-likelihood function $\ell_{n,w}(\theta) = \sum_{i=1}^n w(y_i) \log f(Y_i, \theta) - nB(\theta)$, with $B(\theta) = \int w(y)f(y, \theta) dy$. We assume the regularity conditions of Ex. 6.9 are in force.

(a) When observations $Y_1, \dots, Y_n$ stem from G , with density g , show that the MWL estimator $\hat{\theta}$ tends to $\theta_0 = \theta_0(G)$, the minimiser of the weighted KL divergence $\text{KL}_w(g, f_\theta)$. When g is not inside the parametric family, this least false parameter value depends also on the weight function. Show also that θ_0 is characterised as the solution to $\int w(y)\{g(y) - f(y, \theta_0)\}u(y, \theta_0) dy = 0$. As for Ex. 6.10 we write f_0, u_0, i_0 for the functions $f(y, \theta_0), u(y, \theta_0), i(y, \theta_0)$, and $\xi_0 = \int wgu_0 dy = \int wf_0u_0 dy$.

(b) Explain that maximising $\ell_{n,w}(\theta)$ is the same as minimising $n^{-1} \sum_{i=1}^n h(Y_i, \theta)$, with $h(y, \theta) = w(y) \log f(y, \theta) - B(\theta)$, or solving $n^{-1} \sum_{i=1}^n h'(Y_i, \theta) = 0$, with $h'(y, \theta) = w(y)u(y, \theta) - \xi(\theta)$, where $\xi(\theta) = \partial B(\theta)/\partial \theta = \int w(y)f(y, \theta)u(y, \theta) dy$. Show that $K_w = \int w^2 gu_0 u_0^t dy - \xi_0 \xi_0^t$. Under model conditions, we have $\xi_0 = 0$, and the above simplifies to $K_w = \int w^2 f_0 u_0 u_0^t dy$.

(c) Then show that $J_w = - \int wgi_0 dy + \partial^2 B(\theta_0)/\partial \theta \partial \theta^t$. Under model conditions, $J_w = - \int wf_0 i_0 dy$.

(d) (xx find sandwich, and compare with J^{-1} under model. xx)

Ex. 6.13 *Maximum weighted likelihood for multinomial models.* The weighted likelihood ideas of Ex. 6.6, 6.12 can also be used to estimate parameters in models for multinomial probabilities, allowing different weights of importance for different outcomes.

(a) Show first that $p \log(p/q) - (p - q)$ is always nonnegative, for p, q in $(0, 1)$. Now consider $(Y_1, \dots, Y_k)$ being multinomial with count n and probability vector $p = (p_1, \dots, p_k)$, as with Ex. 1.5. Then

$$d_w(p, p_\theta) = \sum_{j=1}^k w_j \{p_j \log(p_j/p_{j,\theta}) - (p_j - p_{j,\theta})\},$$

is a proper nonnegative divergence, from p to some p_θ , as long as the weights $w_1, \dots, w_k$ are nonnegative. With equal weights, show that this is the Kullback–Leibler divergence $\text{KL}(p, p_\theta)$. If $p_\theta = (p_1(\theta), \dots, p_k(\theta))$ is some postulated model, with θ of dimension r , say, show that the minimum divergence method amounts to maximising the weighted and modified log-likelihood

$$\ell_{n,w}(\theta) = n \sum_{j=1}^k w_j \{\hat{p}_j \log p_j(\theta) - p_j(\theta)\} = \sum_{j=1}^k w_j Y_j \log p_j(\theta) - n \sum_{j=1}^k w_j p_j(\theta),$$

with $\hat{p}_j = Y_j/n$. This generalises the usual ML method, which corresponds to equal weights.

(b) Let $p = (p_1, \dots, p_k)$ denote the true probability vector. Explain that the weighted likelihood estimator $\hat{\theta}_w$, maximising $\ell_{n,w}(\theta)$, tends in probability to the least false parameter maximising $\sum_{j=1}^k w_j \{p_j \log p_j(\theta) - p_j(\theta)\}$, i.e. to the minimiser of the weighted KL distance $d_w(p, p_\theta)$. Writing $u_j(\theta) = \partial \log p_j(\theta) / \partial \theta$, show that the least false parameter is also the solution to $\sum_{j=1}^k w_j \{p_j - p_j(\theta)\} u_j(\theta) = 0$, assumed here to be unique. Show furthermore that

$$U_n = n^{-1/2} \frac{\partial \ell_{n,w}(\theta_0)}{\partial \theta} = \sum_{j=1}^k w_j \sqrt{n} \{\hat{p}_j - p_j(\theta_0)\} u_j(\theta_0) \rightarrow_d U = \sum_{j=1}^k w_j Z_j u_j(\theta_0),$$

using results and notation from the multinomial CLT worked with in Ex. 2.51. Deduce that $U \sim N_r(0, K_w)$, with

$$K_w = \sum_{j=1}^k w_j^2 p_j u_j(\theta_0) u_j(\theta_0)^t - \xi_0 \xi_0^t, \quad \text{where } \xi_0 = \sum_{j=1}^k w_j p_j u_j(\theta_0).$$

(c) Next demonstrate that the normalised Hessian matrix $\hat{J}_{n,w} = -n^{-1} \partial^2 \ell_{n,w}(\hat{\theta}) / \partial \theta \partial \theta^t$ tends in probability to a well-defined J_w . In fact, writing $i_j(\theta) = \partial^2 \log p_j(\theta) / \partial \theta \partial \theta^t$, show that $J_w = J_w(\theta_0)$, where $J_w(\theta)$ is

$$-\frac{\partial^2}{\partial \theta \partial \theta^t} \sum_{j=1}^k w_j \{p_j \log p_j(\theta) - p_j(\theta)\} = \sum_{j=1}^k w_j [p_j(\theta) u_j(\theta) u_j(\theta)^t - \{p_j - p_j(\theta)\} i_j(\theta)],$$

with the expression simplifying under model conditions. Use this in conjunction with the general minimum divergence theory to establish that $\sqrt{n}(\hat{\theta}_w - \theta_0) \rightarrow_d J_w^{-1} U \sim N_r(0, \Sigma_w)$, with the sandwich matrix $\Sigma_w = J_w^{-1} K_w J_w^{-1}$.

Profiling, and minimum divergence in practice

Ex. 6.14 *Profiling a minimum divergence function, I.* For data $Y_1, \dots, Y_n$ from some distribution G , we have in Ex. 6.3 considered estimating a parameter $\theta_0 = \text{argmin}(H)$,

for $H(\theta) = E_G h(Y, \theta)$, by minimising the distance function $H_n(\theta) = n^{-1} \sum_{i=1}^n h(Y_i, \theta)$. For a focus parameter $\mu = k(\theta)$, a smooth function of $\theta = (\theta_1, \dots, \theta_p)$, it is useful to work with the associated profile function

$$H_{n,\text{prof}}(\mu) = \min\{H_n(\theta) : k(\theta) = \mu\}.$$

profiling a
distance
function

(a) As an introductory illustration, consider estimating the parameters (a, b) of a Gamma distribution via minimum L_2 , as in Ex. 6.1. Simulate 100 datapoints from a gamma; compute $(\hat{a}, \hat{b})$; and compute and display also the profile function $H_{n,\text{prof}}(\mu)$ for the mean parameter $\mu = a/b$.

(b) From the full model, with ensuing minimum divergence function estimator $\hat{\theta}$, show that the consequent $\hat{\mu} = k(\hat{\theta})$ becomes normal. With setup and notation as in Ex. 6.9, prove indeed that $\sqrt{n}(\hat{\mu} - \mu_0) \rightarrow_d c^t J^{-1} U$, with $c = \partial k(\theta_0)/\partial \theta$, and that the limit is a zero-mean normal with variance $c^t J^{-1} K J^{-1} c$. Show also that this $\hat{\mu}$ is identical to the minimiser of the profile function.

(c) In other words we already know the basic story for any focus parameter estimator $\hat{\mu}$, thanks to the delta method. It is however fruitful to work with representations and approximations stemming from examining the associated profile function. With methods from Ex. 6.9, show that

$$n\{H_n(\hat{\theta} + s/\sqrt{n}) - H_n(\hat{\theta})\} = \frac{1}{2} s^t J_n s + r_n(s), \quad \text{with } r_n(s) = O_{\text{pr}}(\|s\|^3/\sqrt{n}).$$

For the profiling, therefore, we must minimise this expression over all s such that $k(\theta) = k(\hat{\theta} + s/\sqrt{n}) = \mu$. With $k(\theta) = \hat{\mu} + c_n^t s/\sqrt{n} + O_{\text{pr}}(\|s\|^2/n)$, here writing $c_n = \partial k(\hat{\theta})/\partial \theta$, the essence is to minimise $\frac{1}{2} s^t J_n s$ under $c_n^t s = \sqrt{n}(\mu - \hat{\mu}) = x_n$, say. Appealing to Ex. 5.12, show that this minimum becomes $\frac{1}{2} x_n^2 / c_n^t J_n^{-1} c_n = n(\hat{\mu} - \mu)^2 / c_n^t J_n^{-1} c_n$. Fill in more details to prove that with $\mu_0 = k(\theta_0)$ the true parameter in question,

$$2n\{H_{n,\text{prof}}(\mu_0) - H_{n,\text{prof}}(\hat{\mu})\} = \frac{n(\hat{\mu} - \mu_0)^2}{c_n^t J_n^{-1} c_n} + o_{\text{pr}}(1) \rightarrow_d \frac{(c^t J^{-1} U)^2}{c^t J^{-1} c} \sim \kappa \chi_1^2,$$

with $\kappa = c^t J^{-1} K J^{-1} c / c^t J^{-1} c$.

Ex. 6.15 *ML and minimum divergence function estimators in practice.* In previous exercises we have learned that $\hat{\theta}$, the minimiser of $H_n(\theta) = n^{-1} \sum_{i=1}^n h(Y_i, \theta)$, is a natural estimator for θ_0 , the minimiser of $E_G h(Y, \theta)$, and that its distribution approaches normality. In order to use such results in practice, for testing, setting confidence intervals, etc., we need to estimate the two crucial matrices $J = E_G h''(Y, \theta_0)$ and $K = \text{Var}_G h'(Y, \theta_0)$. These comments apply in particular to ML estimation, with $J = -E_G i(Y, \theta_0)$ and $K = \text{Var}_G u(Y, \theta_0)$, see Ex. 5.10.

(a) Via technical details of Ex. 5.28, give conditions under which the natural estimators

$$\hat{J} = n^{-1} \sum_{i=1}^n h''(Y_i, \hat{\theta}) \quad \text{and} \quad \hat{K} = n^{-1} \sum_{i=1}^n h'(Y_i, \hat{\theta}) h'(Y_i, \hat{\theta})^t$$

are consistent for J and K . Deduce that when such hold, the empirical sandwich matrix $\hat{\Sigma} = \hat{J}^{-1} \hat{K} \hat{J}^{-1}$ is consistent for $\Sigma = J^{-1} K J^{-1}$. Note also that $\hat{J} = H_n''(\hat{\theta})$ is the Hessian matrix of the criterion function H_n , often computed directly when using numerical minimisation methods for finding $\hat{\theta}$ in the first place. Note that ML estimation is a special case of this setup; see Ex. 5.10 and 5.28.

this is
analogous to
ML machinery,
see Ex. 5.28

(b) Explain how confidence intervals for the components of θ may be read off from this. More generally, for any focus parameter $\mu = k(\theta)$, with estimator $\hat{\mu} = k(\hat{\theta})$, show that $\hat{\mu} \pm 1.96 \hat{\kappa}/\sqrt{n}$ is an approximate 95 percent interval for μ , where $\hat{\kappa}^2 = \hat{c}^t \hat{J}^{-1} \hat{K} \hat{J}^{-1} \hat{c}$, and $\hat{c} = \partial k(\hat{\theta})/\partial \theta$.

(c) To illustrate how the above machinery works in practice, simulate 100 points from the standard normal, and then estimate the two normal parameters (ξ, σ) via minimum L_2 , as in Ex. 6.3. Explain that this means minimising the empirical distance function

$$H_n(\xi, \sigma) = \int f(y, \xi, \sigma)^2 dy - \frac{2}{n} \sum_{i=1}^n f(y_i, \xi, \sigma) = \frac{\frac{1}{2}/\pi^{1/2}}{\sigma} - \frac{2}{n} \sum_{i=1}^n \frac{1}{\sigma} \phi\left(\frac{y_i - \xi}{\sigma}\right).$$

Carry out this minimisation using e.g. `nlm` in R, a non-linear minimisation algorithm, which finds both $(\hat{\xi}, \hat{\sigma})$ and the Hessian $\hat{J}$. Compute also $\hat{K}$, and find confidence intervals for ξ , for σ , and for $p(y_0) = \Pr(Y \geq y_0)$, with say $y_0 = 1.00$.

(d) Change one or two of your simulated datapoints to somewhat far-off values, e.g. $y_{99} = d$ and $y_{100} = d$, with $d = 5.00$ (which indeed is really far off for the standard normal). Observe what then happens to the ordinary ML estimators, and compare with what happens with the minimum L_2 estimators. The point is the the minimum L_2 method is much more robust than the ML method.

(e) (xx profiling too, to illustrate, for $p(y_0)$. again handled by the general theory. intervals need $\kappa = c^t J^{-1} K J^{-1} c / c^t J^{-1} c$. xx)

Ex. 6.16 BHHJ analysis in practice. IN: We have seen how the behaviour of the model robust BHHJ estimators, defined in Ex. 6.4, is described via the two crucial matrices J_a and K_a worked with in Ex. 6.10. Here we go into the details of their estimation. The setting is having i.i.d. data $Y_1, \dots, Y_n$ from some density g , fitted to a parametric $f(y, \theta)$ by minimising $H_{n,a}(\theta)$ of (6.3). We let $\hat{\theta}$ be this BHHJ estimator, computed for the given balance parameter a . Below, we write $\hat{u}_i = u(Y_i, \hat{\theta})$ and $\hat{f}_i = f(Y_i, \hat{\theta})$.

(a) Show that $\hat{\xi}_a = n^{-1} \sum_{i=1}^n \hat{f}_i^a \hat{u}_i$ is equal to $\int f(y, \hat{\theta})^{1+a} u(y, \hat{\theta}) dy$, and that it is consistent for the ξ_a defined in Ex. 6.10.

(b) Show next that $\hat{J}_a = H''_{n,a}(\hat{\theta})$, the Hessian matrix associated with minimisation of the criterion function, is consistent for the J_a matrix.

(c) Show that

$$\hat{K}_a = (1+a)^2 \left\{ n^{-1} \sum_{i=1}^n \hat{f}_i^{2a} \hat{u}_i \hat{u}_i^t - \hat{\xi}_a \hat{\xi}_a^t \right\} = (1+a)^2 n^{-1} \sum_{i=1}^n (\hat{f}_i^a \hat{u}_i - \hat{\xi}_a) (\hat{f}_i^a \hat{u}_i - \hat{\xi}_a)^t$$

is consistent for K_a .

Divergences in regression models

(xx fill in two or three exercises, BHHJ regression. xx)

Influence functions

Ex. 6.17 Influence functions. For a distribution function F , consider some associated parameter, say $\theta = T(F)$, with T the appropriate functional mapping the distribution to the parameter value in question. Examples include the mean, the standard deviation,

the skewness, the interquartile range, a threshold probability. The *influence function* for $\theta = T(F)$ is a very useful quantity, as we shall see. It is defined as

$$\text{IF}(F, y) = \lim_{\varepsilon \rightarrow 0} (1/\varepsilon) \{T((1-\varepsilon)F + \varepsilon\delta(y)) - T(F)\}. \quad (6.5)$$

Here $\delta(y)$ is the measure putting full mass 1 at the point y , and $(1-\varepsilon)F + \varepsilon\delta(y)$ the consequent mixture distribution. A variable Y_ε drawn from this mixture is from F with probability $1-\varepsilon$ and is equal to y with probability ε .

(a) Consider $\theta(F) = E_G h(Y) = \int h(y) dF(y)$, the mean of $h(Y)$. Show that $\text{IF}(F, y) = h(y) - \theta(F)$. In particular, the influence is bounded when h is, but unbounded e.g. in the case of the plain mean $h(y) = y$, which signifies a potential lack of robustness of this mean parameter functional $\theta = E_G Y$.

(b) Then consider the class of smooth functions of means. For mean type parameters $\gamma_1 = E_G h_1(Y), \dots, \gamma_k = E_G h_k(Y)$, let $\theta = T(F) = A(\gamma_1(F), \dots, \gamma_k(F))$, where $A(u_1, \dots, u_k)$ is smooth in a neighbourhood of $(\gamma_1(F), \dots, \gamma_k(F))$. Show that this parameter has influence function

$$\begin{aligned} \text{IF}(F, y) &= c_1(F) \text{IF}_{\gamma_1}(F, y) + \dots + c_k(F) \text{IF}_{\gamma_k}(F, y) \\ &= c_1(F) \{h_1(y) - \gamma_1(F)\} + \dots + c_k(F) \{h_k(y) - \gamma_k(F)\}, \end{aligned}$$

with $c_j(F)$ is the partial derivative $\partial A(u_1, \dots, u_k)/\partial u_j$, evaluated at $(\gamma_1(F), \dots, \gamma_k(F))$.

(c) Writing $\mu_F = E_G Y$ for the mean, show for the variance parameter $\sigma_F^2 = E_G Y^2 - \mu_F^2$ that its influence function becomes

$$\text{IF}(F, y) = -2\mu_F(y - \mu_F) + y^2 - E_G Y^2 = (y - \mu_F)^2 - \sigma_F^2.$$

Then for the standard deviation parameter $\sigma(F)$ itself, show that its influence function becomes

$$\text{IF}_\sigma(F, y) = \frac{1}{2}(1/\sigma_F) \{(y - \mu_F)^2 - \sigma_F^2\}.$$

(d) For a given parametric family $f(y, \theta)$, consider the ML functional $T(F)$, mapping a given F with density f to the least parameter value $\theta_0 = \theta_0(F)$, the minimiser of the information distance $\text{KL}(f, f(\cdot, \theta))$, or the maximiser of $\int \log f(y, \theta) dF(y)$. With F_n the empirical distribution of the data, placing probability $1/n$ on each data point, cf. Ex. 3.9. Show that $T(F_n)$ is the ML estimator, and that its influence function becomes $\text{IF}(F, y) = J^{-1}u(y, \theta_0)$.

Ex. 6.18 Influence for quantiles. Assume the c.d.f. F has a smooth density f .

(a) Consider $\mu = F^{-1}(\frac{1}{2})$, the median. Show that the influence function $\text{IF}(F, y)$ becomes $-\frac{1}{2}/f(\mu)$ for $y \leq \mu$ and $\frac{1}{2}/f(\mu)$ for $y > \mu$. Verify that the influence function has mean zero and variance $\frac{1}{4}/f(\mu)^2$.

(b) More generally, let $\mu_q = F^{-1}(q)$, for some quantile level $q \in (0, 1)$. Show that $\text{IF}(F, y)$ is $-(1-q)/f(\mu_q)$ for $y \leq \mu_q$ and $q/f(\mu_q)$ for $y > \mu_q$. Comment on the fact that these are bounded.

(c) Then consider a smooth function of quantiles, say $\gamma(F) = A(Q_1(F), \dots, Q_k(F))$, where $Q_j(F) = F^{-1}(q_j)$. Show that its influence function is $\text{IF}(F, y) = c_1(F) \text{IF}_1(F, y) + \dots + c_k(F) \text{IF}_k(F, y)$, in terms of $\text{IF}_j(F, y)$ the influence function of $F^{-1}(q_j)$ calculated as above, and where $c_j(F) = \partial A(Q_1, \dots, Q_k)/\partial Q_j$ computed at $(F^{-1}(q_1), \dots, F^{-1}(q_k))$. As an illustration, find and graph the influence function for $\gamma(F) = F^{-1}(0.90) - F^{-1}(0.10)$.

Ex. 6.19 *Influence function for the maximum likelihood functional.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from some underlying c.d.f. G with density g , and that the data are fitted via ML to a smooth parametric model $f(y, \theta)$. We have seen (xx pointer xx) that $\hat{\theta}_{\text{ml}} \rightarrow_{\text{pr}} \theta_0$, the least false parameter value, minimising $\text{KL}(g, f_\theta)$. This is the ML functional, mapping a c.d.f. G to the KL minimiser $\theta_0 = V(G)$.

(a) Consider first the exponential family case, studied in Ex. 1.57 and 5.38, where $f(y, \theta) = \exp\{\theta^t T(y) - k(\theta)\} h(y)$. We have seen that the ML estimator $\hat{\theta}_{\text{ml}}$ maximises $\theta^t \bar{T} - k(\theta)$, and is also the solution to $\bar{T} = \xi(\theta)$, with $\xi(\theta) = E_\theta T(Y) = \partial k(\theta) / \partial \theta$. It is hence a smooth function of sample means, where Ex. 6.17 applies. Show that $\text{IF}(G, y) = J^{-1}\{T(y) - \xi(\theta_0)\}$, with $J = \partial^2 k(\theta_0) / \partial \theta \partial \theta^t$.

(b) There is a clear enough influence function also in the general case. Show that under mild conditions, including a finite information matrix $J = -E_g \partial^2 \log f(Y, \theta_0) / \partial \theta \partial \theta^t$ and score variance matrix $K = \text{Var}_g u(Y, \theta_0)$, and θ_0 being an inner point in its parameter region, we have

$$\text{IF}(G, y) = J^{-1} u(y, \theta_0).$$

(c) Find the influence function $\text{IF}(G, y)$ for the exponential model with density $\theta \exp(-\theta y)$, and plot it. For the normal (ξ, σ) case, show that

$$\text{IF}(G, y) = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \frac{1}{\sigma} \begin{pmatrix} z \\ z^2 - 1 \end{pmatrix} = \sigma \begin{pmatrix} z \\ \frac{1}{2}(z^2 - 1) \end{pmatrix} \quad \text{where } z = (y - \xi) / \sigma.$$

For a situation outside the exponential family, find and plot also the influence function for the one-parameter Weibull model, with c.d.f. $1 - \exp(-y^\theta)$.

Ex. 6.20 *An estimator represented via its influence function.* IN: Consider an i.i.d. sequence $Y_1, Y_2, \dots$ from F , with $\theta = T(F)$ a parameter of interest. It may be estimated nonparametrically using $\hat{\theta} = T(F_n)$, with F_n the empirical distribution. Here we work towards a representation of $\hat{\theta} - \theta = T(F_n) - T(F)$ in terms of the influence function.

(a) Consider the case of $\theta = A(\gamma(F))$, where $\gamma(F) = E_G h(Y) = \int h dF$. Show that $\hat{\theta} = T(F_n)$ is equal to $A(\bar{h})$, with $\bar{h} = \int h dF_n = n^{-1} \sum_{i=1}^n h(y_i)$. Assuming $A(u)$ smooth, with two derivatives, show that

$$\hat{\theta} = A(\gamma_0) + A'(\gamma_0)(\bar{h} - \gamma_0) + \frac{1}{2} A''(\gamma_0)(\bar{h} - \gamma_0)^2 + o_{\text{pr}}(1/n),$$

with $\gamma_0 = \gamma(F)$. Deduce that $E \hat{\theta} = \theta + \frac{1}{2} A''(\gamma_0) \tau^2 / n + o(1/n)$, in terms of $\tau^2 = \text{Var } h(Y_i)$, and that $\hat{\theta} - \theta = n^{-1} \sum_{i=1}^n \text{IF}(F, y_i) + b/n + o_{\text{pr}}(1/\sqrt{n})$, where $b = \frac{1}{2} A''(\gamma_0) \tau^2$.

(b) Generalise to the case of $T(F) = A(\gamma_1(F), \dots, \gamma_p(F))$ being a smooth function of several means, as studied also in Ex. 6.17, with $\gamma_j(F) = \int h_j dF$. Show that $E \hat{\theta} = \theta + b/n + o(1/n)$, with $b = \frac{1}{2} \text{Tr}(A''(\gamma_0) K)$, with K the variance matrix of $(h_1(Y_i), \dots, h_p(Y_i))^t$, and $A''(\gamma_0)$ the second order derivative matrix of A , computed at γ_0 . Show that

$$\hat{\theta} - \theta = T(F_n) - T(F) = n^{-1} \sum_{i=1}^n \text{IF}(F, Y_i) + b/n + \varepsilon_n, \quad (6.6)$$

with $\varepsilon_n = o_{\text{pr}}(1/\sqrt{n})$, i.e. small enough to have $\sqrt{n} \varepsilon_n \rightarrow_{\text{pr}} 0$.

(c) Use the above to find the limit distribution of $\sqrt{n}(\hat{\theta} - \theta)$. This gives a new and partly simpler proof of things proved in Ex. 2.50.

(d) (xx fix this; used for AIC argument xx) The powerful representation (6.6) actually holds quite generally, as long as $T(F)$ is a moderately smooth functional (xx find refs, Shao (1991), Jullum and Hjort (2017) xx), though with no easy general formula for the bias component b/n . Deduce that $\sqrt{n}(\hat{\theta} - \theta)$ has the limit distribution $N(0, \kappa^2)$, with κ^2 the variance of $\text{IF}(Y_i, \theta)$, and with the bias part b/n disappearing in this normal limit. Show therefore that for the ML estimator, in any smooth regular parametric family with the least false parameter θ_0 being an inner point in the parameter region (xx pointer to above xx), one has the representation

$$\hat{\theta}_{\text{ml}} - \theta_0 = J^{-1}U_n + o_{\text{pr}}(n^{-1/2}), \quad \text{with } U_n = n^{-1} \sum_{i=1}^n u(Y_i, \theta_0).$$

Ex. 6.21 Robust estimation of location. Consider data $Y_1, \dots, Y_n$ from a symmetric density on the line, with unknown location parameter ξ . We may write the density as $f(y, \xi) = f_0(y - \xi)$, with f_0 being symmetric around zero, but otherwise taken unknown. With c a positive fine-tuning parameter, the Huber estimator $\hat{\xi}_c$ is the minimiser of $H_{n,c}(\xi) = n^{-1} \sum_{i=1}^n H_c(y_i - \xi)$, for the particular function

$$H_c(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq c, \\ c(|u| - \frac{1}{2}c) & \text{if } |u| > c. \end{cases}$$

(a) Show that $H_{n,c}(\xi)$ is monotone and smooth, and explain that $\hat{\xi}_c$ is also the unique solution to the equation $\sum_{i=1}^n \psi_c(y_i - \xi) = 0$, with $\psi_c(u)$ being u inside $[-c, c]$, then $-c$ to the left and c to the right of that window. Invent a dataset and construct an algorithm that finds $\hat{\xi}_c$, and experiment a little with different c . Explain that the estimator for big c is equal to the sample mean $\bar{y}$, and for small c equal to the sample median M_n .

(b) Note that the general methods and results of Ex. 6.8 and 6.9 do not work here, since the criterion function $H_c(u)$ is continuous, but not with a smooth derivative. Show that $H_c(u)$ is convex, however, so the convex minimisation methods of Ex. 5.33 apply. Work first through the details of the following:

$$H_c(y_i - \xi_0 - \varepsilon) - H_c(y_i - \xi_0) = \begin{cases} -(y_i - \xi_0)\varepsilon + \frac{1}{2}\varepsilon^2 & \text{if } y_i - \xi_0 \in (-c, c), \\ -\varepsilon & \text{if } y_i - \xi_0 > c, \\ \varepsilon & \text{if } y_i - \xi_0 < -c, \end{cases}$$

for small ε . There is a minor problem when $y_i - \xi_0$ is close to $-c$ or c , but these cases do not disturb the main argument. Use this to reach the representation

$$\begin{aligned} A_{n,c}(s) &= nH_{n,c}(\xi_0 + s/\sqrt{n}) - nH_{n,c}(\xi_0) \\ &= \sum_{i=1}^n \{H_c(y_i - \xi_0 - s/\sqrt{n}) - H_c(y_i - \xi_0)\} = -D_{n,c}s + \frac{1}{2}J_{n,c}s^2 + r_{n,c}(s), \end{aligned}$$

in which $D_{n,c} = n^{-1/2} \sum_{i=1}^n \psi_c(y_i - \xi_0)$, $J_{n,c} = n^{-1} \sum_{i=1}^n I\{|Y_i - \xi_0| < c\}$, and the remainder term goes pointwise to zero in probability. Explain that $\psi_c(Y_i - \xi_0)$ has mean zero, under the symmetry assumption, so the CLT applies, with $D_{n,c} \rightarrow_d D_c \sim N(0, \kappa_c^2)$, where $\kappa_c^2 = \int \psi_c(u)^2 f_0(u) du$.

(c) Deduce from the above, partly as per Ex. 5.33, that $A_n(s) \rightarrow_d A(s) = -D_c s + \frac{1}{2}J_c s^2$, with $J_c = \Pr(|Y - \xi_0| \leq c) = 2F_0(c) - 1$. Explain from this again that this entails (i) $\sqrt{n}(\hat{\xi}_c - \xi_0) \rightarrow_d D_c/J_c$, a normal with variance $v_\tau^2 = \kappa_c^2/J_c^2$, and (ii)

$$2n\{H_{n,c}(\xi_0) - H_{n,c}(\hat{\xi}_c)\} \rightarrow_d D_c^2/J_c \sim v_\tau^2 \chi_1^2.$$

Huber
estimator,
Huber function

(d) (xx round off. xx) estimating τ_c and J_c and $v_\tau = \kappa_c/J_c$. compute for the normal case. new default value $c_{1.05} = 1.023$, making intervals 1.05 bigger than the optimal size, under normality. do the influence function. do the ‘Huber model’ that we invent, where Huber is the ML. two parameters, even three, can read off $\ell_{n,\text{prof}}(c)$. point to nice Story #96.

Ex. 6.22 *Influence functions for BHHJ and for weighted likelihood estimators.* For some smooth parametric model $f_\theta(y) = f(y, \theta)$, consider the BHHJ estimation method of Ex. 6.4 and 6.10. The setup involves the data-generating density g , and for the given tuning parameter a , the defining parameter is $\theta_0 = \theta_{0,a}$, the minimiser of $\int f_\theta^{1+a} dy - (1 + 1/a) \int g f_\theta^a dy$, also the solution to $\int (f_\theta^{1+a} - g f_\theta^a) u_\theta dy = 0$. Here $u_\theta(y) = u(y, \theta)$ is the score function, and we will also need the information function $i_\theta(y) = i(y, \theta)$ below. As in Ex. 6.10 we use f_0, u_0, i_0 for these functions at θ_0 .

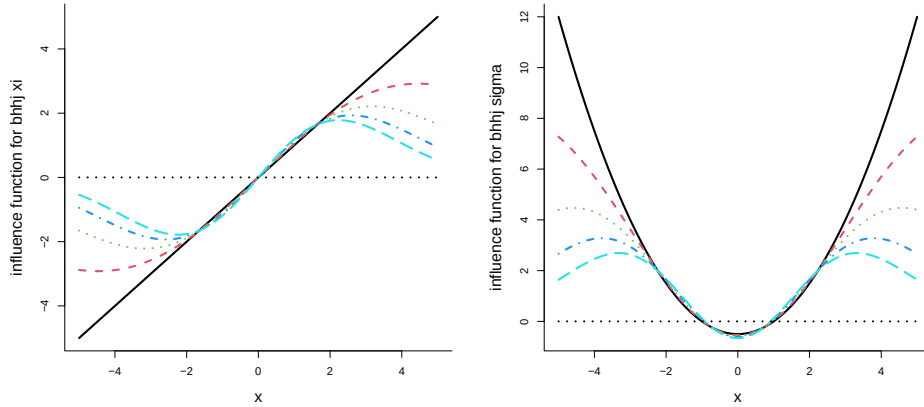


Figure 6.1: Influence functions for the BHHJ estimator, for the normal (ξ, σ^2) model, for values $a = 0, 0.05, 0.10, 0.15, 0.20$; the full black curves are for the ML case $a = 0$. Left panel: for $\hat{\xi}_a$; right panel: for $\hat{\sigma}_a$.

(a) Explain that to find the influence function $IF_a(G, y)$, we need for a small ε to assess the solution $\theta_\varepsilon = \theta_0 + \delta$ to the equation

$$\int f_{\theta_0+\delta}^{1+a} u_{\theta_0+\delta} dy = (1 - \varepsilon) \int g f_{\theta_0+\delta}^a u_{\theta_0+\delta} dy + \varepsilon f(y, \theta_0 + \delta)^a u(y, \theta_0 + \delta).$$

The programme is now for the given small ε to carry out first order Taylor expansion for δ small and then solve for δ . Some details are as follows: show first

$$f_{\theta_0+\delta}^a \doteq f_0^a(1 + a\delta^t u_0), \quad f_{\theta_0+\delta}^{1+a} \doteq f_0^{1+a}\{1 + (1+a)\delta^t u_0\}, \quad u_{\theta_0+\delta} \doteq u_0 + i_0\delta.$$

Again with $\xi_a = \int f_0^{1+a} u_0 du = \int g f_0^a u_0 dy$, use this to write the left and right hand sides of the equation as

$$\begin{aligned} L &\doteq \xi_a + \left\{ (1+a) \int f_0^{1+a} u_0 u_0^t dy + \int f_0^{1+a} i_0 dy \right\} \delta, \\ R &\doteq \xi_a + \left\{ \int g f_0^a (u_0 u_0^t + i_0) dy \right\} \delta + \varepsilon \{ f(y, \theta_0)^a u(y, \theta_0) - \xi_a \}. \end{aligned}$$

Conclude that $\text{IF}_a(G, y) = M_a^{-1}\{f_0(y)^a u_0(y) - \xi_a\}$, with the matrix

$$M_a = (1 + a) \int f_0^{1+a} u_0 u_0^t dy + \int f_0^{1+a} i_0 dy - \int g f_0^a (u_0 u_0^t + i_0) dy.$$

Check with the J_a matrix of Ex. 6.10 that in fact $M_a = J_a/(1 + a)$, so that

$$\text{IF}_a(G, y) = (1 + a) J_a^{-1} \{f(y, \theta_0)^a u(y, \theta_0) - \xi_a\}.$$

Explain that as $a \rightarrow 0$, the limit is $J_0^{-1} u(y, \theta_0)$, the influence function for the ML estimation method. Show also that under conditions, M_a simplifies to $\int f_0^{1+a} u_0 u_0^t dy$.

(b) Consider the case of the normal $N(\xi, \sigma^2)$. Under model conditions, find formulae for and graph the influence functions $\text{IF}_{1,a}(y)$ and $\text{IF}_{2,a}(y)$, for the ξ and the σ , at the standard position $(\xi, \sigma) = (0, 1)$, for a equal to 0, 0.05, 0.10, 0.15, 0.20. Construct a version of Figure 6.1. Note that the BHHJ influence functions mimic those for the ML (i.e. for $a = 0$) for the main expected data range, but that they sensibly redescend down to zero for values far away, via density downweighting. Show in particular that the influence functions are bounded, for each $a > 0$. For illustration of how the BHHJ successfully deals with outliers (without needing to identify them), and for seeing that very little efficiency is lost for small a , see Story #95. Formulae needed here include these, which you should prove, in addition to $\xi_{1,a} = 0$:

$$\begin{aligned} \xi_{2,a} &= -\sigma^{-a} (2\pi)^{-a/2} a / (1 + a)^{3/2}, \\ M_{11,a} &= (1/\sigma^2) \sigma^{-a} (2\pi)^{-a/2} / (1 + a)^{3/2}, \\ M_{22,a} &= (1/\sigma^2) \sigma^{-a} (2\pi)^{-a/2} \{3/(1 + a)^2 - 2/(1 + a) + 1\}. \end{aligned}$$

(c) Carry out similar analysis for the Gamma (α, β) model.

(d) Consider then the MWL method of Ex. 6.6 and 6.12. For a given weight function $w(y)$, the parameter aimed at for the MWL estimator is the maximiser $\theta_0 = \theta_{0,w}$ of $\int w(g \log f_\theta - f_\theta)$. We need $J_w = \int w f_0 u_0 u_0^t dy + \int w(f_0 - g) i_0 dy$ and $\xi_w = \int w f_0 u_0 dy = \int w g u_0 dy$, computed at θ_0 . Show that the influence function becomes

$$\text{IF}(G, y) = J_w^{-1} \{w(y) u(y, \theta_0) - \xi_w\}.$$

Leftovers and further material

Ex. 6.23 *Minimum dispair estimators.* (xx do this, in crisp manner. xx) The *Hellinger distance* between densities g and f_θ is

$$d(g, f_\theta) = \int (g^{1/2} - f_\theta^{1/2})^2 dy = 2 - 2 \int (g f_\theta)^{1/2} dy = 2 - 2 E_g (f_\theta/g)^{1/2} dy.$$

With an i.i.d. sample $Y_1, \dots, Y_n$ from g , setting up an empirical criterion for estimating the Hellinger minimiser θ_0 , is rather more complicated than for the KL and BHHJ and other divergences. We give one version below, thereby defining an estimator $\hat{\theta}_H$ with good properties.

(a) For estimating $\int g(f_\theta/g)^{1/2} dy$, we go for

$$H_n(\theta) = n^{-1} \sum_{i=1}^n \left\{ \frac{f(Y_i, \theta)}{g_n(Y_i)} \right\}^{1/2},$$

in terms of a pilot density estimator $g_n(y)$. Show that the two first derivatives become

$$H'_n(\theta) = n^{-1} \sum_{i=1}^n \frac{1}{2} \left\{ \frac{f(Y_i, \theta)}{g_n(Y_i)} \right\}^{1/2} u(Y_i, \theta),$$

$$H''_n(\theta) = \dots$$

(b) here we do $g_n(y) = f(y, \hat{\theta}, \hat{\gamma})$, a wider model than the one we focus on. full efficiency at the model, but more robust. we don't do more than this.

(c)

Notes and pointers

(xx various. xx) BHHJ and JHHB. Huber. The Basu book. Robustness. Influence functions here and there. Mutatis mutandis.

EL and HEL, with IVK and with ID, [Hjort et al. \(2009, 2018\)](#), [Dæhlen and Hjort \(2025\)](#).

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, 194

Euler, L., 38

Ibsen, H., 23

Røislien, J., 19

Sverdrup, E., 143

Tolstoï, L.N., 23

Subject index

- absolutely summable sequence, 49
- ancillary statistic, 130
- Borel–Cantelli lemma, 45
- bounded in probability, 56
- Cauchy–Schwarz inequality, 21
- conditional sufficiency, 129
- consistency of estimators, 44
- Cramér–Wold theorem, 66
- differentiability in quadratic mean, 176
- Factorisation theorem, 127
- first Bartlett identity, 151
- floor function, 50
- Hölder’s inequality, 21
- Jensen’s inequality, 20
- Kullback–Leibler distance, 149
- kurtosis, 49
- law of large numbers, weak, 50
- Law of Small Numbers, 47
- Levy’s continuity theorem, 66
- Lindeberg CLT, 61
- Markov’s inequality, 45
- method of moments, 98
- Mills ratio, 45
- minimal sufficient statistic, 129
- multivariate CLT, 66
- natural parameter region, 35
- probability integral transform, 55
- Prokhorov’s theorem, 66
- Rashomon effect, 39
- risk, 194
- score function, 150
- second Bartlett identity, 151
- Stirling’s formula, 47
- strong consistency of estimators, 44
- subsequence lemma, 57
- sufficient statistic, 126
- the diagonal method, 57
- tight random vectors, 66
- tightness, 56
- Toeplitz lemma, 64
- triangular array, 63
- weak convergence, 52