Statistical Inference:
777 Exercises, 77 Stories, and Solutions

Nils Lid Hjort
University of Oslo

Emil Aas Stoltenberg
University of Oslo and the BI Norwegian Business School

– This PartTwo version, for STK 4090 / 9090,
with preliminary versions of the 77 Statistical Stories,
last touched by Nils and by Emil, 16-January-2025 –

©Nils Lid Hjort and Emil Aas Stoltenberg, 2025

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.A.S.

Preface

This book builds on Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

(xx then three-four paragraphs here, on the carrying ideas behind and structure of the book: *exercises* and *stories*. a partly flipped classroom, with direct participation from the first pages of each chapter. there will be *solutions to all exercises*, not physically placed inside the book, but rather on the book's website-to-be, perhaps url'd `www.mn.uio.no/math/english/research/projects/HjortStoltenberg`. That website will also have all datasets and code is `R` and `python` to carry out all analyses, for the construction of each of the book's figures, etc.

the Hjort-Stoltenberg book website

(xx if we're clever with the 777 exercises, 77 stories, we should mention Stigler's 7 pillars. x)

(xx briefly on on prerequisties: linear algebra, with matrix theory, etc.; calcululs, with functions of one or more variables, partial derivatives, etc.; programming, in `R` or `Python` or other appropriate language, both for running common algorithms inside relevant pacakages, and for programming one's own functions, for simulation, etc.)

(xx crisp clear prose here, regarding segments of readers and how they can manouevre through the material. overall: from beginning master's level, in statistics, probability theory, data science, machine learning, and upwards, to PhD level and more. xx) (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 high master or PhD level. (ii) The Statistical Stories Readers, for those who already know the basics on statistical models, parameter estimation and testing, some Bayes, etc. (iii) Our book is also for the specialists inside certain themes, who wish to learn even more.

(xx crisp clear prose here, regarding courses and teaching. below we help readers and instructors by also providing short lists of relevant stories, for the different types of courses using our book. xx) Several types of courses can be taught from this book.

five courses which can be taught from our book

(i) Hard-core statistical inference, with parametric models, etc.: Chs. 1, half of 2, then most of 3, 4, half of 5, 6, 7, 8; a selection of Stories.

(ii) Large-sample theory, the careful probablity theory leading to CLT and more, with applications in statistics: Chs. 1, 2, 5; half of 9, a selection of Stories.

(iii) Empirical processes, convergence, approximations, applications in statistics: Chs. 1, 2, 5, 9; a selection of Stories.

(iv) Survival and event history analysis: Chs. 1, the essence of 2, 3, 4, 5, then the full 9; a selection of Stories.

(v) Model selection and model averaging: Chs. 1, the essence of 2, 3, 4, 5, then the full 11; a selection of Stories.

(vi) Bayesian statistics and confidence distributions: Chs. 1, the essence of 3, 5, then the full 6, 7, 15; a selection of Stories.

(vii) A special course on statistical applications can also be given, with emphasis on the stories, and less concern for all the technical details. This is also a way to apprciate and to learn about modern statistics, concepts, modelling, computation, scientific context.

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 Cristensen, Ingrid Dæhlen, Arnoldo Frigessi, Ingrid Glad, Håvard Hegre, Aliaksandr Hubin, Ingrid Hobæk Haff, Kristoffer Hellton, Anders Dahl Hjort, Bjørn Jamtveit, Martin Jullum, Vinnie Ko, Alexander Koning, George Lowther, Ian McKeague, Per Mykland, Per August Moen, Jonas Moss, Håvard Mokleiv Nygård, Lars Olsen, Sam Power, Oskar Høgberg Simensen, Catharina Stoltenberg, Gunnar Taraldsen, Ingrid Van Keilegom, Lars Walløe, Jonathan Williams, Lan Zhang.

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 2025

Contents

Preface iii

Contents v

I Short & crisp 1

1 Statistical models 3

2 Large-sample theory 49

3 Parameters, estimators, precision, confidence 105

4 Testing, sufficiency, power 139

5 Minimum divergence and maximum likelihood 173

6 Bayesian inference and computation 233

7 CDs, confidence curves, combining information 257

8 Loss, risk, performance, optimality 287

9 Brownian motion and empirical processes 311

10 Survival and event history analysis 349

11 Model selection 363

12 Markov chains, Markov processes, and time series 389

13 Estimating densities, hazard rates, regression curves 407

14 Bootstrapping 425

15 Bayesian nonparametrics 427

16 Statistical learning 429

II Stories **433**

i Demography, Epidemiology, Medicine **435**

ii Art, History, Literature, Music **475**

iii Economics, Political Science, Sociology **515**

iv Biology, Climate, Ecology **553**

v Sports **569**

vi Simulated stories **601**

vii Miscellaneous stories **623**

III Appendix **651**

A Mini-primer on measure and integration theory **653**

B Overview of stories and data **697**

References **717**

Name index **729**

Subject index **731**

Part I

Short & crisp

I.14

Bootstrapping

Part II

Stories

II.i

Demography, Epidemiology, Medicine

(xx WELL: various things to fix, as of 16-January-2025, polish and finish. nils tries to tend to these: (i) round off Oslo quantile babies story. (ii) round off overdispersed and Markov children. (iii) nils goes for Story i.11, with PCI versus only usual medical treatment for coronary disease. (iv) things to calibarte from nils Time to second child and Emil Third child stories. xx)

Story i.1 *Cooling of newborns*
Data: `in text below`
Code: `(xx to come, com106* com109* of nilswork22`
Seminal work carried out by Marianne Thoresen and coworkers (see work pointed to in Walløe et al. (2019a)) has demonstrated that when a newborn has been deprived of oxygen during birth, an emergency intervention involving cooling (therapeutic hypothermia) can save its life, with no loss of motoric or mental abilities later on – provided this is implemented within six hours. Is it still helpful, or not at all, when the cooling scheme starts later than six hours? Laptook (2017) report on a wide and elaborate study, combining information from many registries across several U.S. state, pertaining to this and related question. In particular, one counts the number of events, in the cooled and non-cooled groups, the event in question being death or disability (with a precise definition of disability, assessed when the child is about 18 months old). The essential relevant summary, from all these life-and-death efforts, lies in the two times two table

```
non-cooled infants   m0: y0 and m0-y0:   79: 22 and 57
hypothermic infants: m1: y1 and m1-y1:   78: 19 and 59
```

Seeing these as two biomial experiments, $y_0 \sim \text{binom}(m_0, p_0)$ and $y_1 \sim \text{binom}(m_1, p_1)$, the statistical question is what inferences we may make, for comparing p_0 and p_1.

(a) First, give ordinary (and perhaps approximate) 95 percent confidence intervals for p_0 and p_1, and comment. Then compute and display in the same diagram the confidence distributions $\text{cc}_0(p_0)$ and $\text{cc}_1(p_1)$, associated with the optimal binomial confidence distributions $C(p) = \Pr_p(Y > y_{\text{obs}}) + \frac{1}{2}\Pr_p(Y = y_{\text{obs}})$, as for the left panel of Figure i.1; see Ex. 7.34.

(b) To analyse the degree to which p_0 and p_1 might be different, transform to the logistic scale, with $p_0 = \exp(\theta_0)/\{1 + \exp(\theta_0)\}$ and $p_1 = \exp(\theta_0 + \gamma)/\{1 + \exp(\theta_0 + \gamma)\}$. Note that γ can be seen as the log-odds difference $\log(p_1/(1 - p_1)) - \log(p_0/(1 - p_0))$; results

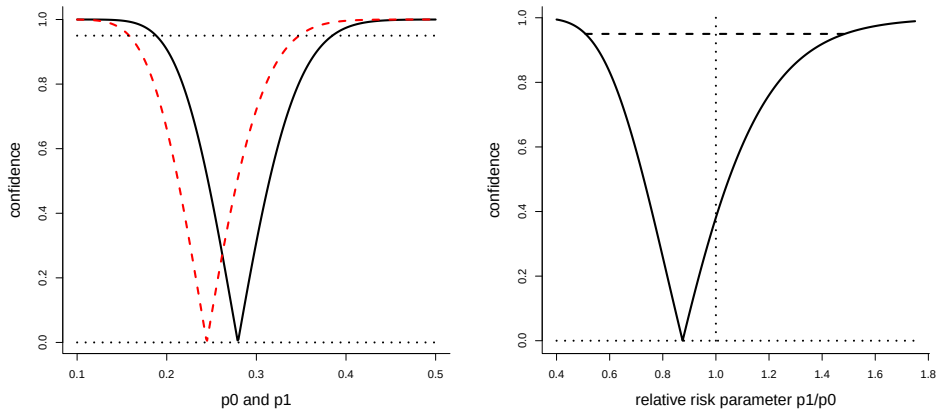


Figure i.1: *Left panel: confidence curves for p_0 and p_1, with 95 intervals [0.189,0.384] and [0.159,0.347]. Right panel: confidence curve for* $\mathrm{rr} = p_1/p_0$, *with 95 interval [0.509,1.485]. There is no indication that p_0 and p_1 really differ.*

may also be given in terms of the odds ratio $\rho = \exp(\gamma)$. Use Ex. 7.35 to compute the optimal confidence curve $\mathrm{cc}(\rho)$, and give a 95 percent interval for the parameter. The Laptook (2017) article reported mainly in terms of the relative risk parameter $\mathrm{rr} = p_1/p_0$, however. Construct therefore a confidence distribution cc(rr) also for that parameter, using the Wilks theorem based recipe of Ex. 5.29, as in the right panel of Figure i.1. Give the median confidence estimate $\widehat{\mathrm{rr}}_{0.50}$ and a 95 percent interval.

(c) The Laptook (2017) report framed results in terms of Bayesian priors and posteriors. With priors $p_0 \sim \mathrm{Beta}(a_0, b_0)$ and $p_1 \sim \mathrm{Beta}(a_1, b_1)$, show that $\mathrm{rr} = p_1/p_0$ given data is a ratio of independent $\mathrm{Beta}(a_0 + y_0, b_0 + m_0 - y_0)$ and $\mathrm{Beta}(a_1 + y_1, b_1 + m_1 - y_1)$. Deduce that the posterior distribution has cumulative

$$F(v \,|\, \mathrm{data}) = \int_0^1 G(vp_0, a_1 + y_1, b_1 + m_1 - y_1) g(p_0, a_0 + y_0, b_0 + m_0 - y_0) \,\mathrm{d}p_0,$$

in terms of the density g and c.d.f. G for Beta distributions. Compute and display this $F(\mathrm{rr} \,|\, \mathrm{data})$, using the Jeffreys prior $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ (xx pointer to Ch7 with this detail xx), Show that it becomes very close to the prior-free CD $C(\mathrm{rr})$, constructed from the cc(rr) above as $\frac{1}{2} - \frac{1}{2}\mathrm{cc}(\mathrm{rr})$ for $\mathrm{rr} \le \widehat{\mathrm{rr}}_{0.50}$ and $\frac{1}{2} + \frac{1}{2}\mathrm{cc}(\mathrm{rr})$ for $\mathrm{rr} \ge \widehat{\mathrm{rr}}_{0.50}$ (xx pointer to that thing in Ch8 xx).

(d) Laptook (2017) used several informative priors for their analyses, including one called by them a neutral prior, with mean zero and standard deviataion 0.35 for log rr. Translate this to two equal Beta priors $(a, a), (a, a)$, finding the a matching their 0.35 standard deviation, perhaps via simulations; you should find $a \doteq 8.95$. For this neutral prior, display the posterior c.d.f. and density for rr. Compute also $\Pr(\mathrm{rr} \le v \,|\, \mathrm{data})$ for $v = 0.90, 0.95, 1.00$.

(e) Above we have computed $\Pr(\mathrm{rr} < 1 \,|\, y_0 = 22, y_1 = 19) = 0.664$ with the neutral prior. Compute $\Pr(\mathrm{rr} < 1 \,|\, y_0 = 22, y_1)$ for imagined data sets with $y_1 = 19, 18, \ldots, 5$, say, keeping the other aspects of the data fixed, including $y_0 = 22$. How small ought y_1 to have been, in order for the $rr < 1$ scenario to have posterior probabilty above 0.95?

(f) (xx round off. point to Hjort blog story Hjort (2017a), big JAMA paper Laptook (2017), short critical follow-up papers Walløe et al. (2019a,b). xx)

Story i.2 *Overdispersed children*
Data: `in text below`
Code: `(xx to come, com75* in nilswork22 xx)`
Some one and a half century ago, there were as many as $n = 38495$ plentiful 8-or-more children families living in Sachsen, with Geißler (1889) dutifully counting and reporting about them and the number of girls and boys. The little table to the left here gives the number $N(y)$ of these families having y girls and $8 - y$ boys, for $y = 0, 1, \ldots, 8$. In the course of this and the following Story i.3 we will work through models 1, 2, 3, say, producing expected numbers $E_1(y), E_2(y), E_3(y)$ to match the $N(y)$, along with what we term Pearson residuals $\{N(y) - E_j(y)\}/E_j(y)^{1/2}$.

y	N	E1	pear1	E2	pear2	E3	pear3
0	264	192.325	5.168	255.621	0.524	255.210	0.550
1	1655	1445.384	5.514	1657.032	-0.050	1655.181	-0.004
2	4948	4752.364	2.838	4909.686	0.547	4901.376	0.666
3	8498	8928.902	-4.560	8683.213	-1.988	8692.383	-2.085
4	10263	10484.952	-2.168	10024.863	2.378	10034.318	2.283
5	7603	7879.792	-3.118	7735.975	-1.512	7736.379	-1.516
6	3951	3701.205	4.106	3896.509	0.873	3890.978	0.962
7	1152	993.421	5.031	1171.238	-0.562	1167.280	-0.447
8	161	116.655	4.106	160.865	0.011	161.895	-0.070

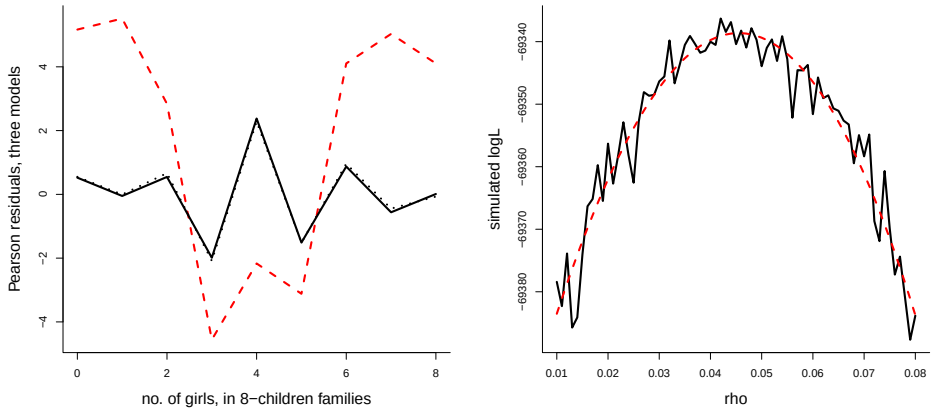


Figure i.2: *Left panel: Pearson residuals* $\{N(y) - E(y)\}/E(y)^{1/2}$, *for three models: the simple binomial (red, dashed curve); the betabinomial (black, full curve); and the Markov model (black, dotted curve). Here* $N(y)$ *is the observed number of girls* y, *and* $E_1(y), E_2(y), E_3(y)$ *the expected number under models 1, 2, 3. Models 2 and 3 produce very similar fitted values. Right panel: simulated log-likelihood* $\widehat{\ell}_n(\rho)$, *using* 10^5 *simulations for each value of the Markovian correlation parameter* ρ, *for 8-children families, along with a cubic smoother, to read off the ML estimate* $\widehat{\rho} = 0.044$.

(a) Compute the overall fraction of girls, among the $mn = 307860$ children, as $\widehat{p} = \sum_{y=0}^{m} N(y)y/(mn) = 0.4844$. Show that the null hypothesis $p = 0.50$ must be soundly rejected here.

(b) Of course a statistician can't always expect to see the difference between 0.500 and 0.485 as a clearly significant one – as this very much depends on the sample size. Suppose you go out on the street sampling, counting a binomial $B \sim \text{binom}(k, p)$ after having studied k objects or persons. How large must k be, in order for your 0.05-level test of $p = 0.50$ against $p \neq 0.50$ to have detection power say 0.95, if the truth is $p = 0.485$? What if you use a 0.01-level test and need detection power 0.99?

(c) Assume that the binomial model $Y \sim \text{binom}(8, p_0)$ holds, with the same p_0 across all families. Find point estimates and 99 percent confidence intervals for $nf_1(0, p_0)$, the expected number of all-boys families, and for $nf_1(8, p_0)$, the expected number of all-girls families, among the $n = 38495$ families with eight children. Then check with the real world.

(d) Under the assumption that the girl-probability p is constant, across families, we would have $Y \sim \text{binom}(8, p)$, for these $n = 38495$ Sachsen families. Compute $E_1(y) = nf_1(y, \widehat{p})$, the expected number of y-girls families, under this model, with $f_1(y, p)$ the usual binomial. Compute also the Pearson residual, say $P_1(y) = \{N(y) - E_1(y)\}/E_1(y)^{1/2}$, for $y = 0, 1, \ldots, 8$. These should roughly be standard normal, if the model used is good. Check with Figure i.2 (left panel). Discuss what you find: in particular, it appears that the real world exhibits significantly more 'extreme families', those with all boys or all girls, than what is predicted under the straight binomial model.

(e) Suppose rather that each family has its own girl-probability p, but that this p varies across families, according to some distribution with overall mean p_0 and positive standard deviation τ_0. Show that $\text{E}\, Y = mp_0$ and that the extra-binomial variability manifests itself by $\text{Var}\, Y = mp_0(1 - p_0) + m(m - 1)\tau_0^2$; this is also clear from Ex. 1.22. Compute the empirical variance $S^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n - 1) = \sum_{y=0}^m N(y)(y - m\widehat{p})^2/(n - 1)$, set the extra-binomial variance $S^2 - m\widehat{p}_0(1 - \widehat{p}_0)$ equal to $m(m - 1)\tau_0^2$, and show that this leads to $\widehat{\tau}_0 = 0.0538$.

(f) Establish that this extra-variation, with $\widehat{\tau}_0 = 0.0538$, is indeed very significantly positive. Again, we would not always be able to identify a standard deviation of this size as being significantly present, but we are, of course, helped by the enormous sample size.

(g) A natural two-parameter model, to explain also the extra-binomial variability, is to take $Y \,|\, p \sim \text{binom}(m, p)$ and $p \sim \text{Beta}(a, b)$; see Ex. 1.22. Show that this leads to

$$f_2(y, a, b) = \binom{m}{y} \frac{\Gamma(a + y)\Gamma(b + m - y)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + b)}{\Gamma(a + b + m)} \quad \text{for } y = 0, 1, \ldots, m.$$

Representing (a, b) as $(kp_0, k(1 - p_0))$, estimate k from the overdispersion number $\widehat{\tau}_0 = 0.0538$; the result should be $\widehat{k} = 85.1961$. Draw the resulting Beta density in a diagram. How many families in the world have their girl-probabilities outside the interval $[0.40, 0.60]$?

(h) Compute the expected numbers $E_2(y)$ and the Pearson residuals $P_2(y) = \{N(y) - E_2(y)\}/E_2(y)^{1/2}$ also for this two-parameter model, and reconstruct the relevant parts of the table above and Figure i.2 (left panel). Compute the Pearson statistics $\sum_{y=0}^m P_1(y)^2 = 159.41$, way too big for the binomial, but $\sum_{y=0}^m P_2(y)^2 = 13.55$, close to acceptable, for the beta-binomial.

(i) To assess whether a proposed model $f(y,\theta)$ for the probabilities $\Pr(Y=y)$ is adequate, one may in addition to Pearson residuals inspection use the likelihood theory of Ch. 5. Show that the log-likelihood function is $\ell_n(\theta)=\sum_{y=0}^m N(y)\log f(y,\theta)$, with an ensuing ML estimator $\widehat\theta$. Deduce that the consequent Wilks testing statistic takes the form $W_n=2\sum_{y=0}^m N(y)\log\{\widehat r(y)/f(y,\widehat\theta)\}$, with $\widehat r(y)=N(y)/n$ the raw estimates. Explain that if the parametric model holds, then the distribution of W_n is very close to a $\chi^2_{\rm df}$, with $\mathrm{df}=m-\dim(\theta)$. Compute the W_n for the two models considered here, the binomial and the beta-binomial, and comment. – For the beta-binomial model, this involves finding the ML estimates via numerical optimisation, leading to $(\widehat a,\widehat b)=(41.244,43.904)$.

Story i.3 *Markovian children*
Data: `in text below`
Code: `(xx to come, com75* in nilswork22 xx)`
Let us now introduce a model for Markovian children, where the gender of your next child depends (but only slightly, as it turns out) on the gender of your currently last child. Let $(q_0,p_0)=(1-p_0,p_0)$ be the long-term frequencies for boys and girls. Instead of taking births to be fully independent of each other, consider the Markov chain $x_1,x_2,\ldots$ of births (again with 0 for boy and 1 for girl), with x_1 drawn from the (q_0,p_0) coin of reproduction, and then following the two-stage transition probability matrix

$$P=\begin{pmatrix} q_0+\rho p_0, & p_0-\rho p_0\\ q_0-\rho q_0, & p_0+\rho q_0\end{pmatrix}=\begin{pmatrix} q_0, & p_0\\ q_0, & p_0\end{pmatrix}+\rho\begin{pmatrix} p_0, & -p_0\\ -q_0, & q_0\end{pmatrix}, \tag{i.1}$$

with ρ a fine-tuning Markov dependence parameter.

(a) Argue first that with $\rho=0$ we're back to ordinary independence, with a binomial distribution for the number of girls in a family of a given size. Show that (q_0,p_0) indeed is the equiblibrium distribution for this Markov chain. Work out the covariance of 'it's a girl', 'it's a girl', and show that ρ is the correlation, also for boy, boy. What is the parameter region for ρ? Below we shall find 'same gender twice in a row' correlation ρ around 0.05.

(b) Klotz (1972) discussed girl-boy sibling sequences among 195 Amish families, in a society where presumably birth control mechanisms were not used or indeed thought of (the data stem from observations recorded before 1910). The sibling flock sizes were from 2 to 16. In total these 195 families had 716 girls and 742 boys, with girl-ratio $716/1458=0.491$. We now fit the two-parameter Markov model (i.1) to these data. Write $(x_{i,1},\ldots,x_{i,m_i})$ for the gender sequence in family i, with 1 for girls and 0 for boys. Show that the full likelihood for the $n=195$ sibling chains of children may be expressed as

$$L(p_0,\rho)=\prod_{i=1}^n\Big\{(1-p_0)^{1-x_{i,1}}p_0^{x_{i,1}}\prod_{j=2}^{m_i}\Pr_{p_0,\rho}(x_{i,j}\,|\,x_{i,j-1})\Big\},$$

and that this leads to the log-likelihood

$$\begin{aligned}\ell(p_0,\rho)\;=\;&M\log p_0+(n-M)\log(1-p_0)\\ &+N_{0,0}\log p_{0,0}+N_{0,1}\log p_{0,1}+N_{1,0}\log p_{1,0}+N_{1,1}\log p_{1,1},\end{aligned}$$

in terms of the components $p_{0,0},p_{0,1},p_{1,0},p_{1,1}$ of the transition probability matrix (i.1). Here M is the number of families with first child girl, $n-M$ the number of families with

first child boy, and $N_{a,b} = \sum_{i=1}^{n} \sum_{j=2}^{m_i} I((x_{i,j-1}, x_{i,j}) = (a,b))$ the number of transitions seen from a to b. Using ML theory methods from Ex. 12.18, establish that the point estimate for ρ becomes a significant nonzero 0.074, with standard error 0.028; the 95 percent interval is $[0.019, 0.127]$. Construct also a confidence curve $\mathrm{cc}(\rho)$. Give also the estimated transition probability matrix. What is the estimated probability of having five girls in a row? – Remark (i): Note the practical point that we can carry out full analysis via maximisation of $\ell(p_0, \rho)$, or of its profiled $\ell_{\mathrm{prof}}(\rho)$; we do not need extra details for this or similar models, per se, though such are given in Klotz (1972, 1973); Lindqvist (1978). Remark (ii): It took us considerable efforts and patience to transcribe the Klotz (1972) data, which he gave in compressed octogonal form, into sibling gender sequencess of 0s and 1s. Here it is sufficient to inform our readers about the required summary statitics: $M = 85$, $(N_{0,0}, N_{0,1}, N_{1,0}, N_{1,1}) = (345, 298, 287, 333)$. Explain that you for the Markov model worked with here do not need more detail for the $n = 195$ variable-length 0-1 time series.

(c) For the Klotz (1972) data, we have the precise gender sequences for the 195 families. For the Geißler data we have only the number of girls and boys in a sibling flock, however, making it harder but not impossible to unearth Markovian dependence. Consider any parametric model for $f(y, \rho) = \Pr_\rho(Y = y)$, the number of girls among m siblings. As we have seen in Story i.2, the log-likelihood function is $\ell_n(\rho) = \sum_{y=0}^{m} N(y) \log f(y, \rho)$, with $N(y)$ the number of times $Y = y$ is observed in the data. For the Markov model there is no easy formula for $f(y, \rho) = \Pr_\rho(X_1 + \cdots + X_m = y)$, but we may for each ρ simulate a high number of chains $(x_1, \ldots, x_m)$ and estimate $f(y, \rho)$ with the relative proportion of these chains which have sum y. Argue that this leads to the simulated log-likelihood function

$$\widehat{\ell}_n(\rho) = \sum_{y=0}^{m} N(y) \log \widehat{f}(y, \rho).$$

Implement such a scheme, and construct a version of Figure i.2, right panel (which used 10^5 simulations for each value in a grid of ρ). Supplement the $\widehat{\ell}_n$ curve with a cubic regression smoother, of the type $\widehat{\rho}(\rho) = \beta_0 + \beta_1\rho + \beta_2\rho^2 + \beta_3\rho^3$; this is the smooth approximation curve in the figure's right panel. Read off $\widehat{\rho} = 0.044$ as a good numerical approximation to the ML estimator. Use the Wilks based construction recipe of Ex. 7.11 to produce a confidence curve $\mathrm{cc}(\rho)$, and read off a 95 percent interval (which turns out to be $[0.037, 0.051]$).

(d) For the estimated $\widehat{\rho} = 0.044$, simulate a high number of $(x_1, \ldots, x_m)$ Markovian children in m-children families, to get $\widehat{f}_3(y) = f(y, \widehat{\rho})$. Compute from this expected numbers $E_3(y) = n\widehat{f}_3(y)$, Pearson residuals $(N - E_3)/E_3^{1/2}$, as with the table of Story i.2. Compute also the Pearson type statistic $\sum_{y=0}^{m} P_3(y)^2$, which will be close to the corresponding statistic for the beta-binomial model. As with Story i.2(i), compute the Wilks statistic W_n for the present Markov model, and comment on your findings.

Story i.4 *IUD expulsion*
Data: `iud-data`
Code: `(xx to come com88* com90* of nilswork22 xx)`
Data have been collected for a certain intrauterine device for $n = 100$ women (xx pointer here to data overview, there mentioning Peterson (1975) and Aalen, but mention that

this is 1970ies in U.S. xx). The `iud-data` file has three columns: the index $i = 1, \ldots, n$; the time t_i to 'event', measured in days, from the first day of use; and an index for 'event', specifically with '2' denoting expulsion, and '3' or '4' being removal for pains or bleeding. Here we shall be concerned with modelling the time to expulsion of the IUD, and let δ_i be a 1-0 indicator for this event. For most users this never happens (there are 11 cases of $\delta_i = 1$), so this means heavy censoring, in the survival analysis language of Ch. 10.

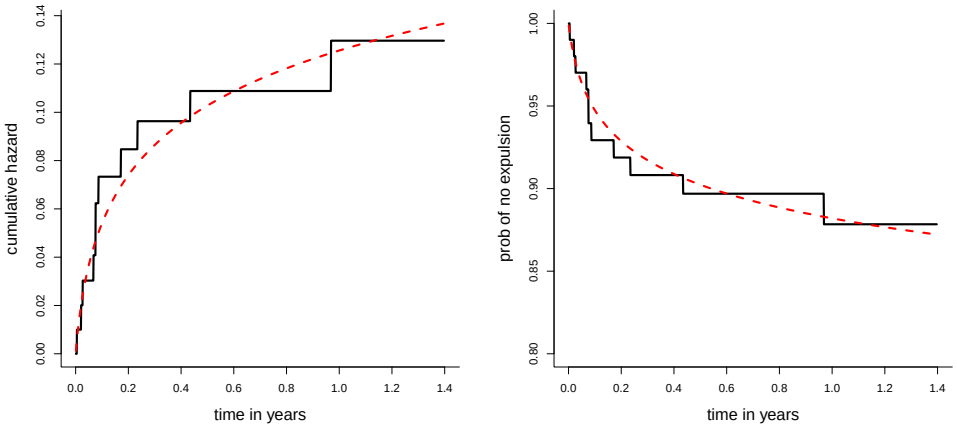


Figure i.3: For the IUD expulsion data: cumulative hazards (left panel); survival curves Pr(no expulsion so far) (right panel), where the scale is above 0.80. Nonparametric and parametric.

(a) (xx check similarity with substory for Better Angels. xx) Assume each woman has her constant hazard rate θ, in the process leading to IUD expulsion, but that these rates vary in the population of IUD users, according to a Gam(a, b) distribution. With the convenient reparametrisation $(a, b) = (\theta_0/c, 1/c)$, show that θ then has mean θ_0 and variance $c\theta_0$. Show, perhaps using Ex. 1.10, that the hazard rate for a randomly sampled woman then becomes $\alpha(t, \theta_0, c) = \theta_0/(1 + ct)$. The c parameter might be called a frailty parameter, since it models the degree to which IUD users are different, with the frail ones tending to experience expulsion early, and with yet others not having problems at all. Comment on the case $c \to 0$.

(b) Show that the log-likelihood function becomes

$$\ell(\theta_0, c) = \sum_{i=1}^{n} \big[\{\log \theta_0 - \log(1 + ct_i)\}\delta_i - (\theta_0/c)\log(1 + ct_i)\big],$$

and give also an expression for the profiled log-likelihood $\ell_{\rm prof}(c)$. Display that function, and compute the ML estimates $(1.320, 38.710)$ for (θ_0, c). Construct versions of the left and right panels of Figure i.3.

(c) Compute estimated standard deviations, for the estimators portrayed in Figure i.3, and construct additional figures with pointwise 90 percent confidence bands.

(d) The ML estimates above, along with their estimated standard deviations, are found based on data for $n = 100$ women followed over 510 days, or 1.397 years, during which 11 users experienced expulsion. Simulate more data, from the estimated model, say for another 510 days, for the 89 users still at risk for expulsion. Assess how much more precise the model parameters can be estimated now. (xx here i also have in mind the approximate variance expression $(1/n)J^{-1}_{[0,\tau]}$, where we now increase τ. xx)

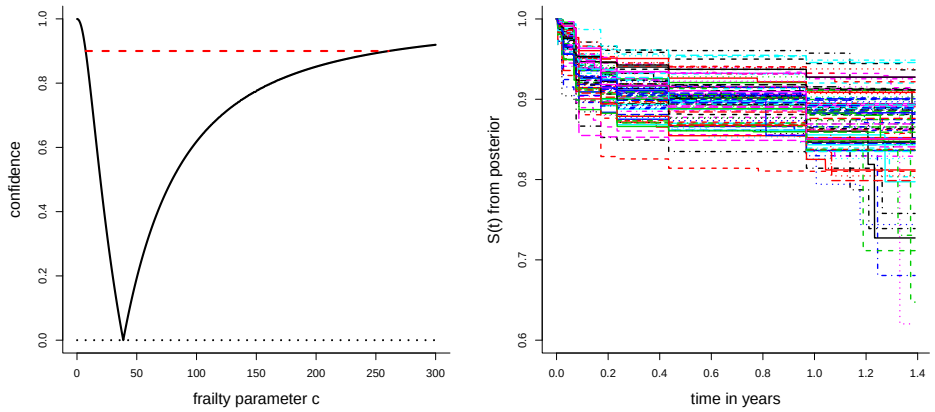


Figure i.4: Left panel: confidence curve $\mathrm{cc}(c)$ for the frailty parameter c. The ML estimate is 38.71, and the 90 percent interval is quite skewed, from 7.27 to 260.81. Right panel: 100 simulated $S(t)$ from posterior distribution.

(e) Compute and display a confidence curve $\mathrm{cc}(c)$ for the dispersion parameter c, as in Figure i.4, left panel, using the Wilks theorems, see Ex. 7.11, and give in particular a 90 percent confidence interval. Note that the $\mathrm{cc}(c)$ is strongly skewed to the right, giving asymmetric confidence intervals.

(f) We now disregard the parametric model worked with above, and enter Bayesian nonparametric terrain from Ch. 15, aiming at using Beta processes for the cumulative hazard $A(t)$ to reach posterior inference for both A and the survival function $S(t) = \prod_{[0,t]}\{1 - \mathrm{d}A(s)\}$, see (xx pointer to exercises xx). The Beta process prior involves setting up the prior mean function $\mathrm{d}A_0(s) = \alpha_0(s)\,\mathrm{d}s$ and prior precision function $k(s)$, with increments $\mathrm{d}A(s)$ having Beta distributions with parameters $(k(s)\alpha_0(s)\,\mathrm{d}s, k(s)\{1-\alpha_0(s)\,\mathrm{d}s\})$. Here we do this by taking $A_0(t) = \alpha_0 t$ with constant rate, supplemented with a constant precision function k. The prior construction task is then to choose rate α_0 and precision k so that Beta increments $(k\alpha_0\,\mathrm{d}s, k(1-\alpha_0\,\mathrm{d}s))$, with means $\alpha_0\,\mathrm{d}s$ and variances $\alpha_0\,\mathrm{d}s/(k+1)$, are reasonable, but still allow flexibility in order for the data to point us in other directions, if appropriate, in the given IUD expulsion context. For the present illustration, aiming at a somewhat informative prior, we take into account that around 90 percent of IUD users do not experience expulsion during a year of use, and decide somewhat arbitrarily to translate the prior desiderata into (i) having $S_0(t) = \exp\{-A_0(t)\} = 0.90$, at time $t_0 = 1.00$ years; (ii) having a distribution for $S(t_0)$ which has 0.95 of its probability mass above 0.60. Show that the first demand means $\alpha_0 = -\log 0.90 = 0.105$, and then carry out a search over a grid of k values, checking for

each the implied distribution for $S(t_0)$. Make a plot of the 0.05 quantile for $S(t_0)$, as a function of k, and show that this leads to $k_0 = 2.20$ (computed without full precision). Simulate a high number of $S(t_0)$, for this choice of (k_0, α_0), and display a fine histogram.

(g) Having completed the Beta process prior construction, carry out posterior inference. Display say 100 simulated paths of A and of $S = \prod_{[0,t]}(1 - \mathrm{d}A)$, as with Figure i.4, right panel. Comment on what this conveys. Construct also a figure showing, as a function of time up to 1.50 years, (i) the 0.05 and 0.95 quantiles for the posterior of $S(t)$, (ii) a 90 percent frequentist confidence band, around the Kaplan–Meier curve (xx pointer to Ch11 things xx). This demonstrates that the Bayesian nonparametric approach, with the Beta process prior with modest prior strength, produces bands with the right frequentist coverage.

(h) (xx with your code, try other schemes for the Beta process prior. in particular, show the results for the case of the noninformative case, with $k(s)$ close to zero. compare the Bayesian posterior 90 percent pointwise bands, here computed via simulations but no large-sample approximations, with those obtained via Nelson–Aalen bands. they are pretty similar, so frequentist coverage of Bayesian bands pretty accurate. point to BvM theorem here. xx)

Story i.5 *Boys are born slightly bigger than girls*
Data: `birthweight-boys, birthweights-girls`
Code: `(xx to come com110* com328* nilswork22 com80* nilswork23 xx)`
(xx quantile things, to be told. perestroika required. first separate quantiles, boys and girls, then ratios of quantiles. xx) $n_b = 548$ boys and $n_g = 480$ girls born in oslo. ratio of quantiles. let $f_b(x)$ and $f_g(x)$ be the birthweight densities, for boys and for girls, with cumulative distribution functions $F_b(x)$ and $F_g(x)$. here we shall compare quantiles for boys and girls, $\mu_{b,q} = F_b^{-1}(q)$ and $\mu_{g,q} = F_g^{-1}(\mu_g)$, at different levels q. may give a figure of estimated densities, standard kernel methods from Ch. 13. see Figures.

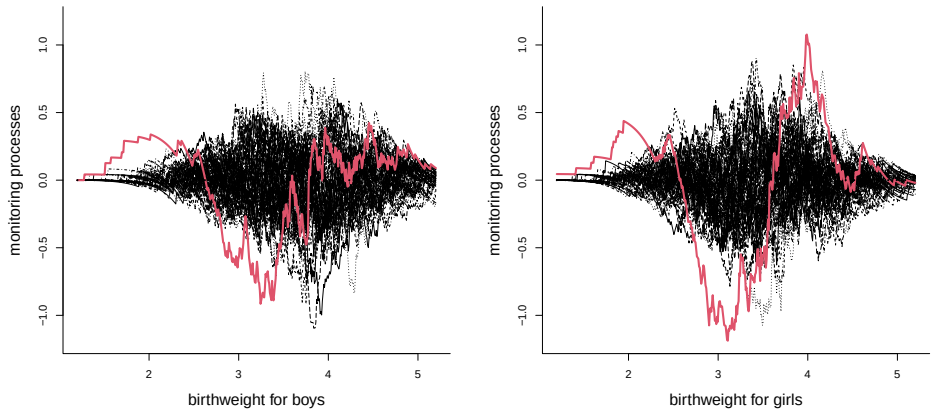


Figure i.5: *For the 548 boys and 480 girls born at Rikshospitalet in Oslo, during 2001–2008, normal monitoring process plots* $Z_n(y) = \sqrt{n}\{F_n(y) - F(y, \widehat{\xi}, \widehat{\sigma})\}$ *(fat lines), along with 50 simulated processes under the normality assumption. Left and right panels for boys and girls.*

(a) Show that boys are significantly bigger than girls, but that there is no clear indication that they have different variances in their birthweight distributions.

(b) Could the birthweights actually be nicely normally distributed? Following Ex. 9.27, compute and display the normal monitoring process $Z_n(y) = \sqrt{n}\{F_n(y) - \Phi((y - \widehat{\xi})/\widehat{\sigma})\}$, for boys and girls, and compute also the Kolmogorov–Smirnov maxima $\max_y |Z_n(y)|$. In order to check whether these plots look normal or not, simulate such Z_n paths under normality. Construct a version of Figure i.5, there with 50 simulated paths. Compute p-values for the Kolmogorov-Smirnov tests, via a higher number of simulations, and from these draw the conclusion that no, birthweights are not normally distributed, a fact we are able to spot thanks to the relatively large sample sizes. Comment on what you learn from the shapes of the Z_n processes, attempting to quantify in which ways the distributions are not normal.

(c) The non-normality of the distributions motivates nonparametric assessment of quantiles. For each of the five quantile levels $0.1, 0.3, 0.5, 0.7, 0.9$, construct a CD for the $F^{-1}(q)$, for boys and for girls, using the order statistic method of Ex. 7.19. Compute also the consequent confidence curves $\mathrm{cc}(\mu_q)$, and make a version of Figure i.6. (xx should get a better plot from nils com16c. xx)

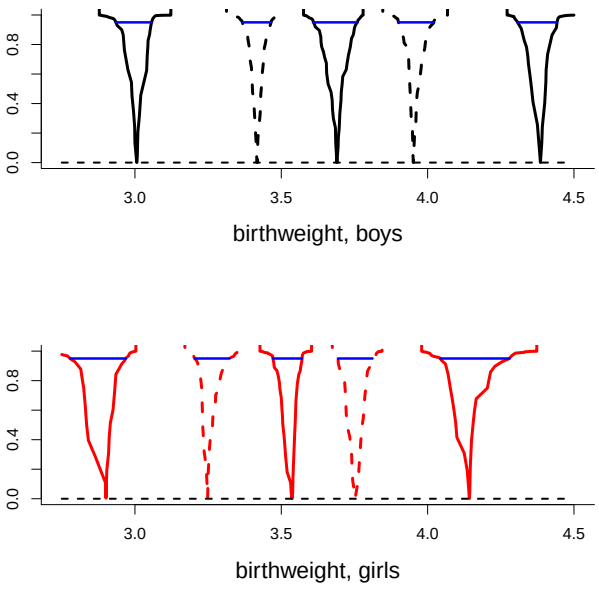


Figure i.6: *Confidence curves for the five deciles $F^{-1}(q)$, for levels 0.1, 0.3, 0.5, 0.7, 0.9, for the birthweight distributions of boys (upper panel) and girls (lower panel), in kg. 95 percent confidence intervals for the 5 + 5 quantities are indicated with the blue horizontal lines.*

(d) In the following, keep one quantile level q fixed, to avoid a too heavily subscripted notation. From Ex. 3.21 we know that $n_b^{1/2}(Q_b - \mu_b) \to_d \mathrm{N}(0, \kappa_b^2)$ and $n_g^{1/2}(Q_g - \mu_g) \to_d$

$\mathrm{N}(0,\kappa_g^2)$, with $\kappa_b = \{q(1-q)\}^{1/2}/f_b(\mu_{b,q})$ and with $\kappa_g = \{q(1-q)\}^{1/2}/f_g(\mu_{b,g})$. Show that this entails

$$Q_b = \mu_b + \kappa_b/n_b^{1/2} Z_b \quad \text{and} \quad Q_g = \mu_g + \kappa_g/n_g^{1/2} Z_g,$$

where $Z_b = Z_{b,n_b}$ and $Z_g = Z_{g,n_g}$ have distributions coming (very) close to the standard normal.

(e) Estimate the difference between the boys and girls distributions, as a function of the quantile q, along with a confidence band. Construct a version of Figure i.7, using gram. The horizonal line represents the estimated overall difference $d = \xi_b - \xi_g$.

(f) We now wish to estimate the ratio of quantiles function $\rho = \mu_b/\mu_g$, nonparametrically, using $\widehat{\rho} = Q_b/Q_g$. Use delta method arguments to deduce that

$$\widehat{\rho} = \rho\{1 + (1/\mu_b)(\kappa_b/n_b^{1/2})Z_b - (1/\mu_g)\kappa_g/n_g^{1/2} Z_g\},$$

and from this that $\widehat{\rho} \approx_d \mathrm{N}(\rho, v^2)$, with variance

$$v^2 = \Big(\frac{\mu_b}{\mu_g}\Big)^2 \Big(\frac{1}{\mu_b^2}\frac{\kappa_b^2}{n_b} + \frac{1}{\mu_g^2}\frac{\kappa_g^2}{n_g}\Big) = \frac{1}{\mu_g^2}\Big(\frac{\kappa_b^2}{n_b} + \rho^2\frac{\kappa_g^2}{n_g}\Big).$$

Construct a version of Figure i.7. (xx conclude that across quantiles, boys tend to be about 5 percent bigger than girls. Attempt to build a model for how $F_b^{-1}(q)$ for boys relates to $F_g^{-1}(q)$ for girls. xx)

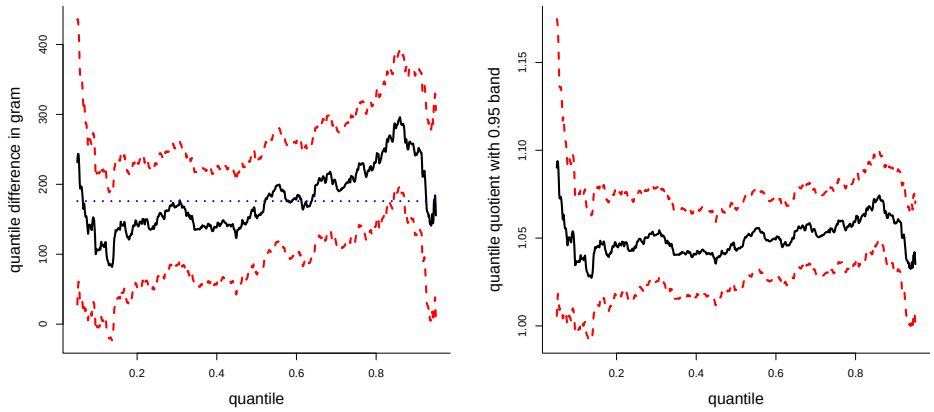


Figure i.7: *For the 548 boys and 480 girls born at Rikshospitalet in Oslo, during 2001–2008, and for quantiles in [0.05,0.95], the plot displays the estimated difference of quantiles (left panel), and the estimated quotient of quantiles (right panel), along with a 95 percent confidence band.*

(g) (xx just a bit more. xx) [xx can ask for estimates with bands of the ratio $f_1(y)/f_2(y)$, perhaps constructed by first estimating the log difference, finding band there, and exping home. could also bake a Type B Story from the birthweights of Oslo boys and Oslo girls, 2001–2008, with other natural analyses. see (xx Data Story B.2.B xx). xx]

Story i.6 *Mothers, babies, birthweights, factors*
Data: `smallbabies-data`
Code: `(xx to come com75* nilswork24 com12* nilswork23 xx)`
(xx polish. we include perhaps as many as six figures. xx) Here we work with the dataset for $n = 189$ mothers and their newborns, collected for a study at Baystate Medical Center, Springfield, Massachusetts, during 1986. The focus is on y, the birthweight (in kg), with covariates including x_1, weight before pregnancy (in kg); x_2, age; x_3, indicator for smoking or not; x_4, indicator for ethnic group 1 (white), x_5, indicator for ethnic group 2 (black), x_6, indicator for ethnic group 3 (other). Of particular interest in this study was their potential influence on the probability for having a small birthweight, defined as $y \leq y_0 = 2.500$. One may therefore work with e.g. logistic regressions, for the 1-0 variables $z_i = I(y_i \leq y_0)$, or with other regression models with the continuous outcomes y_i.

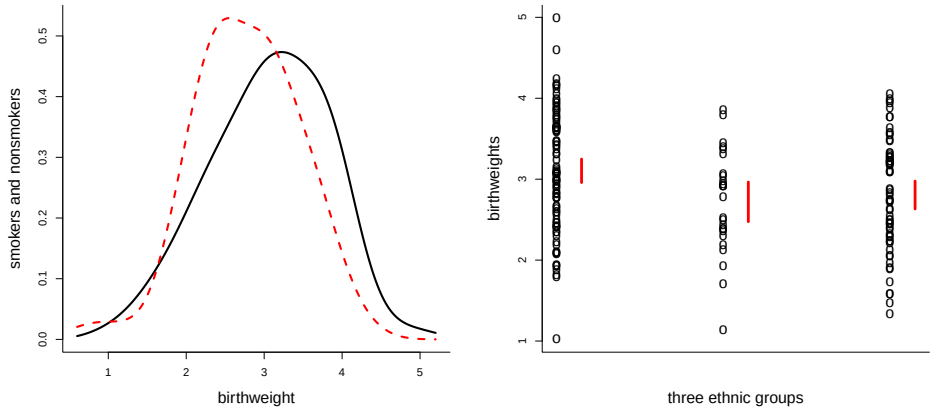


Figure i.8: *Left panel: density estimates* $\widehat{f}_0$ *and* $\widehat{f}_1$*, for weights of newborns, in kg, from mothers who are respectively nonsmokers (full curve) and smokers (broken curve). Right panel: birthweights for the three ethnic groups, with 99 percent confidence intervals for their means.*

(a) Construct and display nonparametric density estimates $\widehat{f}_0$ and $\widehat{f}_1$ for the birthweights, for non-smoking and smoking mothers, as in Figure i.8, left panel. Test equality of means for the two groups; give a 95 percent confidence interval for the mean difference, say $\mu_0 - \mu_1$; and test also equality of spread.

(b) Plot the birthweights for the three ethnic groups, as in Figure i.8, right panel, along with 99 percent confidence intervals for the three population means, say ξ_1, ξ_2, ξ_3. Carry out a one-way layout test to assess the hypothesis $\xi_1 = \xi_2 = \xi_3$, using Ex. 4.42. Use similar methods to check whether the weight of mothers are about the same in the smoking and non-smoking groups.

(c) As a side issue, not connected to the main story concerning influence of covariates, plot the empirical c.d.f. F_n for the weight of mothers, and use the methods of Ex. 4.3 to read off 95 confidence intervals for the quantiles $F^{-1}(q)$, at levels 0.10, 0.50, 0.90. Construct a version of Figure i.9, left panel. (xx answers: 43.10 to 46.20 for 0.10; 54.44 to 56.69 for median; 74.85 to 84.36 for 0.90. xx)

(d) Clearly the weights of newborns carry significant variability, here with standard deviation around 0.730, and it is not to be expected that this variability can be fully explained via taking age, weight, smoking habits, ethnicity into a model. Use methods of Ex. 4.37 to assess how much of the overall variability is being explained by the five covariates $x_1, \ldots, x_5$. This is most conveniently done via the linear regression model $y_i = \beta_0 + \beta_1 x^*_{i,1} + \cdots + \beta_5 x^*_{i,5} + \varepsilon_i$, using the normalised covariates $x^*_{i,j} = x_{i,j} - \bar{x}_j$ for $j = 1, \ldots, 5$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Compute as in the exercise pointed to $\Sigma_n = (1/n)\sum_{i=1}^n x^*_i (x^*_i)^{\mathrm{t}}$, the least squares estimates $\widehat{\beta}_0, \widehat{\beta}$, along with the terms in the variance decomposition

$$(1/n)\sum_{i=1}^n (y_i - \bar{y})^2 = (1/n)\sum_{i=1}^n (y_i - \widehat{\mu}_i)^2 + \widehat{\beta}^{\mathrm{t}} \Sigma_n \widehat{\beta},$$

in which $\widehat{\mu}_i = \widehat{\beta}_0 + x_i^{\mathrm{t}} \widehat{\beta}$. Use this to compute also

$$R^2 = \frac{\widehat{\beta}^{\mathrm{t}} \Sigma_n \widehat{\beta}}{\widehat{\beta}^{\mathrm{t}} \Sigma_n \widehat{\beta} + \widetilde{\sigma}^2} = 1 - \frac{(1/n)\sum_{i=1}^n (y_i - \widehat{\mu}_i)^2}{(1/n)\sum_{i=1}^n (y_i - \bar{y})^2},$$

interpreted as the part of the variance explained via the covariates. The result is 14.84 percent. Then do more, finding a full confidence curve for the population parameter $\rho = \beta^{\mathrm{t}} \Sigma_n \beta / (\beta^{\mathrm{t}} \Sigma_n \beta + \sigma^2)$ for which R^2 is an estimate. Construct a version of Figure i.9, right panel, which has this $\mathrm{cc}(\rho)$, using all five covariates, alongside corresponding one-at-a-time confidence curves for the five separate covariates. Note that the implied σ parameter, in these submodels, changes value and interpretation, depending on which covariates are included in the regression equation. We learn that the five individual covariates only explain 3-5 percent of the overall variability each, whereas the full information of the five covariates explains about 14 percent. A 90 percent interval for that ρ is $[6.31, 21.23]$, indicated via the horizontal 0.90 line in the confidence figure.

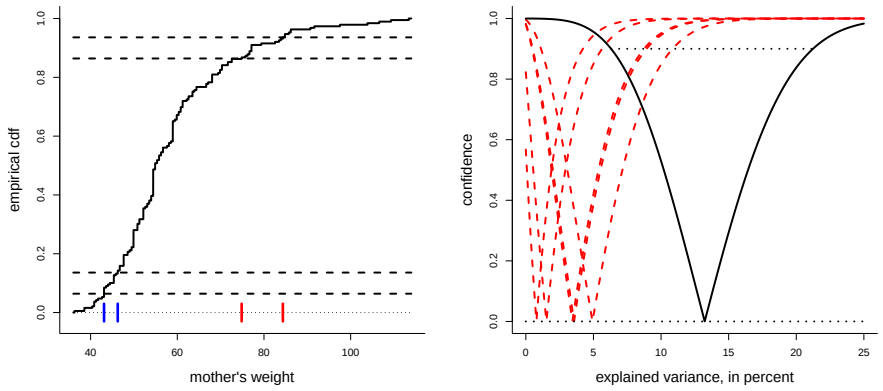


Figure i.9: *Left panel: the empirical c.d.f. for the pre-pregnancy weight of 189 mothers, with bands to read off 90 percent confidence intervals for the 0.10 level and 0.90 level quantiles. Right panel: confidence curves* $\mathrm{cc}(\rho)$ *for the ratio of variation explained by the five covariates, in percent (full curve), along with sumilar one-at-a-time confidence curves for the five individual covariates.*

Story i.7 *Mothers, babies: logistic regression or regression with logistic distribution?*
Data: `smallbabies-data`
Code: `(xx to come com102* of nilswork23 xx)`
For data of the type worked with in Story i.6, with interest focused on 'is $y_i \leq y_0$ or not', one may use logistic regression on the $z_i = I(y_i \leq y_0)$, but also exploit the continuous pre-thresholded raw data themseves. Here we compare these two approaches, with a logistic distribution model for the y_i. The regression coefficients β_j are then the same, for the two models, making the comparison exercise a clear one.

(a) Carry out logistic regression with the five covariates, for the 1-0 outcomes z_i, i.e. using the model $p_i = H(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_5 x_{i,5})$ with $H(u) = \exp(u)/\{1+\exp(u)\}$ the logistic transform. This is easily accomplished in R using `doit = glm(y ~ x1 + x2 + x3 + x4 + x5, family=binomial)` and `summary(doit)`, producing a table as with the four left columns below. Explain how these estimates can be interpreted; carry out Wald tests, as per Ex. 5.50; and give 95 percent intervals for the five regression coefficients.

	logistic regression:					with full birthweight outcomes:				
	estimate	stderr	z value	Pr(>|z|)		estimate	stderr	z value	Pr(>|z|)	
	1.27571	1.01663	1.255	0.20954						
x1	-0.02761	0.01408	-1.961	0.04982	*	-0.0255	0.0096	-2.6577	0.0079	** beta1
x2	-0.02248	0.03417	-0.658	0.51065		0.0013	0.0265	0.0500	0.9601	beta2
x3	1.05444	0.38000	2.775	0.00552	**	1.0415	0.2889	3.6057	0.0003	*** beta3
x4	-0.94326	0.41623	-2.266	0.02344	*	-0.9687	0.3187	-3.0400	0.0024	* beta4
x5	0.28841	0.52676	0.548	0.58402		0.2790	0.4157	0.6711	0.5022	beta5

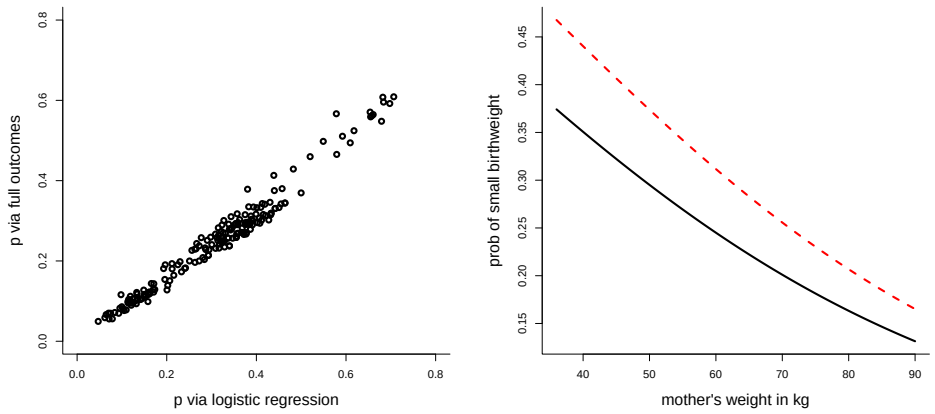


Figure i.10: *Left panel: plot of* $(\widehat{p}_i, p_i^*)$*, estimated probabilities for having birthweight less than 2.500 kg, for the 189 mothers, using respectively the logistic regression model and the full-data model. The* p_i^* *estimates are more precise. Right panel: probability of small birthweight, for a white smoking mother of mean age, as a function of her weight pre pregnancy, estimated with logistic regression (upper curve) and the full-data model (lower curve). The full-data model is more precise.*

(b) Hosmer and Lemeshow (1999) and other researchers have analysed these data with an emphasis on logistic regression, i.e. using the indicator outcomes $z_i = I(y_i \leq 2.500)$,

somehow throwing away part of the statistical information. To see if it pays off to use the full birthweight outcomes y_i, consider the seven-parameter model where $y_i \,|\, x_i$ has c.d.f. $G(y_i \,|\, x_i) = H(y_i/\tau + \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_5 x_{i,5})$. Argue that the $\beta_1, \ldots, \beta_5$ coefficients have the same interpretation as in the logistic setup. Fit this model, by programming and optimising the log-likelihood, reaching coefficients, standard deviations, Wald ratios, p-values as in the four right columns in the table. Note that the β_j estimates are similar, but they are now more precisely estimated, with tighter confidence intervals, sharper Wald ratios, and smaller p-values where they matter.

(c) With the two models we may now compute and compare estimates of the $p_i = \Pr(y_i \leq 2.500 \,|\, x_i)$. For the full-data model, these are $p_i^* = H(2.500/\widehat{\tau} + \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_5 x_{i,5})$. Construct a version of Figure i.10, left panel. The correlation is as high as 0.98, but estimates from the full-data model are more precise. Construct also a version of Figure i.10, right panel, giving two estimates of the probability curve $\Pr(Y \leq 2.500 \,|\, x)$, for a white smoking woman, of mean age 23.24, as a function of her weight.

Story i.8 *Mothers, babies, birthweights, birth order*
Data: `thousand-babies`
Code: `(xx to come com90* of nilswork24 xx)`
A high number of standard models and methods for statistical inference presuppose *independence*, that measurements taken do not influence the others. Here we consider a dataset of consecutive birthweights, for 200 women having had five children. At the outset also these thousand birthweights could be independent, but we learn via modelling, testing, estimation that (i) there is indeed dependence, inside each quintuple of siblings, and (ii) that the birthweight is slowly increasing as a function of birth order. The dataset consists of $(y_{i,1}, \ldots, y_{i,5}, x_{i,1}, \ldots, x_{i,5})$, with birthweights, then mother's age at these births.

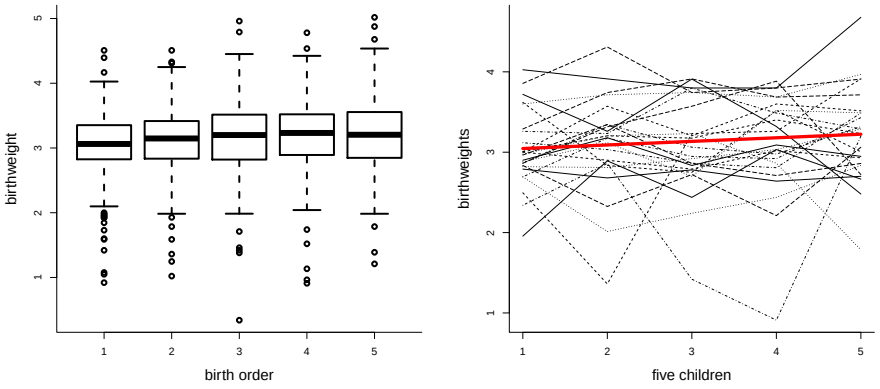


Figure i.11: *Left panel: boxplots for birthweights (in kg), for children 1 to 5, for the 199 mothers. Right panel: the birthweights, for children 1 to 5, for 25 of the 199 mothers, along with the estimated mean curve* $\widehat{a} + \widehat{b}(j-1)$, *with about 45 grams increase per new child.*

(a) Let $Y_i = (Y_{i,1}, \ldots, Y_{i,5})$ be the 5-vector of birthweights for mother $i = 1, \ldots, n$. Produce versions of Figure i.11, with so-called boxplots in the left panel and a sam-

ple of 25 such Y_i plotted in the right panel (with weight in kg). The boxplots are quick and informative summaries of unidimensional datasets, giving minimum and maximum, with the median shown inside the 0.25 to 0.75 quantiles, and are plotted using `boxplot(cbind(y1,y2,y3,y4,y5))` commands in `R`.

(b) For the 200 mothers, plot their ages for when they gave birth. You will then detect that mother no. 142 mysteriously had her 4th and 5th child at age 99. Not quite willing to believe this, push her out of the dataset, so that the following analyses are carried out for the remaining 199 mothers.

(c) For testing whether the mean parameters $\xi_1, \ldots, \xi_5$ for five groups are equal, Ex. 4.42 gives a clear recipe, via an F test, with between-group variability divided by within-group variability. Explain why this might not be appropriate here. To take maternal dependence into account, consider instead the model

$$Y_{i,j} = a + b(j-1) + M_i + \varepsilon_{i,j} \quad \text{for } i = 1, \ldots, n,\ j = 1, \ldots, 5,$$

where the mother components $M_1, \ldots, M_n$ are seen as i.i.d. $\mathrm{N}(0, \tau^2)$, along with the extra variation $\varepsilon_{i,j}$, seen as i.i.d. $\mathrm{N}(0, \sigma^2)$, and independent of the M_i. Explain that the model may be written $Y_i \sim \mathrm{N}_5(\xi_i, \Sigma^2)$, with

$$\xi_i = \begin{pmatrix} a \\ a+b \\ a+2b \\ a+3b \\ a+4b \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \tau^2+\sigma^2 & \tau^2 & \tau^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2+\sigma^2 & \tau^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2+\sigma^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 & \tau^2+\sigma^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 & \tau^2 & \tau^2+\sigma^2 \end{pmatrix}.$$

In particular, the inter-mother correlation between siblings is then $\rho = \tau^2/(\sigma^2+\tau^2)$.

(d) Show that the log-likelihood function can be written

$$\ell_n(a, b, \sigma, \tau) = \sum_{i=1}^{n} \{-\tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}(y_i - \xi_i)^{\mathrm{t}}\Sigma^{-1}(y_i - \xi_i) - (5/2)\log(2\pi)\}.$$

Maximise this function, using the general likelihood theory of Ch. 5 to find both ML estimators and their estimated standard deviations. Produce a table like that under 'model A' below. Explain that this leads to preliminary findings (i) the τ is very clearly present, with estimated intra-mother correlation 0.4201; (ii) that the b parameter is tiny, though highly significant, with about 45 grams added per new child.

	model A			model B		
	para	se	wald	para	se	wald
a	3.0454	0.0354		2.6726	0.1215	
b	0.0446	0.0097	4.5837	0.0031	0.0162	0.1932
c				0.0212	0.0066	3.2046
sigma	0.4343	0.0109		0.4343	0.0109	
tau	0.3697	0.0238		0.3581	0.0234	
logLmax	-734.371			-729.354		

(e) So birth order, with no further information taken into account, is significant. We may formulate and analyse a somewhat bigger model, however, which takes in also the

age at which the mother had her five children. With $x_{i,j}$ her age when having her child j, fit the model

$$Y_{i,j} = a + b(j-1) + cx_{i,j} + M_i + \varepsilon_{i,j} \quad \text{for } i = 1, \ldots, n,\ j = 1, 2, 3, 4, 5,$$

and produce a table like that under 'model B'. Here c for moter's age looks much more important than the b for birth order; explain that this shifts the interpretation of the previous finding from 'the birth order matters' to 'but what really goes on is the age of the mother'. The two explanations, birth order and mother's age, are clearly correlated; compute the estimated correlation between $\widehat{b}$ and $\widehat{c}$.

(f) Carry out model comparison here, using the Wilks test, or the AIC, to demonstrate that the 5-parameter model with both b and c is clearly better than the 4-parameter model with only b.

(g) Working with the 5-parameter model, find and display a confidence curve for the inter-mother correlation. The point estimate is 0.404, and the 95 percent interval is $[0.338, 0.474]$.

Story i.9 *Time to second child after stillbirth*
Data: `time-to-2ndbirth`
Code: `xx (xx to come com28* com39* nilswork23 xx)`
Stillbirths luckily occur more rarely than for earlier generations. Here we consider data extracted from the Norwegian Medical Birth Registry, pertaining to the time to next birth for young women having experienced a stillbirth (death or loss of a baby after 20 weeks of pregnancy). The dataset consists of all the 451 Norwegian women who had their first birth during 1967 to 1971, who were at the time of this birth below 25 years of age and married, and for whom the child was stillborn. The data available are grouped into time windows $[l_j, r_j]$, in months, given in the table below, with Y_j indicating the number of women that at the start of that time window have not yet had a second child, and the number ΔN_j of those who actually gave birth inside that time interval. Below we shall work through two different models for such data, and in the process also estimate the fraction of couples for whom a second child will not occur. For convenience the time scale of months has been converted to years in calculations below.

left	right	Y	DN	left	right	Y	DN
9	10	451	5	36	42	92	7
10	11	446	15	42	48	85	13
11	12	431	27	48	54	72	10
12	13	404	31	54	60	62	4
13	14	373	38	60	72	58	4
14	15	335	29	72	84	54	5
15	18	306	79	84	96	49	1
18	21	227	44	96	108	48	2
21	24	183	36	108	120	46	1
24	27	147	24	120	144	34	2
27	30	123	12	144	156	17	0
30	36	111	19	156	180	6	0

(a) Suppose the time T to the second child has waiting time function $S(t) = \Pr(T \geq t)$, with density $f(t)$. On the time-continuous scale, there is an intensity function or hazard rate $\alpha(t) = f(t)/S(t)$. For our grouped data, show that

$$h_j = \Pr(T \in [l_j, r_j] \,|\, T \geq l_j) = \{S(l_j) - S(r_j)\}/S(l_j) = 1 - S(r_j)/S(l_j).$$

Compute the nonparametric estimates $\widehat{h}_j = (\Delta N_j/Y_j)/(r_j - l_j)$, and plot them, with $\widehat{h}_j$ tied to the the midpoint $m_j = \frac{1}{2}(l_j + r_j)$.

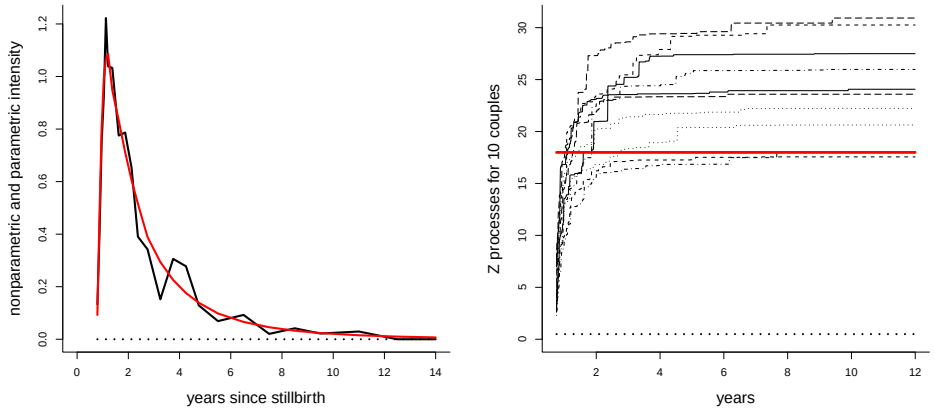


Figure i.12: *Left panel: nonparametric and parametric estimates of the intensity curves for time to 2nd child after stillborn 1st child, with the gamma process threshold crossing model. Right panel: simulated gamma processes, for ten envisaged married couples; these need to cross* $\widehat{d} = 17.494$ *to have a 2nd child.*

(b) Then consider a parametric model $S(t, \theta)$ for the waiting time, with ensuing intensitis $h_j(\theta) = 1 - S(r_j, \theta)/S(l_j, \theta)$ for the n time windowss $[l_j, r_j]$. Show that the log-likelihood function can be written

$$\ell(\theta) = \sum_{i=1}^{n} [\Delta N_j \log h_j(\theta) + (Y_j - \Delta N_j) \log\{1 - h_j(\theta)\}].$$

(xx a bit more here. the same as a sequence of nested binomials, though these are not independent. but large-sample theory works. Hjort and Lumley (1993). calibrate with how we present things elsewhere. xx)

(c) The model we work with now is of the Gamma process threshold crossing type, a variant of what we used for the Roman era Egypt lifetime analyses in Story ii.11. Specifically, the time T to 2nd child is seen as the time it takes a gamma process $Z(t)$ to cross threshold d, where $Z(t) \sim \text{Gam}(aM(t), 1)$, and the motor function is taken as a Weibull: $M(t) = 1 - \exp[-\{(t - t_0)/b\}^c]$, where $t_0 = 9/12 = 0.75$. With $G(x, a, 1)$ the c.d.f. of a $\text{Gam}(a, 1)$, show that $S(t) = \Pr(T \geq t) = G(d, aM(t), 1)$, leading to $h_j(\theta) = 1 - G(d, aM(r_j), 1)/G(d, aM(l_j), 1)$. Maximise the log-likelihood; you should find parameter estimates $(23.721, 0.400, 0.495, 17.494)$ for (a, b, c, d). Construct a version of Figure i.12, left panel. Then simulate Gamma processes, checking for each whether they cross $\widehat{d}$, and make a version of the figure's right panel.

(d) Plot $\widehat{F}(t) = F(t, \widehat{a}, \widehat{b}, \widehat{c}, \widehat{d})$, a cumulative curve that will not reach 1; not all gamma processes $\text{Gam}(aM(t), 1)$ will succeed in crossing the threshold d. Show that the fraction of couples who not have a second child, with this model, is $p = G(d, a, 1)$, here estimated at 8.9 percent. Find a 90 percent confidence interval for this p.

(e) There are clearly several different models which might work well for the nested binomials above, perhaps motivated from different perspectives. A rich class of waiting time distributions emerges via the multiplicative frailty constructions worked with in Ex. 10.15. Assuming that couples have intensity function $Z\alpha_0(s)$ for having a second child, with $Z = \sum_{i=1}^{N} X_i$ of compound Poisson type, show that

$$S(t) = \exp[-\lambda\{1 - L_0(A(t))\}], \quad \text{in terms of } L_0(u) = \mathrm{E}\exp(-uX_i),$$

with λ the Poisson parameter. If the X_i are positive, show that the fraction of couples not having second child is $p = \exp(-\lambda)$.

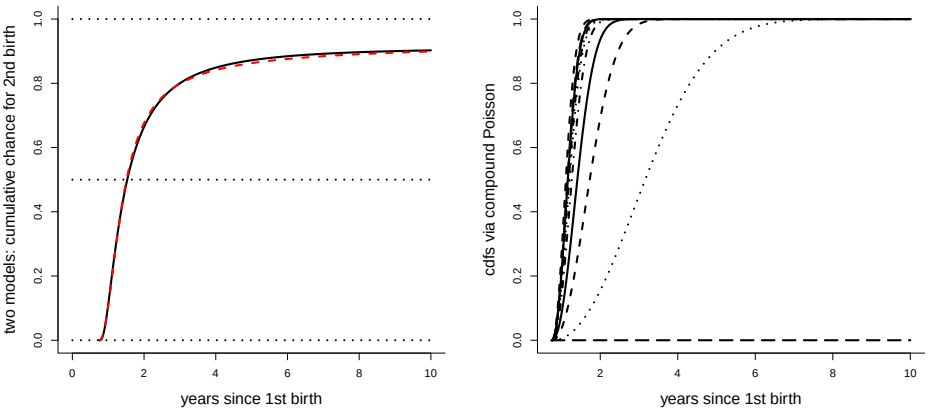


Figure i.13: *Left panel: the Rashomon syndrome: two rather different models for the cumulative waiting time function* $F(t) = \Pr(T \le t)$ *lead to almost identical estimated curves. The median time for having a second child as about 1.52 years. Right panel: ten simulated c.d.f.s* $F(t) = 1 - \exp\{-ZA_0(t)\}$*, via the compound Poisson frailty hazard model. One of these is flat at zero, implying no 2nd child will come.*

(f) Now construct a four-parameter model as follows. We start with a Weibull distribution, with intensity $\alpha_0(t) = a(t-t_0)^k$ and cumulative intensity $A_0(t) = a(t-t_0)^{k+1}/(k+1)$. Then postulate that the couples in the population studied have intensity functions $Z\alpha_0(t)$, where $Z = \sum_{i=1}^{N} X_i$ is compound Poisson, with the $X_i \sim \mathrm{Gam}(c, c)$ having a fixed mean 1, to avoid overparametrisation. Show that

$$S(t) = \exp\Big(-\lambda\Big[1 - \Big\{\frac{c}{c + A_0(t)}\Big\}^c\Big]\Big).$$

(This model is similar to the one proposed by Aalen (1992) for these data, but here we use a perhaps more straightforward model construction, with a different parametrisation.) Programme and maximise the log-likelihood function, which should give parameter estimates $(1.865, 1.152, 0.305, 2.756)$ for (a, k, c, λ). Estimate the not having a second child fraction p, with a confidence interval, and compare with what was obtained above for the gamma process threshold crossing model. Also, plot the estimated cumulative waiting time function $F(t) = \Pr(T \le t)$, for both the gamma process and the compound Poisson frailty models, producing Figure i.13, left panel. Observe and discuss aspects of the

the Rashomon effect

fact that these two curves are almost identical. In a rather diffierent context and setup, Breiman (2001, Section 8) names this *the Rashomon effect*, or syndrome; quite different models, built from different perspectives, might in the end fit data almost equally well. In the eponymous 1950 Japanese film, different eyewitnesses report very different and partly contradictory interpretations of the very same event. The lesson is that even though a model fits and 'explains' data, the cautious quotation marks for the explain part might be needed. We cannot know, from the data alone, whether the processes leading to a second child somehow reflect underlying ingredients better explained by cumulative gamma processes or by frailties following a compound Poisson. (xx check and do a bit cross-pointing for other places where we use the compound poisson model, Ch1, Ch10. xx)

(g) For the compound Poisson frailty hazard function model above, simulate say 100 realisations of $Z = \sum_{i=1}^{N} X_i$, and from these realisations of c.d.f.s $F(t) = 1-\exp\{-ZA_0(t)\}$ for the time to 2nd child, using the estimated values of (a, k, c, λ). Make a version of Figure i.13, right panel, which shows ten such c.d.f.s., one of which is simply zero, indicating there will not be a 2nd child for that couple.

(h) (xx should put in just a little more. how do we know that the $\mathrm{N}_p(\theta_9, \widehat{J}^{-1})$ recipe works here; perhaps martingales. also: study time to 2nd child given that there will be one, i.e. condition on $T < \infty$. this is conditioning on $N \geq 1$ for the aalen model, and conditioning on $Z(\infty) > d$ for the nils model. can simulate from these two estimated conditional distributions. xx)

Story i.10 *A third child?* [xx the R-script for this story is `survivalstory_emil1.R` xx] Do parents want to mix the sexes? The units in the data set `third_births.txt` are mothers of two (the data stems from the Medical Birth Registry of Norway, and is openly available on the website of the book Aalen et al. (2008)). The data set contains the age of the mother at the first birth; the number of days between the births of the first and second child; the genders of her first two children; the number of days from the second to the third birth or censoring, which will be our outcome of interest; and an indicator of noncensoring.

(a) To investigate whether parents want both girls and boys, we can compare the probability of having a third child for the families with two girls or two boys, to the probability of having a third child among those with a girl and a boy. The time to a third child or censoring is given in days, so it is perhaps natural to stay in discrete time. Let $x = I\{\text{two girls or two boys}\}$, and $t = \min(t^*, c)$ where t^* is the true time to a third child, and c a censoring time. Assume that our data are n independent replicates of (t, δ, x), where δ is an idicator of noncensoring. Estimate the two survival functions $\Pr(t \geq j \,|\, x = 1)$ and $\Pr(t \geq j \,|\, x = 0)$, and also the two cumulative hazards $A_j(x) = \sum_{i=1}^{j} \Pr(t = i \,|\, t \geq i, x)$ for $x = 0, 1$, both using the maximiser of the likelihood the discrete time likelihood in Ex. 10.8(c). These two estimators are the discrete time analogues of the Kaplan–Meier estimator and the Nelson–Aalen estimator. Plot these in two different plots, and comment on your findings.

(b) (a) xx] [xx Greenwood's formula and add pointwise confidence bands to the two plots from

(c) Instead of estimating separate hazards for the two groups, as in (a), we'll now try out a few different regression approaches. For all the regression models, $\alpha_{i,j}$ is the hazard

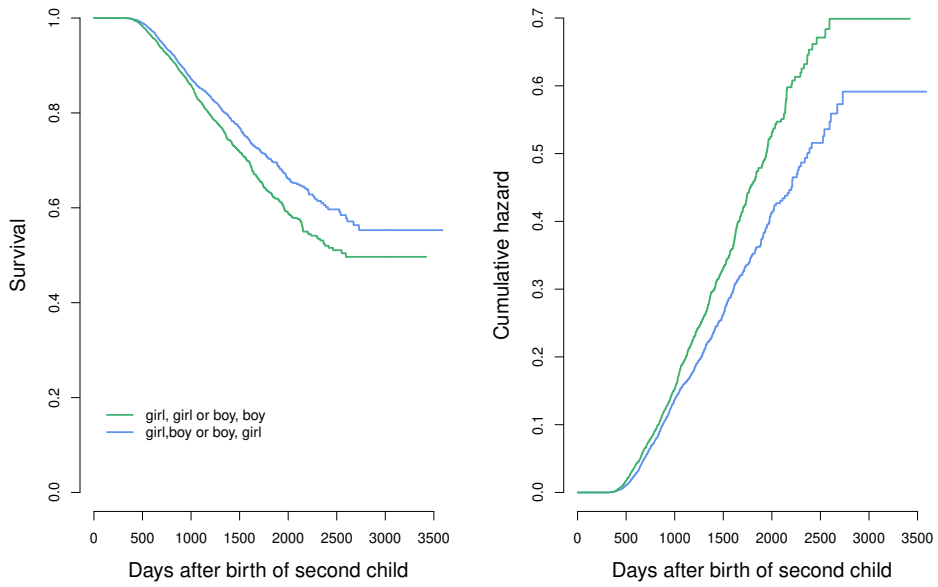


Figure i.14: *[xx text here. perhaps also add confidence bands to these xx]*

of the ith mother at time j, we think of γ_j as a baseline hazard at time j. To allow for some generality of the results we reach, we let $x_i = (x_{i,1}, \ldots, x_{i,p})^{\rm t}$ be a vector of covariates for the ith mother, though, in the setting of this exercise, we'll simply have $x_i = I\{\text{two girls or two boys}\}$. We start out with discrete time relative risk regression models, that is, models where the hazard of the ith mother at time j is

$$\alpha_{i,j} = \gamma_j r(x_i^{\rm t}\beta). \tag{i.2}$$

Taking $r(z) = \exp(z)$ leads to a Cox type model. In the discrete time setup studied here, however, where hazard is a probability (not just a nonnegative function, as in the continuous case), models where $r(z)$ takes values between zero and one might be more theoretically sound. The model $r(z) = 1/\{1 + \exp(-z)\}$ is one such (a continuous time version of this latter model was studied in a Bayesian framework in De Blasi and Hjort (2007)). The parameters of models of the form (i.2) may in principle be estimated by findind the maximisers of the likelihood function

$$L(\gamma, \beta) = \prod_{j \geq 0} \prod_{i=1}^{n} \alpha_{i,j}^{\Delta N_{i,j}} (1 - \alpha_{i,j})^{Y_{i,j} - \Delta N_{i,j}}, \tag{i.3}$$

since this product is really a finite one, every factor after the largest survival time being equal to 1. With $x_i = I\{\text{two girls or two boys}\}$, try to estimate of the parameters of (i.2) with $r(z) = \exp(z)$ and $r(z) = 1/\{1 + \exp(-z)\}$ using the maximum likelihood method. Compare the estimated survival functions from the two models. Proceed directly to the next subpoint if you don't find sensible estimates.

(d) Estimating $\gamma_0, \gamma_1, \ldots$ and β in one go can be quite burdensome. A procedure that is operationally more smooth, is to first estimate the β parameter(s) using a discrete time

version of Cox's partial likelihood. That is, first we find the minimiser $\widehat{\beta}$ of (evocatively, we here use notation from Ex. **??**)

$$H_n(\beta) = -n^{-1} \sum_{j\geq 0} \sum_{i=1}^{n} \{\log r(x_i^{\mathrm{t}}\beta) - \log\big(nS_{n,j}^{(0)}(\beta)\big)\}\Delta N_{i,j},$$

where $S_{n,j}^{(0)}(\beta) = n^{-1}\sum_{i=1}^{n} Y_{i,j} r(x_i^{\mathrm{t}}\beta)$, and then estimate γ_j with a discrete time analogue of the Breslow-estimator, $\widehat{\gamma}_j = \Delta N_j/\{\sum_{i=1}^{n} Y_{i,j} r(x_i^{\mathrm{t}}\widehat{\beta})\}$. Note that, contrary to the continuous time (no-ties) scenario under which Cox's partial likelihood was derived in Ex. 10.12, the function $m(\beta)$ is neither a partial- nor a profile- likelihood. That does not prevent it, however, from producing good estimates. Fit one of the two models from (c) to the data, and plot the two survival functions (i.e. for the $x_i = 0$ and $x_i = 1$ mothers) based on the Breslow-type estimates.

(e) We'll now see why and in what sense the estimation method from (d) works. For some nonnegative function $r(z)$, assume that the data truly stem from a model with hazard rate $\alpha_{i,j} = \gamma_j^\circ r(x_i^{\mathrm{t}}\beta^\circ)$. Introduce $S_{n,j}^{(1)}(\beta) = \sum_{i=1}^{n} x_i Y_{i,j} r'(x_i^{\mathrm{t}}\beta)$, and $S_{n,j}^{(2)}(\beta) = \sum_{i=1}^{n} x_i x_i^{\mathrm{t}} Y_{i,j} r'(x_i^{\mathrm{t}}\beta)/r(x_i^{\mathrm{t}}\beta)$ where $r'(z) = \partial r(z)/\partial z$.

[xx hit xx] Show that the first derivative of $m(\beta)$ evaluated in the true value β° is the mean zero martingale

$$\frac{\partial}{\partial\beta}\log m(\beta^\circ)_\tau = \sum_{i=1}^{n}\sum_{j=0}^{\tau}\left(x_i\frac{r'(x_i^{\mathrm{t}}\beta^\circ)}{r(x_i^{\mathrm{t}}\beta^\circ)} - \frac{s_{n,j}^{(1)}(\beta^\circ)}{s_{n,j}^{(0)}(\beta^\circ)}\right)\Delta M_{i,j},$$

where $M_{i,j}$ is the discrete time martingale $M_{i,j} = N_{i,j} - \sum_{\ell=0}^{j} Y_{i,j}\alpha_{i,j}$, as defined in (10.2). You may also show that the second derivative of $m(\beta)$ evaluated in the true parameter values is

$$\frac{\partial^2}{\partial\beta\partial\beta^{\mathrm{t}}}\log m(\beta)_\tau = -\sum_{j=0}^{\tau}\Big\{\frac{s_{n,j}^{(2)}(\beta^\circ)}{s_{n,j}^{(0)}(\beta^\circ)} - \frac{s_{n,j}^{(1)}(\beta^\circ)(s_{n,j}^{(1)}(\beta^\circ))^{\mathrm{t}}}{s_{n,j}^{(0)}(\beta^\circ)^2}\Big\}s_{n,j}^{(0)}(\beta^\circ)\gamma_j,$$

which gives the following relation between the predictable variation of $\partial\log m(\beta^\circ)/\partial\beta$ and its derivative,

$$\langle\frac{\partial}{\partial\beta}\log m(\beta^\circ), \frac{\partial}{\partial\beta}\log m(\beta^\circ)\rangle_\tau = -\frac{\partial^2}{\partial\beta\partial\beta^{\mathrm{t}}}\log m(\beta^\circ)_\tau + \varepsilon_{n,\tau},$$

where $\varepsilon_{n,\tau}$ is a certain mean zero martingale. Look at Lenglart's inequality in Ex. 10.7 and the central limit theorem in Ex. 10.9, and sketch a limit-in-distribution result for the maximiser $\widehat{\beta}$ of the $m(\beta)$.

(f) Back to our application with $x_i = I\{\text{two girls or two boys}\}$. Let the hazard of the ith mother at time j be $\alpha_{i,j} = \gamma_j \exp(x_i\beta)/\{1+\exp(x_i\beta)\}$, or with some other $r(x_i\beta)$ if you like. Use the results from (e) to test the hypothesis $\beta^\circ = 0$ versus its two sided alternative. Make a plot of the survival functions $\Pr(T \geq j \,|\, x = 0)$ and $\Pr(T \geq j \,|\, x = 1)$.

(g) Is there a difference between the two girls and the two boys families? Let now $x_i = I\{\text{two girls}\}$ and $z_i = I\{\text{two girls}\}$, and suppose the hazard of the ith mother at time j be $\alpha_{i,j} = \gamma_j \exp(x_i\beta + z_i\eta)/\{1+\exp(x_i\beta + z_i\eta)\}$, or, again with your favourite $r(\cdot)$ function. Test the relevant hypotheses. [xx be more precise xx]

Story i.11 *PCI and rare events confidence fusion of thirteen studies*
Data: `thirteen-tables-data` (xx to be organised xx)
Code: `(xx com131* com132* com133* nilswork22 xx)`
Percutaneous Coronary Intervention is a so-called minimally invasive procedure, used for certain patients with stable coronary disease, to open clogged arteries. There is a literature pertaining to whether PCI lowers the risk of future death cardiac death, compared to recieving medical treatment alone. A very firm conclusion has apparently not been reached. Note that several traditional statistical methods do not work well here, since it is harder to compare probabilities for rare events than for events with probabilities say bigger than 0.15.

Organising information from Schömig et al. (2008) we find the table below, from thirteen different studies; see also Zabriskie et al. (2021). We translate this to thirteen two by two tables, with $y_{i,0}$ and $y_{i,1}$ the number of cardiac deaths, out of respectively $m_{i,0}$ and $m_{i,1}$ patients. Group 0 is that receiving medical treatment alone, and Group 1 that with PCI. The overall question addressed here is whether the risk has been reduced for Group 1, compared to Group 0, also taking on board that the implied probabilities $p_{i,0}$ and $p_{i,1}$ may vary across studies. The overall rates of cardiac death, for patient groups 0 and 1, are 5.4 and 4.1 percent. (xx need to calibrate and cross-check. we do have (p_0, p_1) for one table in Ex. 4.30; perhaps we touch this in Ch5; then to be rounded off with CD and cc in Ex. 7.35. question is whether we have the many-tables thing in Ch7 or do it directly here. xx)

	y1	m1	y0	m0
Sievers et al.	0	44	1	44
Dakik et al.	1	21	1	23
AVERT	1	177	1	164
MASS	4	72	2	72
Bech et al.	1	90	2	91
ALKK	4	149	14	151
RITA-2	20	504	24	514
TIME	32	153	33	148
Hambrecht et al.	0	50	0	51
INSPIRE	2	104	1	101
MASS II	24	205	25	203
SWISSI II	3	96	22	105
COURAGE	23	1149	25	1138

(a) Under natural rules and conditions for how patients enter treatment and are then being sorted into the two groups, adhered to in these studies, we may take $y_{i,0}$ and $y_{i,1}$ as outcomes of $Y_{i,0} \sim \text{binom}(m_{i,0}, p_{i,0})$ and $Y_{i,1} \sim \text{binom}(m_{i,1}, p_{i,1})$. Starting then with one 2×2 table of such binomial outcomes at the time, we use the logistic representation, as in Ex. 4.30 and 7.35, to write $p_{i,0} = H(\theta_i)$ and $p_{i,1} = H(\theta_i + \gamma_i)$, with $H(u) = \exp(u)/\{1 + \exp(u)\}$. The focus is on γ_i, the log-odds difference, or equivalently on $\rho_i = \exp(\gamma_i)$, the odds ratio, and whether it is negative or not. Compute and display what is according to Ex. 7.35 the optimal CD for γ_i,

$$C_i(\gamma_i) = \Pr_{\gamma_i}(Y_i > y_{i,\text{obs}} \,|\, Z_i = z_{i,\text{obs}}) + \tfrac{1}{2} \Pr_{\gamma_i}(Y_i = y_{i,\text{obs}} \,|\, Z_i = z_{i,\text{obs}}),$$

with $Z_i = Y_{i,0} + Y_{i,1}$; this involves the excentric hypergeometric distribution. Convert these to confidence curves $\text{cc}_i(\gamma_i)$, as in the left panel of Figure i.15. Note that most of

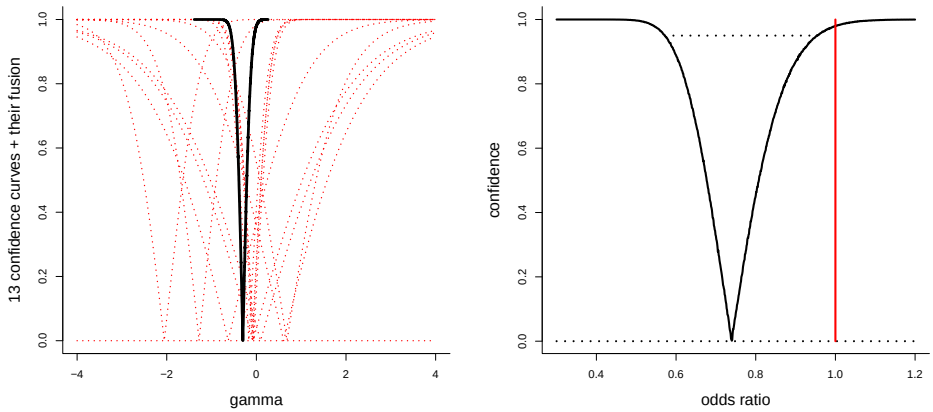


Figure i.15: *Left panel: confidence curves* $\mathrm{cc}(\gamma)$ *from thirteen separate studies, along with the fusion (black, fuller curve); the question is if* $\gamma < 0$ *or not. Right panel: confidence cruve* $\mathrm{cc}^*(\rho)$ *for the odds ratio* $\rho = \exp(\gamma)$*; the question is if* $\rho < 1$ *or not. The 95 percent interval is* $[0.572, 0.954]$*, and the median confidence estimate is 0.740.*

the 95 percent confidence intervals include zero, and with median confidence estimates on both sides of zero; these studies when taken separately do not lend strong support to PCI being beneficial, i.e. that $\gamma_i < 0$.

(b) Now make the additional assumption that the γ_i are about equal, to some common γ; the probabilities $(p_{i,0}, p_{i,1})$ are still seen as varying across studies. With $k = 13$ for the number of 2×2 tables, set up the log-likelihood for the $2k$ observations, in terms of the $k + 1$ parameters. Show from (xx stuff in Ch7 xx) that there is an optimal CD for γ, taking the form

$$\begin{aligned} C^*(\gamma) \,=\, & \Pr_\gamma\{S > S_{\mathrm{obs}} \,|\, Z_1 = z_{1,\mathrm{obs}}, \ldots, Z_k = z_{k,\mathrm{obs}}\} \\ & + \tfrac{1}{2} \Pr_\gamma\{S = S_{\mathrm{obs}} \,|\, Z_1 = z_{1,\mathrm{obs}}, \ldots, Z_k = z_{k,\mathrm{obs}}\}. \end{aligned}$$

Here $S = \sum_{i=1}^k Y_{i,1}$, with $S_{\mathrm{obs}} = 115$. Computing these probabilities is for each given γ achieved by simulating $Y_{i,1} \,|\, z_{i,\mathrm{obs}}$, for $i = 1, \ldots, k$, then summing these to get S, and repeating this a high number of times. Carry out this, producing both the tight fusion confidence curve $\mathrm{cc}^*(\gamma)$, shown in bold in the left panel of Figure i.15, and the corresponding fusion confidence curve $\mathrm{cc}^*(\rho)$, in the right panel. You should find $C^*(0) = 0.99$. Discuss the implications of this, and of seeing most of the mass to the left of zero.

(c) Binomials with small probabilities are close to Poisson, see Ex. 2.8, so one may alternatively view the meta-analysis data as Poisson pairs $(Y_{i,0}, Y_{i,1})$. It is practical to normalise the parameters as $(e_{i,0}\lambda_i, e_{i,1}\lambda_i\kappa_i)$, in terms of $(e_{i,0}, e_{i,1}) = (m_{i,0}/100, m_{i,1}/100)$, so that λ_i and $\lambda_i\kappa_i$ are the rates per 100 patients, for the two groups; see also Cunen and Hjort (2015). Show that the log-likelihood for pair i takes the form $\ell_i = -\lambda_i(e_{i,0} + e_{i,1}\kappa_i) + z_i \log\lambda_i + y_{i,1}\log\kappa_i$, plus terms not depending on the parameters. Use (xx pointer to Ch8 xx) to argue that the optimal confidence distribution for κ_i, based on this data pair, takes the form

$$C_i(\kappa_i) = \Pr_{\kappa_i}(Y_i > y_{i,\mathrm{obs}} \,|\, Z_i = z_{i,\mathrm{obs}}) + \tfrac{1}{2} \Pr_{\kappa_i}(Y_i = y_{i,\mathrm{obs}} \,|\, Z_i = z_{i,\mathrm{obs}}).$$

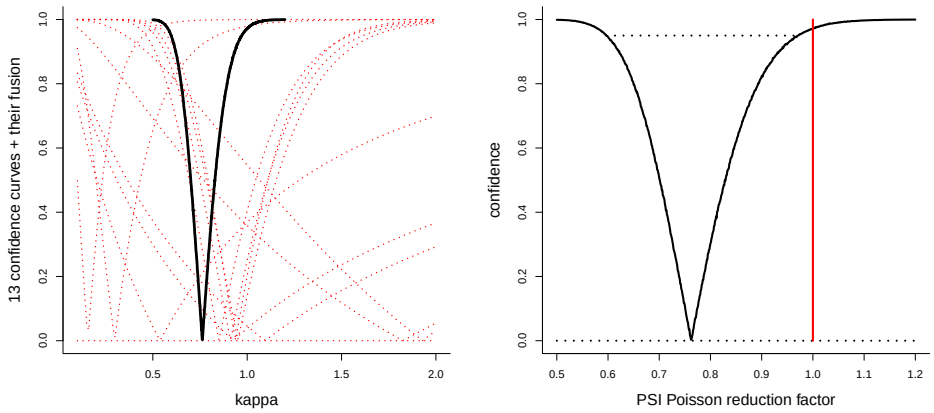


Figure i.16: *As with Figure i.15, but now with the meta-analysis data seen as Poisson pairs* $(\lambda_i, \lambda_i\kappa)$. *Left panel: confidence curves* $\mathrm{cc}(\kappa)$ *from thirteen separate studies, along with the fusion (black, fuller curve); the question is if* $\kappa < 1$ *or not. Right panel, zooming in: confidence curve* $\mathrm{cc}^*(\kappa)$, *with 95 percent interval* $[0.597, 0.971]$, *and median confidence estimate 0.762.*

This is partly as for the γ_i case above, but show now that the conditional distributions $Y_{i,1} \,|\, z_i$ are binomial $(z_i, e_{i,1}\kappa_i/(e_{i,0} + e_{i,1}\kappa_i))$. Compute and display the confidence distributions and the confidence curves $\mathrm{cc}_i(\kappa_i)$. (xx check Figure i.16. xx)

(d) Consider then the fixed effects Poisson pairs model, with $k+1$ parameters, with a common PCI Poisson factor κ. Show that the log-likelihood function becomes

$$\ell = \sum_{i=1}^{k} \{-(e_{i,0} + e_{i,1}\kappa)\lambda_i + z_i \log \lambda_i + y_{i,1} \log \kappa\}.$$

Show that the optimal CD for this κ becomes

$$\begin{aligned} C^*(\kappa) \,=\, & \Pr_\kappa\{S > S_{\mathrm{obs}} \,|\, Z_1 = z_{1,\mathrm{obs}}, \ldots, Z_k = z_{k,\mathrm{obs}}\} \\ & + \tfrac{1}{2} \Pr_\kappa\{S = S_{\mathrm{obs}} \,|\, Z_1 = z_{1,\mathrm{obs}}, \ldots, Z_k = z_{k,\mathrm{obs}}\}. \end{aligned}$$

As above, to compute this, one needs to simulate a high number of $S = \sum_{i=1}^{k} Y_{i,1}$ conditional on $z_{1,\mathrm{obs}}, \ldots, z_{k,\mathrm{obs}}$. (xx compute, display, comment, find interval, construct a version of Figure i.16, compute $C^*(1)$. xx)

(e) (xx something on not-constant κ, with variance heterogeneity there too; nils rant, so far. xx) Suppose now that the κ_i also vary across studies. To model this, take these to stem from a Gamma distribution, with parameters $(a, b) = (\kappa_0/c, 1/c)$, such that the mean is κ_0 and the variance $c\kappa_0$; the homogeneous case corresponds to $c = 0$. Writing $g(y, \theta)$ for the Poisson probabilities with parameter θ, show for the distribution of the i the pair that

$$\begin{aligned} \bar f(y_{i,0}, y_{i,1}) \,&=\, \int_0^\infty g(y_{i,0}, e_{i,0}\lambda_i) g(y_{i,1}, e_{i,1}\lambda_i\kappa_i) \pi(\kappa_i)\, \mathrm{d}\kappa_i \\ &=\, \exp\{-(e_{i,0}\lambda_i)\} \lambda_i^{z_i} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + y_{i,1})}{(b + e_{i,1}\lambda_i)^{a+y_{i,1}}} \frac{e_{i,0}^{y_{i,0}}\, e_{i,1}^{y_{i,1}}}{y_{i,0}!\, y_{i,1}!}. \end{aligned}$$

Write down the resulting log-likelihood function $\ell(\lambda_1, \ldots, \lambda_k, \kappa_0, c)$, for the k+2-parameter model. Use this to also show that the profiled log-likelihood $\ell_{\rm prof}(\kappa_0, c)$, where one maximises of $\lambda_1, \ldots, \lambda_k$, can be written $\sum_{i=1}^k \widehat{A}_i(\kappa_0, c)$, say, where $\widehat{A}_i(\kappa_0, c)$ is

$$A_i = -e_{i,0}\lambda_i + z_i \log \lambda_i + \log \Gamma(a + y_{i,1}) - \log \Gamma(a) + a \log b - (a + y_{i,1}) \log(b + e_{i,1}\lambda_i),$$

inserting the maximiser $\widehat{\lambda}_i = \widehat{\lambda}_i(\kappa_0, c)$ for λ_i. Find indeed an expression for this maximiser, by solving a quadratic equation; implement the strategy, and compute $\ell_{\rm prof}(\kappa_0, c)$ for some values. (xx nils jots down details, part of the solutions, i suppose. xx) taking the derivative, $\widehat{\lambda}_i$ is the solution to

$$\frac{z_i}{\lambda_i} - \frac{a + y_{i,1}}{b + e_{i,1}\lambda_i} = e_{i,0}, \quad \text{or} \quad z_i(b + e_{i,1}\lambda_i) - (a + y_{i,1})\lambda_i = e_{i,0}\lambda_i(b + e_{i,1}\lambda_i),$$

which can be organised to (xx check this xx)

$$e_{i,0}e_{i,1}\lambda_i^2 + \lambda_i(e_{i,0}b + a + y_{i,1} - e_{i,1}z_i) - z_i b = 0.$$

special case $z_i = 0$: then contribution is zero. Compute the $\ell_{\rm prof}(c)$, and demonstrate that for this particular dataset, there is no sign of variance heterogeneity for κ; the maximum likelihood estimate of $\widehat{c} = 0$. (xx hm, things to check, could be fine with an example with positive c, to see the estimated $\pi(\kappa)$, etc. check ML in $k + 1$-parameter model. xx)

(f) (xx a separate little point about 'what do do with zero cases', where different opinions have been voiced in the literature. here it comes from clear math: the $(0, 0)$ cases can be discarded from the analysis. xx)

(g)

Story i.12 *Suicide attempt rates for Paroxetine vs. placebo*
Data: `in text below`
Code: `xx (xx to come com96* nilswork22 xx)`
There are several studies of the effects and side effects of the antidepressant Paroxetine (sold under brand names Seroxat, Paxil, and yet others, since 1992). While beneficial for hundreds of thousands of users, serious concerns are also part of the broader picture, with one particularly disturbing aspect being its potential association with suicidal thoughts and actions. Here we use data and information from Aursnes et al. (2005, 2006), who used Bayesian analyses with informative priors, based on data and other information available to those authors in respectively 2005 and 2006. Below we discuss these priors to posteriors calculations, but also include other non-Bayesian methods.

The data are as simple as two Poisson counts, $Y_0 \sim \text{Pois}(m_0\theta_0)$ and $Y_1 \sim \text{Pois}(m_1\theta_1)$, for the placebo and the drug groups, with m_0 and m_1 cumulative exposure time, here conveniently counted as patient years. The parameter of primary interest is $\gamma = \theta_1/\theta_0$. The articles pointed to concentrate on the probability that $\gamma > 1$, or, equivalently, that $\kappa > 0$, where $\kappa = \log(\theta_1/\theta_0)$ is a more convenient scale for computation and summary reporting, due the inherent strong right skewnesses involved on the γ scale.

For the studies in question, the 2005 article had $(y_0, y_1) = (1, 7)$, after $(m_0, m_1) = (73.3, 190.7)$ patient years, and used informative priors based on previous literature to conclude that it was rather likely that $\theta_1 > \theta_0$, i.e. an increased suicide attempt risk in the Paroxetine group. This was followed by media exposure and debate, along with

critical comments from both individual researchers and from GlaxoSmithKline plc, the multinational pharmaceutical and biotechnology company manufacturing the drug. This again led to the 2006 article, by the same four authors, with more extensive data collection and also further care for accuracy. In summary, the data now had $(y_0, y_1) = (1, 11)$, after $(m_0, m_1) = (333.0, 601.0)$ patient years. The 2005 data are to be seen as part of the extended and more accurately curated 2006 dataset.

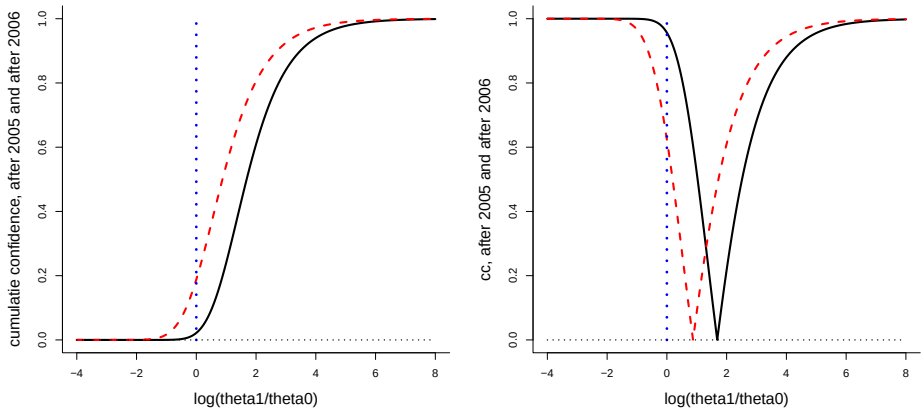


Figure i.17: *Cumulative confidence distributions (left panel) and confidence curves (right panel), for* $\kappa = \log(\theta_1/\theta_0)$*, based on information in the 2005 article (red, slanted) and in the 2006 article (black, full). The central question is whether* $\theta_1 > \theta_0$*, i.e. whether* $\kappa > 0$*. The evidence for this is much clearer with the 2006 information.*

(a) With $Z = Y_0 + Y_1$, show that $Y_1 \,|\, z \sim \text{binom}(z, m_1\gamma/(m_0 + m_1\gamma))$, with $\gamma = \exp(\kappa)$. Compute and display what is according to Ex. 7.37 the optimal confidence distribution,

$$C(\kappa) = \Pr_\kappa(Y_1 > y_{1,\text{obs}} \,|\, Y_0 + Y_1 = z_{\text{obs}}) + \tfrac{1}{2}\Pr_\kappa(Y_1 = y_{1,\text{obs}} \,|\, Y_0 + Y_1 = z_{\text{obs}}),$$

with the 2005-information and the 2006-information; construct versions of Figure i.17, both on the κ and γ scales. Verify in particular that $C_{2005}(0) = 0.188$ and $C_{2006}(0) = 0.022$. Explain how these can be seen as p-values for testing $\theta_0 \le \theta_1$ against the drastic alternative that the antidepressant in question increases the suicide attempt risk. Discuss also how the complementary numbers 0.812 and 0.978 can be seen as epistemic probabiities for $\theta_1 > \theta_0$. Give also 95 percent confidence intervals, first for κ and then transformed back to the scale of θ_1/θ_0.

(b) Suppose now that adequate prior distributions are set of the type $\theta_0 \sim \text{Gam}(a_0, b_0)$ and $\theta_1 \sim \text{Gam}(a_1, b_1)$. Show that this leads to clear posterior distributions

$$\theta_0 \,|\, \text{data} \sim \text{Gam}(a_0 + y_0, b_0 + m_0), \qquad \theta_1 \,|\, \text{data} \sim \text{Gam}(a_1 + y_1, b_1 + m_1).$$

Show that the posterior cumulative and density functions for $\kappa = \log(\theta_1/\theta_0)$ can be expressed as

$$F(\kappa \,|\, \text{data}) = \int_0^\infty G(\exp(\kappa)\theta_0, a_1 + y_1, b_1 + m_1)\, g(\theta_0, a_0 + y_0, b_0 + m_0)\, \mathrm{d}\theta_0,$$
$$f(\kappa \,|\, \text{data}) = \int_0^\infty g(\exp(\kappa)\theta_0, a_1 + y_1, b_1 + m_1) \exp(\kappa)\theta_0\, g(\theta_0, a_0 + y_0, b_0 + m_0)\, \mathrm{d}\theta_0,$$

in terms of the cumulative and density $G(\cdot, a, b)$ and $g(\cdot, a, b)$ of the Gam(a, b). In particular, explain that $p_B = 1 - F(0 \,|\, \text{data})$ is the posterior probability for the dramatic $\theta_1 > \theta_0$ scenario, building on both the priors and the Poisson counts (y_0, y_1) with (m_0, m_1) patient years. Give also corresponding expressions for the posterior cumulative and density on the direct scale of $\gamma = \theta_1/\theta_0$.

(c) For each of the 2005-infomation and 2006-information cases, compute and display

$$p_B = 1 - \int_0^\infty G(\theta_0, a + y_1, 50 + m_1)\, g(\theta_0, a + y_0, 50 + m_1)\, \mathrm{d}\theta_0$$

as a function of a, a common parameter in gamma prior parameters $(a, 50), (a, 50)$ for θ_0, θ_1, interpreted as the expected number of suicide attempts in the course of 50 patient years, for either the placebo or drug groups of patients. Comment on your findings.

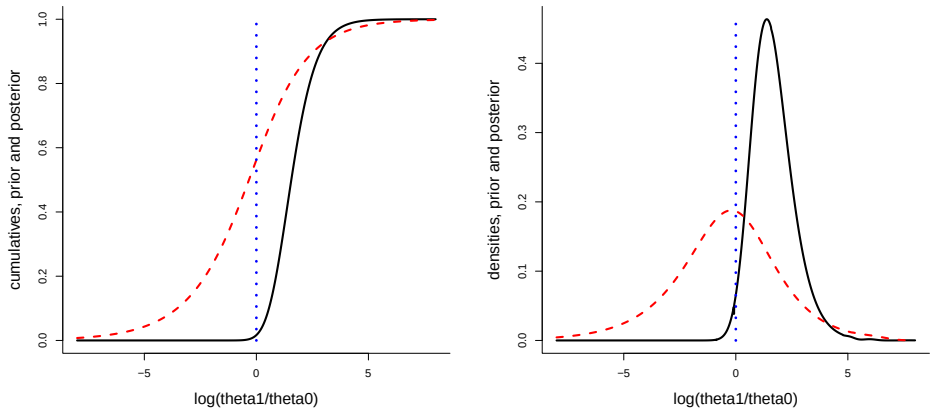


Figure i.18: *From the informative slightly optimistic prior (red, slanted) to the posterior (black, full), using the 2006 data; cumulatives in left panel, densities in right panel. The evidence is very strong that* $\theta_1 > \theta_0$.

(d) Several informative priors are carefully argued for and worked with in Aursnes et al. (2005, 2006). "This does not mean that these parameters are to be interpreted as random variables, but our knowledge of the parameters is uncertain and we describe this uncertainty with the help of probability distributions," as they write, when setting their priors, in fact by attempting to match conclusions of earlier meta-analysis publications to gamma prior parameters. For illustration in the present story we are content with using one of these, called by them in their 2006 article the slightly optimistic prior, having $(a_0, b_0) = (0.71, 50)$ and $(a_1, b_1) = (0.58, 50)$. The idea was to quantify the expected number of suicide attempts for the placebo and the drug groups, in the course of 50 patient years, and with these expected numbers adding up to 1.29 attempts per 100 patient years, matching information recorded in previous literature. For this prior, work through the numerics, and display the prior and posterior densities, as well as the prior and posterior cumulatives, for the focus parameter $\log(\theta_1/\theta_0)$; construct versions of Figure i.18. Find the 95 percent posterior interval for κ, and by transformation for θ_1/θ_0. Also, record the Bayesian answer p_B to the question of how likely we should think it is that the drug increases suicide attempt risk.

(e) Results reported on in Aursnes et al. (2005, 2006) were not reached via the precise integration tools above; the authors resorted rather to simulation. Carry out such work too, simulating say 10^5 realisations of $\log(\theta_2/\theta_1)$ from the posterior distribution, followed by simple density estimation, to reach a simulated version of Figure i.18. As explained via the integration details above, however, there is no real need for simulation here.

(f) (xx something more. try noninformative priors, of the type $(0.1, 0.1)$ and $(0.1, 0.1)$ for θ_0 and θ_1. and something more neutral, like $(1, 50)$ and $(1, 50)$. xx)

(g) (xx something iiccff using Cunen and Hjort (2022). combining one of these informative priors with the new data. several paths possible. (i) using the informative priors for θ_0 and θ_1, then log-likelihoods for θ_0 and θ_1, i.e. four sources combined to find $\text{cc}^*(\kappa)$. (ii) using the log-prior for κ coming out of two priors, and the converted log-likelihood for κ coming from $\text{cc}(\kappa)$. both should work, but four sources in detail might be a bit more precise. xx)

(h) (xx could push this to Notes and pointers: the analyses above have presumed that θ_0 and θ_1 are somehow well-defined overall rate parameters, one for the Paroxetine users and one for the placebo group. more realistically, these suicide attempt rate parameters would vary in the population, e.g. with gender and age. argue that this could lead to negative binomial models for the final counts (Y_0, Y_1). perhaps are conclusions above too sharp. but we can't well answer this since we do not have data divided into any subcategories. xx)

Story i.13 *Brazilian children.* [xx do we have exercises on M- and Z-estimation somewhere? This story can, among other things, be a nice case study in M-estimation? xx] As part of a sanitation program in the metropolitan area of Salvador, Brazil, the Institute of Public Health at the Federal University of Bahia conducted several studies and data gathering efforts. One of these consisted of surveying the extent to which infants in the Salvador area suffered from episodes of diarrhoea. Data collectors were assigned to households and conducted home visits over a period of 455 days from October 2000 to January 2002. One child aged under 3 years at entry was monitored from each household. Being monitored, here means that a data collector went on a home visit and checked whether the child had diarrhoea or not that very day. In other words, for days at which no data collector went to the home of the child, the zero-one diarrhoea indicator is censored. In the data set, there are periods during which data collectors go on vacation, during which they are themselves home with sickness, and also periods during which they are on strike. For these and surely other more or less mundane reasons, there are longer periods for which they some children are not observed at all. In this story, we seek to model the 955 sequences of zero-one data, model the influence of covariates on a child's tendency to fall ill with diarrhoea, all in a manner that accounts for the censoring described above.

(a) The plot in the left panel of Figure i.19 shows the data for one-tenth of the children in the study (the data are plotted for one tenth of the children due to legibility). Each horizontal sequence of dots show the data for one child: Black dots indicate diarrhoea, grey dots indicate that the child was in good health, and the white dots indicate censoring (that is, we can think of the white dot as hiding a black or a grey dot from view). Read in the data and reproduce one version of this plot.

Probit model

(b) The processes-to-models idea worked with in Stoltenberg and Hjort (2021) is a natural extension of the probit regression model. Recall that in the probit model independent

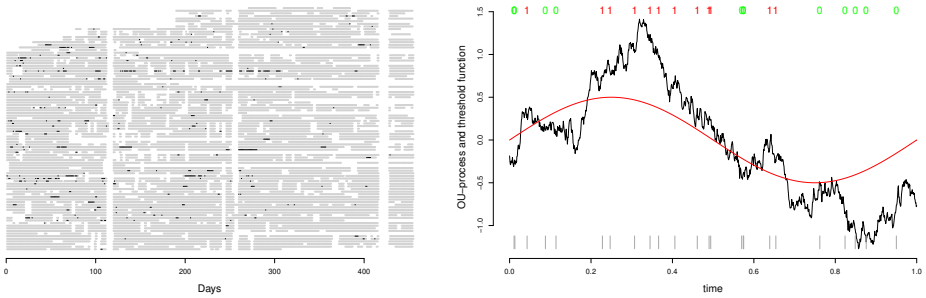


Figure i.19: *Left panel: observation pattern and actual observations for diarrhoea data (for every 10th child in the sample of 925). The grey dots indicate that the child was healthy at the observation time, and the black dots indicate that the child was sick. The white areas are time points at which no observations were made. Right-panel: a sample path of a stationary Ornstein–Uhlenbeck process with the values of $Y_{i,0}, \ldots, Y_{i,k_i}$ superimposed. The red sine-curve is the time-varying threshold. The grey ticks on the x-axis are the times at which observations were made*

zero-one data $Y_1, \ldots, Y_n$ are seen as generated by latent random variables $\eta_i = z_i^{\mathrm{t}}\gamma + \xi_i$ being above or below zero, that is $Y_i = I\{\eta_i \geq 0\}$ for $j = 1, \ldots, n$, where z_i is a vector of covariates, γ is a vector of unknown cofficients, and ξ_i are independent standard normal random variables. Show that the likelihood function for γ based on observing $(z_1, Y_1), \ldots, (z_n, Y_n)$ is $L_n(\gamma) = \prod_{i=1}^n \Phi(z_i^{\mathrm{t}}\gamma)^{Y_i}\{1 - \Phi(z_i^{\mathrm{t}}\gamma)\}^{1-Y_i}$. [xx and something more? xx]

(c) With n children observed over time, we not only have cross-section of data, as in (b), but n zero-one sequences $(Y_{i,0}, \ldots, Y_{i,k_i})_{1 \leq i \leq n}$ of various lengths k_i, and since the health of a child today is probably the best predictor of the childs health tomorrow, we ought to take this time dependence into account. Suppose that $\xi_1(t), \ldots, \xi_n(t)$ are independent Gaussian processes with covariance function $\mathrm{cov}(\xi_i(t), \xi_i(s)) = \exp(-a|t-s|)$ for some parameter $a > 0$. Now $\eta_i(t) = z_i(t)^{\mathrm{t}} + \xi_i(t)$ and the zero-one processes $Y_i(t) = I\{\eta_i(t) \geq 0\}$ are only observed at the possibly random and child specific times $t_{i,0}, t_{i,1}, \ldots, t_{i,k_i}$, so that the observed data are $Y_{i,j} = Y_i(t_{i,j})$ and $z_i(t_{i,j})$ for $j = 1, \ldots, k_i$ and $i = 1, \ldots, n$. Simulate such data with $t \in [0,1]$, $a = 1$, $z_i(t) = z_i^* \sin(2\pi t)$ for independent standard normals $z_1^*, \ldots, z_n^*$, and $t_{i,0}, \ldots, t_{i,k_i}$ sampled from some Poisson process over $[0,1]$ (practically, sample Bernoulli random variables over the fine grid you are simulating $\xi_i(t)$ on). Make a version of the plot in the right panel of Figure i.19.

(d) Suppose that the covariance function and the parameter a are known, so that the only unknown to make inference for is γ. Let $\widehat{\gamma}_n$ be the maximiser of the 'probit likelihood' $\widetilde{L}_n(\gamma) = \prod_{i=1}^n \prod_{j=1}^{k_i} \Phi(z_i^{\mathrm{t}}(t_{i,j})\gamma)^{Y_{i,j}}\{1-\Phi(z_i^{\mathrm{t}}(t_{i,j})\gamma)\}^{1-Y_{i,j}}$, and work out the large sample theory for the $\widehat{\gamma}_n$ estimator. You may treat the covariates and the observation times as fixed, i.e., not random variables, and also suppose that the covariates are constant in time, $z_i(t_{i,j}) = z_i$. Check your findings on simulated data.

(e) Cases where both a and γ are unknown are more complicated. The sequences $Y_{i,0}, Y_{i,1}, \ldots, Y_{i,k_i}$ are *not* Markov (see Slud (1989)), which entails that the full likelihood $L_n(a, \gamma) = \prod_{i=1}^n \Pr(Y_{i,0} = y_{i,0}, \ldots, Y_{i,k_i} = y_{i,k_i})$ does not factorise. For the time

being, this is problematic because maximising $L_n(a,\gamma)$ can take forever on your computer. A fix making everything more computationally convenient is to only consider the probabilities of adjacent pairs, that is, find the maximisers of the function

$$Q_n(a,\gamma) = \prod_{i=1}^{n}\prod_{j=1}^{k_i} \Pr\{(Y_{i,j-1},Y_{i,j}) = (y_{i,j-1},y_{i,j})\}.$$

Try this out on simulated data and 'see' that it works. In view of $Q_n(a,\gamma)$ can you propose other methods that might yield more efficient estimators, while at the same time not being too burdensome on your computer? [xx pointers to quasi- or composite likelihood literature xx]

(f)

(g) (xx perhaps include something on how these data are modelled in Borgan et al. (2007) xx)

(h) In Stoltenberg and Hjort (2021) ...

(i)

Story i.14 *Onset of menarche*
Data: `polish-girls-data`
Code: `(xx to come com95* nilswork23 xx)`
(xx Data from 2.B. xx) When is the onset of menarche, what is the underlying distribution? For estimating $F(x) = \Pr(T \le x)$, where T is the precise age at which menarche starts, for a given population, it would have been statistically easiest and best if one could ask a high number of women about the precise date in question, giving, in that case, data points $t_1,\ldots,t_n$. Since this is impractical, and might also lead to uncertainties in the data, a simpler and robust approach is to rather ask for a yes-no answer to the question, has it happened or not, for women sampled across a range of ages. For $n = 25$ age groups, represented in this dataset via their midpoints x_j, one has recorded that y_j out of m_j Warszawa girls have experienced onset of menarche. We view y_j as a binomial (m_j,p_j), with $p_j = F(x_j)$, and wish to estimate F.

(a) Introduce for numerical convenience $z_j = x_j - 9.00$ as age variable. Fit first the simple logistic model, where $F_1(t) = H(\beta_0+\beta_1 z)$, with the logistic transform H; see Ex. 5.45. Construct a version of Figure i.20, left panel.

(b) Then fit the logistic models also of order 2, 3, 4:

$$\begin{aligned}
F_2(t) &= H(\beta_0+\beta_1 z+\beta_2 z^2),\\
F_3(t) &= H(\beta_0+\beta_1 z+\beta_2 z^2+\beta_3 z^3),\\
F_4(t) &= H(\beta_0+\beta_1 z+\beta_2 z^2+\beta_3 z^3+\beta_4 z^4).
\end{aligned}$$

Plot these in a diagram, alongside $\widehat{p}_j = y_j/m_j$. For each model, compute the log-likelihood maximum, and the AIC scores. Which order is best?

(c) Fit also a fifth model, a skewed logistic model, where

$$p_j = H(\beta_0+\beta_1 x_j)^\kappa = \Big\{\frac{\exp(\beta_0+\beta_1 x_j)}{1+\exp(\beta_0+\beta_1 x_j)}\Big\}^\kappa.$$

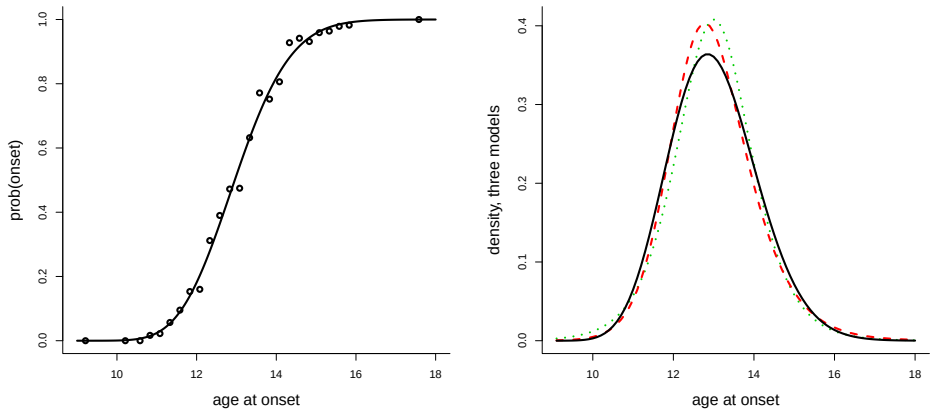


Figure i.20: *Left panel: raw data y_j/m_j, with the fitted logistic model of order 1, for the age at onset of menarche, for the Polish girls dataset. The model fits well, but there are better models. Right panel: estimated density of onset, via the Gamma process model (full curve), the three-parameter skewed logistic (dashed curve), and the simpler logistic (dotted curve).*

Again, compute the log-likelihood maximum, for which one finds $\widehat{\kappa} = 2.021$; the Wilks statistic for testing the two-parameter model against the three-parameter skewed model; and the AIC score. Explain that the evidence clearly favours the three-parameter skewed model. Also carry out CD analysis, using the Wilks based method, see Ex. 7.11, to compute and display the confidence curve $\mathrm{cc}(\kappa)$. This involves the log-likelihood profile

$$\ell_{\mathrm{prof}}(\kappa) = \max_{\text{all } \beta_0,\beta_1} \ell(\beta_0, \beta_1, \kappa).$$

You should find the asymmetric 95 percent interval $[1.202, 4.483]$, clearly to the right on $\kappa = 1$.

(d) Assume the event takes place for an individual when a stochastic nondrecreasing process $\{R(x)\colon x > 0\}$ associated with her crosses a threshold d. Show for such a setup that $F(x) = \Pr(T \le x) = \Pr(R(x) > d)$. Different models for R then lead to different and potentially new models for F. In particular, take R a Gamma process, with $R(x) \sim \mathrm{Gam}(a(x - x_0), 1)$, for $x_0 = 9.00$. Show that then $F(x) = 1 - G(d, a(x - x_0), 1)$, with the c.d.f. for the $\mathrm{Gam}(a(x - x_0), 1)$. Fit this two-parameter model, draw the estimated F and f alongside the others, as with Figure i.20, right panel. You should also check the AIC scores, finding that this Gamma process model works the best. For this best model, slightly more probability is placed on the age interval $[11, 13]$ than for the logistic models, and the peak, estimated at 12.85 years, is less sharp.

Story i.15 *Diabetic retinopathy study*
Data: `wesdr98-data`
Code: `(xx com78* of nilswork24 xx)`
In the broad study Klein et al. (2008), the aim was to examine the 25-year cumulative progression and regression of diabetic retinopathy, in the light of various risk factors. It involved following insulin-taking persons living in Wisconsin with type 1 diabetes

diagnosed before age 30. The main outcome y is an indicator for moderate to severe nonproliferate retinopathy, or proliferate retinopathy, for one or both eyes; out of $n = 691$ individuals with no missing covariates, there are 134 with $y = 1$ and 557 with $y = 0$. We have organised data into rows of $(x_1, x_2, z_1, z_2, z_3, z_4, z_5, y)$, with x_1, duration since diagnosis; x_2, indicator for presence of macular edema in one or both eyes; z_1, glycosylated hemoglobin level; z_2, body-mass index bmi; z_3, pulse rate; z_4, gender (1 for male, 0 for female); z_5, indicator for presence of urine protein. We treat the covariates x_1, x_2 as protected, i.e. they need to be included in each candidate model, whereas $z_1, \ldots, z_5$ are open, i.e. can be included or excluded in submodels, for different purposes. This leads to searching through $2^5 = 32$ logistic regression models, submodels of the wide model of the form

$$p(x, z) = \Pr(Y = 1 \,|\, x, z) = H(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4 + \gamma_5 z_5),$$

with $H(u) = \exp(u)/\{1 + \exp(u)\}$ the logistic transform. An earlier analysis of these data has been given in Claeskens and Hjort (2008a), but the treatment below is partly different.

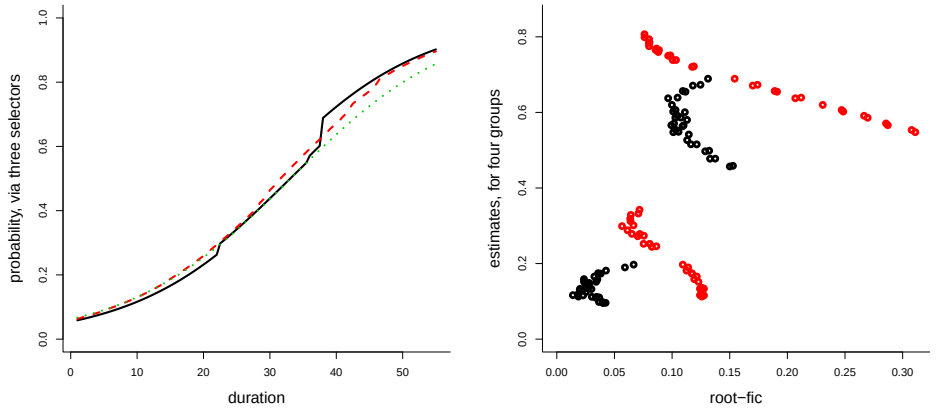


Figure i.21: *Left panel: estimating the probabiity of developing retinopathy, as a function of time since diabetes diagnosis, for individuals at quantile levels 0.50, 0.90, 0.90 for z_1, z_2, z_3, with no edema and no urine condition; the FIC winner (black full curve), the AIC winner (red slanted), the average of best ten FIC (dotted). Right panel: FIC plots for four strata of individuals, corresponding to having x_1, z_1 set equal to their median values, z_2, z_3 to their 90 percent quantile levels, $z_4 = 1$ for male, and values of (x_2, z_5) set equal to $(0, 0)$ (best estimate 0.116); $(0, 1)$ (best estimate 0.299); $(1, 0)$ (best estimate 0.637); $(1, 1)$ (best estimate 0.807). Estimates are on the vertical axis, with root-fic on the horizontal, i.e. estimated root-mse. The more to the left in the plot, the better the submodel and its estimate. There are different best models for the different strata.*

(a) First carry out logistic regression in the wide model, with all $3 + 5 = 8$ parameters, and comment on what you find. In particular, covariate z_5 is seen as strong, increasing the $Y = 1$ probability when present, whereas covariate z_3 appears to be significant, with higher pulse associated with $Y = 1$. In the FICology terminology of Ch. 11, see

Ex. (11.24) and related exercises, compute the 8×8 normalised Fisher information matrix $\widehat{J}$, and the crucial 5×5 matrix $\widehat{Q}$.

(b) Suppose we wish to estimate $\phi = p(x_0, z_0)$ accurately, for a given person, with covariates $x_0 = (1, x_{0,1}, x_{0,2})^{\rm t}$ and $z_0 = (z_{0,1}, \ldots, z_{0,5})^{\rm t}$. With partial derivatives of $p(x_0, z_0)$, with respect to β and γ, show that

$$\omega = J_{10}J_{00}^{-1}\frac{\partial\phi}{\partial\theta} - \frac{\partial\phi}{\partial\gamma} = h(x_0^{\rm t}\beta + z_0^{\rm t}\gamma)(J_{10}J_{00}^{-1}x_0 - z_0),$$

with $h(u) = \exp(u)/\{1+\exp(u)\}^2$ the derivative of H; explain how these can be estimated from data, needed for the FIC calculus.

(c) Set up clear formulae for fic_S, as in Ex. 11.24 and related exercises, involving also as many as 32 5×5 matrices $\widehat{G}_S$, for the subsets S of $\{1, \ldots, 5\}$.

(d) Construct a version of Figure i.21, left panel, as follows. For this illustration, we focus on individuals at quantile level 0.50, 0.90, 0.90 for z_1, z_2, z_3, with no edema and no urine condition. For each duration x_1 there are $2^5 = 32$ different estimates of the logistic probability of developing retinopathy before that time. The red slanted curve is for the AIC best model, which includes z_3, z_5 but not z_1, z_2, z_4. The full curve is for the FIC best model, which after about 40 years includes z_1, z_2, z_3 but not z_4, z_5. The third dotted curve has for each x_1 taken the average estimate of the ten best FIC models. In this particular case the differences are small, but in other applications there are bigger discrepancies between best FIC and best AIC, and then better chances of letting FIC find out when which covariates are more influential than others.

(e) Attempt to construct a version of Figure i.21, right panel, as follows. Motivated by attempting to understand which factors are important for assessing the risk of diabetic retinopathy, for those with relatively high values of bmi and pulse rate, pick individuals with median value for x_1, z_1 but at 90 percent quantile levels for z_2, z_3, with $z_4 = 1$ (i.e. male), and then the four possibilities for x_2 and z_5 (presence or not of endema; presence or not of urine protein). Carry out FIC analysis for these four positions in covariate space. In addition to finding the best estimate, i.e. the FIC winner, go for each case through the ten best models, and check which of $z_1, \ldots, z_5$ are are included at least 50 percent of the time. Check that this leads to covariates only 1, for $(0, 0)$; 1, 2, 3 5, for $(0, 1)$; only 1, for $(1, 0)$; 2, 3, 5 for $(1, 1)$. The FIC point is that there are different best models in different regions of covariate space.

(f) (xx male vs. female, check if same models are selected via AFIC. xx)

Story i.16 *Cigarettes and cancer*
Data: `cigarette-data`
Code: `(xx to come com151* com152* com155* nilswork24 xx)`
Certain datasets played an important role in US health politics in the 1960ies, pertaining to various cancer rates and how these related to cigarette consumption. For this story, we use (x, y), for 44 US states (actually, 43 states plus the District of Columbia), with x the cigarette consumption, in hundreds of cigarettes smoked per capita, and y, the deaths per 100K population from lung cancer. We also briefly include a point below for the corresponding dataset where y is the deaths per 100K for kidney related cancer.

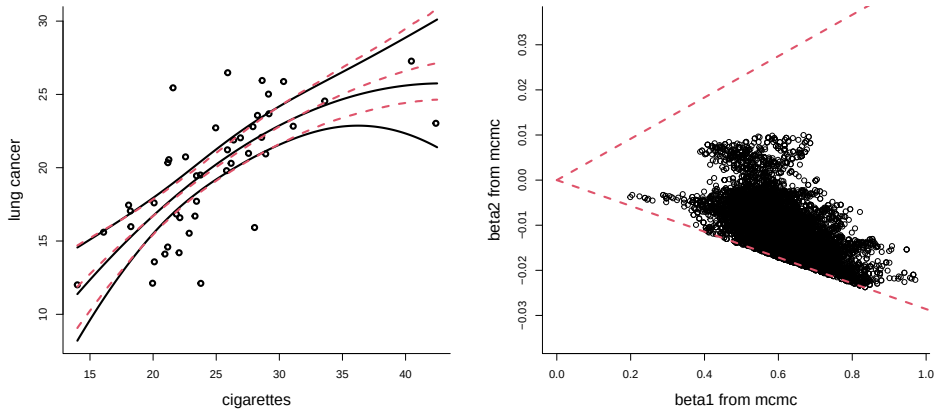


Figure i.22: *Left panel:* (x, y), *with estimates and 95 percent band via ordinary frequentist quadratic regression analysis (full curves), and via Bayesian analysis using the prior respecting monotonicity (dashed lines). Right panel: MCMC realisations from the posterior distribution for* (β_1, β_2).

(a) Work through first regression models of order 1, 2, 3. These are the linear, quadratic, cubic regressions with

$$y_i = \begin{cases} \alpha_0 + \alpha_1(x_i - \bar{x}) + \varepsilon_{i,1}, \\ \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2 + \varepsilon_{i,2}, \\ \gamma_0 + \gamma_1(x_i - \bar{x}) + \gamma_2(x_i - \bar{x})^2 + \gamma_3(x_i - \bar{x})^3 + \varepsilon_{i,3}, \end{cases}$$

with pedantically careful notation for the error terms $\varepsilon_{i,1}, \varepsilon_{i,2}, \varepsilon_{i,3}$, which are taken i.i.d. zero-mean normal, with standard deviations $\sigma_1, \sigma_2, \sigma_3$ for the three models. For the three models, compute estimates and their precision, with emphasis on the sign and size of the derivative of the regression curve, i.e. α_1 for the linear model, $\beta_1 + 2\beta_2(x - \bar{x})$ for the quadratic, and $\gamma_1 + 2\gamma_2(x - \bar{x}) + 3\gamma_3(x - \bar{x})^2$ for the cubic.

(b) Comparing the linear with the quadratic, use the tolerance radius machinery of Ex. 11.27 to judge whether the size of β_2 is small enough for the linear to be better than the quadratic model. Explain that this is a matter of whether the parameter $\gamma_2 = \sqrt{n}\beta_2/(E_n^{1/2}\sigma)$ is inside $[-1, 1]$ or not, where E_n is the $(3, 3)$ element in M_n^{-1}, for M_n the variance matrix of $(1, x_i - \bar{x}, (x_i - \bar{x})^2)$. Estimate γ_2 and comment.

(c) Compute then AIC scores for the three models; conclude that the quadratic and cubic models are judged better than the linear one. For the quadratic model, compute and display the estimated regression curve $\widehat{\beta}_0 + \widehat{\beta}_1(x_i - \bar{x}) + \widehat{\beta}_2(x_i - \bar{x})^2$, along with a 95 percent confidence band, as with Figure i.22, left panel.

(d) Explain that it is natural to require that the mean curve is monotone in x, and that this for the quadratic model means $\beta_1 + 2\beta_2(x - \bar{x}) \geq 0$ for the range of x. Show that this translates to $\beta_1 + 2\beta_2 u \geq 0$ for $u \in [-d_1, d_2]$, with $d_1 = 10.91$ and $d_2 = 17.48$. Explain that this means $\beta_1 \geq 0$ and that β_2 needs to be inside the infinite triangle defined by $-\beta_1/(2d_2) \leq \beta_2 \leq \beta_1/(2d_1)$. A Bayesian analysis ought to take this into account. Use

the prior $\pi(\beta_0, \beta_1, \beta_2, \sigma)$ that takes β_0 flat from 10.0 and above, then (β_1, β_2) uniform inside the triangle, and σ independent with prior $1/\sigma$. For this prior, set up a Metropolis MCMC generating samples $(\beta_0, \beta_1, \beta_2, \sigma)$ from the posterior, and construct from these a version of Figure i.22, right panel, with realisations respecting the boundaries. From the consequent samples of $\beta_0 + \beta_1(x - \bar{x}) + \beta_2(x - \bar{x})^2$, compute the median and a 95 percent credibility band, plotted in Figure i.22, left panel, alongside the frequentist bands. Comment on their differences.

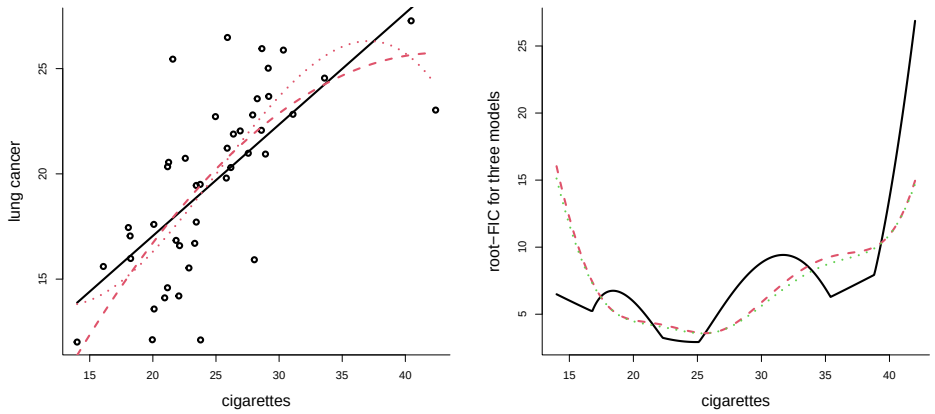


Figure i.23: *Left panel: the points* (x, y)*, with three regression curves, for models 1, 2, 3 (linear, quadratic, cubic). Right panel: root-FIC scores for models 1, 2, 3 for estimating* $\mathrm{E}(Y \mid x)$ *as a function of* x*, with smaller values indicating the best models. The full line is for the linear, the other two are very close. For each* x *this tells us with of the three estimates in the left panel is best.*

(e) Compute the three regression curves for estimating $m(x) = \mathrm{E}(Y \mid x)$, from models 1, 2, 3 (the linear, quadratic, cubic), as in Figure i.23, left panel. For a given state, or for that matter the cigarette consumption level x of any sampled population, it is not clear which of the three models is best. Set therefore up a FIC scheme, via the methods and details of Ex. 11.28. The task is to compute for each x, for the range $[14.0, 42.4]$, not only the regression curve estimates, but the root-FIC scores, via the machinery of estimated variances and biases. This in particular involves the 4×4 variance matrix Σ_n of $(1, x - \bar{x}, (x - \bar{x})^2, (x - \bar{x})^3)$, along with 2×2 matrices $G_{n,1} = 0$, $G_{n,2}$, $G_{n,3} = I_2$ for models 1, 2, 3, and the crucial $\omega(x) = \Sigma_{n,10}\Sigma_{n,00}^{-1}(1, x)^{\mathrm{t}} - ((x - \bar{x})^2, (x - \bar{x})^3)^{\mathrm{t}}$. Construct in this fashion a version of Figure i.23, right panel. We see in particular that the linear curve is not the best predictor for lung cancer volume, when it comes to populations with heavy cigarette consumption.

(f) (xx one more point, now for kidney cancer; to be polished. xx) Fit linear and quadratic regressions to the (x, y) data, with y the rate for kidney cancers, and construct a version of Figure i.24, left panel. Compute AIC values to see that the data clearly prefers to be modelled by quadratic regression, compared to linear regression. Then form the monitoring plots $Z_n(t) = \sqrt{n}\{G_n(t) - \Phi(t)\}$, as per Ex. 9.28, with G_n the empirical

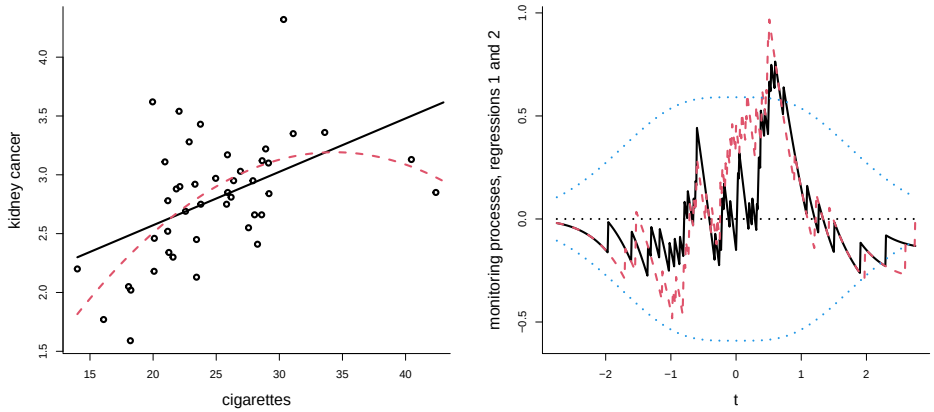


Figure i.24: *Left panel:* (x, y) *for kidney cancer rates, with the linear and quadratic regression fits. Right panel: monitoring plots* $\sqrt{n}\{G_n(t) - \Phi(t)\}$*, for the empirical c.d.f.s for the normalised residuals from linear and quadratic regression, along with a pointwise 95 percent band.*

c.d.f. for the normalised residuals. These are $r_i/\widehat{\sigma}$ with $r_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1(x_i - \bar{x}))$ for the linear case, with a similar expression for the quadratic case. Construct a version of Figure i.24, right panel. That plot includes the curves $\pm 1.96\,\tau(t)$, where

$$\tau(t) = [\Phi(t)\{1 - \Phi(t)\} - \phi(t)^2]^{1/2}.$$

Explain from the exercise pointed to that $\Pr\{Z_n(t) \in \pm 1.96\,\tau(t)\} \to 0.95$, so this is pointwise 95 percent confidence band, provided the model holds. Simulate a high number of Z_n curves, from the estimated regression models, read off their maxima, to compute p-values for the tests based on $M_n = \max_{\text{all } t} Z_n(t)$. Conclude from this that the linear model is barely acceptable, and that the quadratic model is not, with p-value around 0.05 and 0.01. Is this a contradiction, clashing with AIC's choice?

Story i.17 *Life expectancy.* [xx [xx might delete this story xx]Could base story on Borgan and Keilman (2019), but there are many other things here as well. Nice intro in Norwegian is Gran and Stensrud (2022). And also cite the methods protocol Wilmoth et al. (2021). Code for this story is in `lifeexpectancy_emil1.R` xx]. Life expectancy is a common measure of the quality of life in a given country. Countries whose inhabitants are expected to live longer are generally viewed as more fortunate countries to be born in. As is implied by its name, life expectancy is an estimate of the expected lifetime of someone born in a given country in a given year. The challenge in estimating such an expectation is what data to use. There are two main methods: (i) We might wait until all people born in a given year have passed away, then compute the average length of their lifes. This is called the *cohort* method. (ii) The *period* method assumes that the children born in 2023 say, will throughout their lives be exposed to the mortality rates currently observed. Statistically speaking, this means that the lifetimes of the people alive in 2023 are assumed to be sampled from the same distribution.

(a) Let $T \in \{0, 1, \ldots, \tau\}$ be a random variable with hazard $\Pr(T = j \,|\, T \geq j) = \alpha_j$ for $j = 0, 1, \ldots, \tau$. Let $(T_{x,1}, \ldots, T_{x,n})_{0 \leq x \leq \tau}$ be i.i.d. with the same distribution as T. Sup-

pose that n is known and that we are given the data $\Delta N_{\tau-j,j} = \sum_{i=1}^{n} I\{T_{\tau-j,j} = j\}$ and $Y_{\tau-j,j} = \sum_{i=1}^{n} I\{T_{\tau-j,j} \geq j\}$ for $j = 0, 1, \ldots, \tau$. The period method problem is akin to estimating $\mathrm{E}\, T$ based on data of this sort. Since the estimator $\{(\tau+1)n\}^{-1} \sum_{t=0}^{\tau} \sum_{i=1}^{n} T_{t,i}$ is not an option, the trick is to write (show it!)

$$\mathrm{E}\, T = \sum_{\ell=0}^{\tau} \prod_{k=1}^{\ell-1} (1 - \alpha_k),$$

and then estimate the hazard rates $\alpha_0, \alpha_1, \ldots, \alpha_{\tau-1}$. Show that for $j = 0, 1, \ldots, \tau - 1$,

$$\Delta N_{\tau-j,j} \,|\, (Y_{\tau-j,j} = y_{\tau-j,j}) \sim \mathrm{binom}\big(y_{\tau-j,j}, \alpha_j\big),$$

and set $\ell_{\mathrm{cond}}(\alpha_0, \ldots, \alpha_{\tau-1}) = \sum_{j=0}^{\tau} \{x_{\tau-j,j} \log(\alpha_j) + (y_{\tau-j,j} - x_{\tau-j,j}) \log(1 - \alpha_j\}$. Show also that the full likelihood of these data is $\ell_{\mathrm{full}} = \log L_{\mathrm{full}}$ where $\log L_{\mathrm{full}}$ is (proportional to)

$$\begin{aligned} L_{\mathrm{full}}(\alpha_0, \ldots, \alpha_{\tau-1}) &= \{\alpha_0^{\Delta N_{\tau,0}} (1 - \alpha_0)^{n - \Delta N_{\tau,0}}\} \{f_\tau^{Y_{0,\tau}} (1 - f_\tau)^{n - Y_{0,\tau}}\} \\ &\times \{\prod_{j=1}^{\tau-1} \alpha_j^{\Delta N_{\tau-j,j}} (1 - \alpha_j)^{Y_{\tau-j,j} - \Delta N_{\tau-j,j}} S_j^{Y_{\tau-j,j}} (1 - S_j)^{n - Y_{\tau-j,j}}\}, \end{aligned}$$

where, we recall that $S_j = \prod_{k=0}^{j-1} (1 - \alpha_k)$ and $f_j = S_j \alpha_j$. Let $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ and $(\widetilde{\alpha}_0, \widetilde{\alpha}_1)$ be the maximisers of $\ell_{\mathrm{cond}}(\alpha_0, \alpha_1)$ and $\ell_{\mathrm{full}}(\alpha_0, \alpha_1)$, respectively. By way of simulation, compare the means squared errors of the estimators $\widehat{\mu} = (1 - \widehat{\alpha}_0) + (1 - \widehat{\alpha}_0)(1 - \widehat{\alpha}_1)$ and $\widetilde{\mu} = (1 - \widetilde{\alpha}_0) + (1 - \widetilde{\alpha}_0)(1 - \widetilde{\alpha}_1)$. Approximately how much is lost by disregarding the information about α_0 and α_1 contained the at-risk counters?

(b) The standard estimator for $\mathrm{E}\, T$ for a given country in a given year t, disregarding some demographic technicalities (see Wilmoth et al. (2021)), is $\widehat{\mu}^{(t)} = \sum_{\ell=1}^{110} \prod_{j=0}^{\ell-1} (1 - \widehat{\alpha}_j^{(t)})$ where $\widehat{\alpha}_j^{(t)} = \Delta N_{t-j,j} / Y_{t-j,j}$ for $j = 0, 1, \ldots, 110$. One reason for using this estimator and not the one based on the maximisers of the full likelihood of $(\Delta N_{t-j,j}, Y_{t-j,j})_{0 \leq j \leq 110}$, is that the number of people born in year y might, due to immigration, be smaller than the number of people aged a at risk in year $y + a$ (in the expression for ℓ_{full} in (a) this would amount to negative $n - y_{2-j,j}$ terms). Read in the deaths and exposure data sets for Italy, Japan, Sweden, and Norway (called `deathsITA.txt`, `exposureITA.txt`, and so on), and compute $\widehat{\mu}^{(t)}$ for $t = 1947, \ldots, 2019$, or the latest available year, for females in the four countries. Reproduce a version of the plot in Figure i.25.

(c) Among the many data sets provided by the Human Mortality Database, we use the Deaths and the Exposure-to-risk data sets. For example, the two data sets for Japan for the year 2020 have the following form:

(d) The question posed in Borgan and Keilman (2019) is what story the plot in Figure i.25 tells us: Is it correct that women born in Italy and Japan may now expect to live longer than women born in Sweden and Norway, or are the differences seen in the plot an artefact of the estimation method?

(e)

(f)

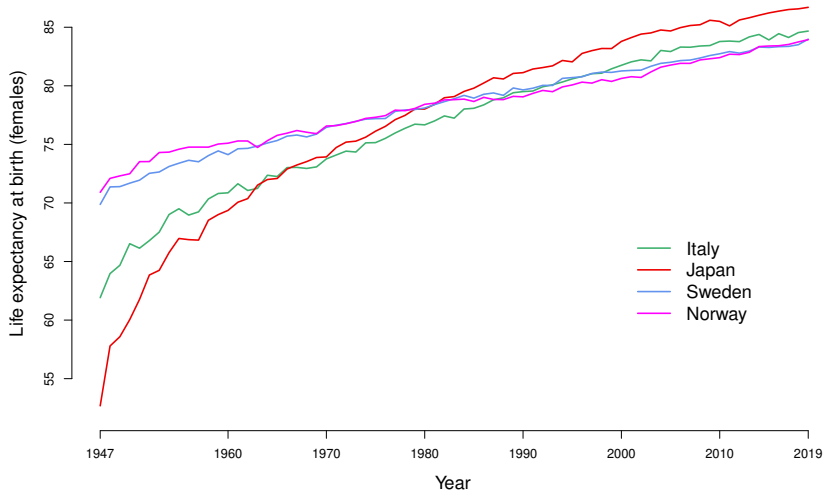


Figure i.25: Estimated life expectancy at birth for females in Italy, Japan, Sweden, and Norway for the period from 1947 to 2019, computed using the period method. Retrieved from The Human Mortality Database `www.mortality.org`

Story i.18 *A cure model.* Some parents are happy with two children. From the plot in Figure i.14 we see that about fifty percent of all mothers of two do not have a third child: the survival curves flattens out. This might be because they make a conscious effort not to have a third child, or because they do not manage to produce one due to age etc. A statistical way of modelling this is to suppose that at the time of the birth of the second child, the mother draws a Bernoulli random variable U. If $U = 1$ she will have a third child, while if $U = 0$ she will not, with unkown $\Pr(U = 1) = p$ unknown. If $U = 1$, then the survival function $S(t) = \Pr(\text{still no third child at time } t)$ will eventually reach zero, while if the mother belong to the no-third-child group $U = 0$, then her 'survival function' is equal to 1 for all time. This gives a population survival function

$$S_{\text{pop.}}(t) = 1 - p + pS(t).$$

Here, $S_{\text{pop.}}(t)$ is a degenerate survival function because it tends to $1 - p > 0$ as $t \to 0$, while $S(t)$ is a bona fide survival function. In general, both p and $S(t)$ may depend on covariates. In this story, we will study a model where p does not depend on covariates, but $S(t)$ does. In particular, the ith mother has survival function $S_{\text{pop.},i}(t) = 1 - p + pS_i(t)$, where $S_i(t) = S_0(t)^{\exp(x_i\beta)}$ for $x_i = I\{\text{two girls or two boys}\}$, as in Story i.10, where $S_0(t) = \exp(-\int_0^t \alpha_0(s)\,\mathrm{d})$ is some baseline survival function, and $\alpha_0(t)$ a baseline hazard. [xx say that we are in continuous time, and introduce N, Y, M xx]

(a) Give the baseline survival function your favourite parametric specification, for example the Weibull $S_0(t) = S_0(t \,|\, a, b) = \exp(-bt^a)$. Show that the population hazard function (i.e., when U is averaged out) is

$$\alpha_{\text{pop.}}(t) = \frac{p\alpha_0(t)S(t)}{1 - p + pS(t)},$$

Estimate the parameters p, β, a, and b, and plot the estimated survival functions $S_0(t\,|\,\widehat{a},\widehat{b})$ and $S_0(t\,|\,\widehat{a},\widehat{b})^{\exp(\widehat{\beta})}$, with the hats indicating the maximum likelihood estimators. Add pointwise 95 percent confidence bands to your plot. See Figure xx.

(b) If the baseline hazard is *not* given a parametric specification, we are no longer in a situation where it can be disregarded in the estimation of the other parameters (that is, using Cox's partial likelihood, see Ex. [xx ch. 10 somewhere xx]).

(c)

Notes and pointers

(xx notes and follow-up things for the stories in this chapter. xx)

for children: see Hjort blog post Hjort (2018a). (xx round off. studying things for sibling flock sizes $m = 2, \ldots, 12$. is the ρ parameter different for smaller families (say $m \leq 5$) compared to bigger families (say $m \geq 10$)? point to blogpost. xx)

(xx iud: mention Peterson (1975) and find Aalen spiralen 1972. data are from the 1970ies. data quality not perfect, since some expulsions are detected around one-year controls. xx)

(xx Re N_ε: point to Hjort and Fenstad (1992); Grønneberg and Hjort (2012). xx)

(xx mention illustrations of Gamma process models, for a couple of the stories, stem from applications presented in Cunen and Hjort (2024). polish girls, old egypt, time to 2nd child. xx)

(xx re Laptook: point to Hjort blog story Hjort (2017a), big JAMA paper Laptook (2017), short critical follow-up papers Walløe et al. (2019a,b). xx)

II.ii

Art, History, Literature, Music

(xx WELL: lots of things to fix, as of 16-January-2025. todo list for nils includes: (i) clean the GoT and WoR stories; some should be scrapped; probably enough with one, not two. revised game plan as of august 2023: nelson-aalen and kaplan-meier; do gompertz, but as benchmark; briefly estimate parameters via quantiles, before turning to ML; throw in one more para model; ask about a focus parameter. (ii) make the Platon story tighter, and round off. point to Markov chapter things. (iii) get the Tirant lo Blanc story started. use changepoint process things from Ch9. xx)

Story ii.1 *Game of Thrones and the Wars of the Roses*
Data: `got-data`
Code: `(xx to come, com1* of nilswork23 xx)`
(xx we have one not two GoT-WoR stories. the present version as of 16-January-2025 will be significantly shortened. need intro sentences, also about the sampling and the populations; we're not sampling the general populations in the usual sense, hence need for care for interpretation. data: (t, δ, x_1, x_2) for the two populations, with x_1 an indicator for belonging to the nobility or not, and x_2 an indicator for gender, $x_2 = 1$ for women and $x_2 = 0$ for men. there are $n_g = 328$ and $n_w = 407$ individuals in the two datasets. For the WoR data there is no censoring, so all $\delta_i = 1$ there, whereas there is about 54 percent censoring for the GoT data. xx) (xx need to calibrate with French presidents. avoid repeated similar use of gompertz, so make sure for this story that we use another interesting model too. xx) In the datasets the lifetimes are essentially recorded as natural numbers, 1.0, 2.0, 3.0, 6.0, 7.0, etc. When computing and displaying the nonparametric cumulative hazard rates and survival curves below we choose for convenience to jitter these, i.e. adding a small level of random noise, to avoid artificial consequences associated with these ties.

(a) For the two datasets, so far disregarding information concerning gender and nobility, compute and display nonparametric Nelson–Aalen cumulative hazard curves $\widehat{A}_{\rm got}$ and $\widehat{A}_{\rm wor}$ and Kaplan–Meier survival curves $\widehat{S}_{\rm got}(t)$ and $\widehat{S}_{\rm wor}$, as in Figure ii.1. Define the empirical q quantile as the first t for which the estimated c.d.f. $\widehat{F}(t) = 1 - \widehat{S}(t)$ is equal to or above q. Show that the 0.25 and 0.75 quantiles are 34.9 and 66.8 for GoT, and 46.1 and 70.0 for WoR.

(b) Consider the two-parameter Gompertz model for survival data, with hazard rate $\alpha(t) = \exp(\gamma_0 + \gamma_1 t)$; see Ex. 1.59. Show that the cumulative hazard can be written

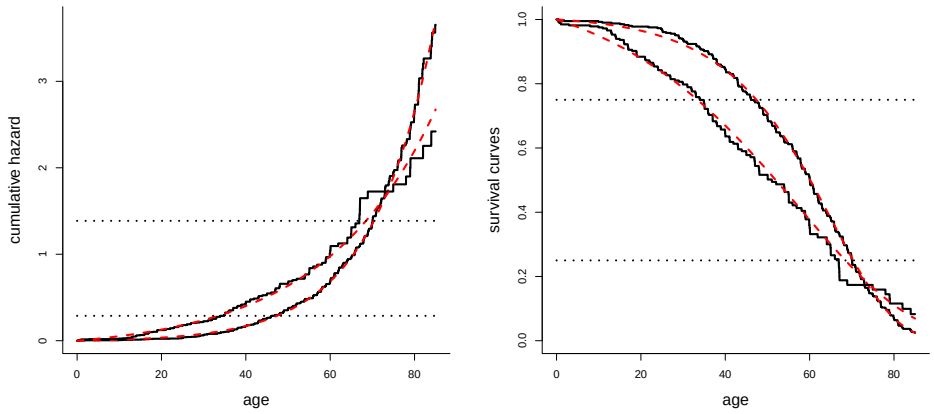


Figure ii.1: *Cumulative hazard rates (left panel) and survival curves (right panel), nonparametric and fitted via the two-parameter Gompertz model, for the GoT and WoR survival data. WoR lives tend to be longer than GoT lives up to about age 75. Horizontal lines indicate 0.25 and 0.75 quantiles. The 0.25 and 0.75 lifetime quantiles are 34.9 and 66.8 for GoT, and 46.1 and 70.0 for WoR.*

$A(t) = \exp(\gamma_0)\{\exp(\gamma_1 t) - 1\}/\gamma_1$, and give a formula for the survival function $S(t)$ and for the quantile $(1 - S)^{-1}(q)$. Before we come to ML estimation, fit such Gompertz models to the GoT and WoR datasets, by equating the 0.25 and 0.75 quantiles (this means solving two equations with two unknowns, for GoT and for WoR). In this fashion, complete a full version of Figure ii.1. Sum up what your comparison between the two worlds tells us, so far.

(c) Turning now to maximum likelihood estimation, show that with the Gompertz model for the hazard rates, the log-likelihood function can be represented as $\sum_{i=1}^n \{(\gamma_0+\gamma_1 t_i)\delta_i - \int_0^{t_i} \exp(\gamma_0 + \gamma_1 s)\,\mathrm{d}s\}$. Make this expression more explicit, and find ML estimates, for GoT and for WoR. Construct a completed version of Figure ii.1, with both nonparametric and parametrically fitted curves.

(d) Then use a regression version of the Gompertz model, for the two populations, with hazard rates

$$\alpha_i(s) = \exp(\gamma_0 + \gamma_1 s)\exp(x_{i,1}\beta_1 + x_{i,2}\beta_2) \quad \text{for } i = 1, \ldots, n.$$

Show that the log-likelihood becomes

$$\ell_n = \sum_{i=1}^n \Big\{(\gamma_0 + \gamma_1 t_i + x_{i,1}\beta_1 + x_{i,2}\beta_2)\delta_i - \exp(\gamma_0)\frac{\exp(\gamma_1 t_i) - 1}{\gamma_1}\exp(x_{i,1}\beta_1 + x_{i,2}\beta_2)\Big\}.$$

Fit this four-parameter model to the two datasets, finding ML estimates, using the pre-covariate ML estimates as starting values for the optimisation. Use the ML theory (xx find right thing in Ch 10 xx) to find confidence intervals for the crucial parameters β_1, β_2, for GoT and for WoR. Explain how this indicates (i) that there is no essential difference between the survival of men and women, for the two populations; (ii) that for WoR, it

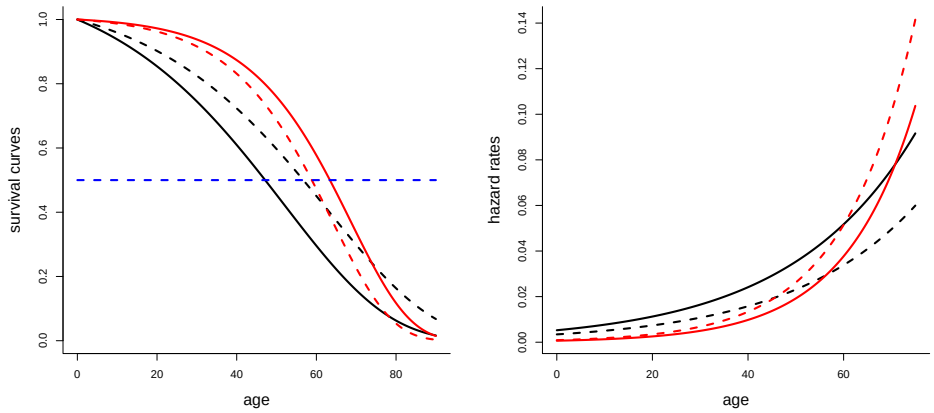


Figure ii.2: *(xx check with com2* of kioskvelterwork22*. black full, black dashed: got0, got1. red full, red dashed: wor0, wor1. with x2=0, i.e. for a man. medians, from smaller to bigger, then: got0, got1, wor1, wor0. xx) Left panel: survival curves. Right panel: hazard rates.*

was worse to belong to nobility than not (β_1 is significantly positive); (iii) that for GoT, it was better to belong to the nobility than not (β_1 is significantly negative).

(e) (xx delta method for a couple of interesting foci. also do AIC for different parametric models, with weibull and gamma in lieu of gompertz. xx) Show that median lifetime t^*, for an individual with covariates x_1 and x_2, is determined by

$$\exp(\gamma_0)\{\exp(\gamma_1 t^*) - 1\}/\gamma_1 \, \exp(x_1\beta_1 + x_2\beta_2) = \log 2.$$

Estimate this median time, for a man belonging to the nobility, and for another man outside nobility, for the GoT and for the WoR worlds. Supply also 90 percent confidence intervals. (xx answers in such a table; estimates and intervals. GoT: best to be noble; WoR: best outside. see Figure ii.2. xx)

```
Got, outside          Got, inside
47.14 42.77 51.51     56.72 52.06 61.38
WoR, outside          WoR, inside
63.39 61.18 65.61     58.85 57.09 60.62
```

(f) (xx compare with classic Cox semiparametric. losing some precision. here we allow crossing hazards, not demanded proportional. point to Jullum and Hjort (2019). xx)

(g) (xx something with interesting focus and then delta method. can also ask readers to redo all with Weibull in leiu of Gompertz, perhaps with aic, etc. xx)

Story ii.2 *Stride towards your bookshelves*
Data: `wordlengths-data`
Code: `(to come com333* of nilswork24 xx)`
As part of the obligatory exercises work for a bachelor level course on statistical methodology at the Department of Mathematics, University of Oslo, we instructed each student

to stride towards her or his bookshelves, to pick one book in Norwegian and one in English, then record the lengths of the first 100 words on page 51. The books could be novels, collections of short stories, poetry, or prose in general, but not technical material (as with mathematics or statistics); the students were also instructed to use page 52 if page 51 didn't have enough words. Do Fløgstad, Kjærstad, Solstad tend to use words with more or less the same lengths as do Miller, Lessing, Munro? And do some students have books tending to have longer words than those of other students?

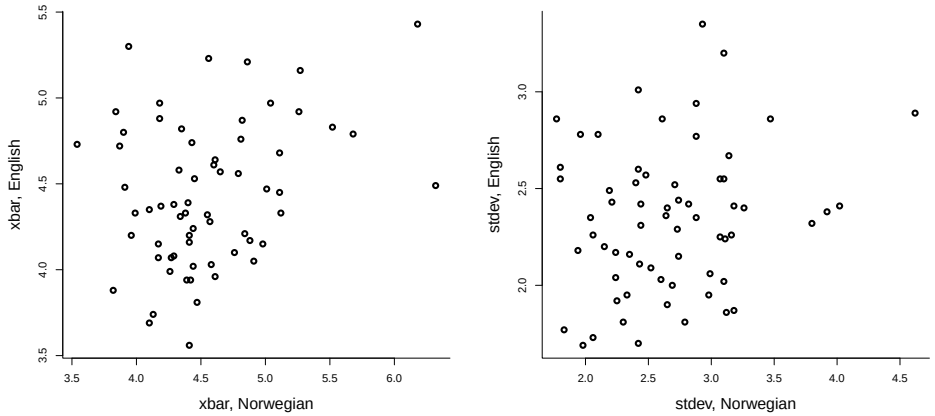


Figure ii.3: *Empirical means* $(x_{i,N}, x_{i,E})$ *(left panel), and empirical standard deviations* $(\widehat{\kappa}_{i,N}, \widehat{\kappa}_{i,E})$ *(right panel), for the Norwegian and English wordlengths found in 64 students' bookshelves, with each student sampling 100 words from their sampled books.*

The students were asked to summarise information and to compare their own two datasets in terms of means and standard deviations. This was expected to involve tests for equality of means and of variances, confidence intervals for differences, perhaps comments on skewnesses, etc. But the experiment also gave us an interesting combined data set, where we recorded the empirical mean and standard deviation for each dataset, for the two languages, for each student. In other words, we have summary statistics data $(x_{i,N}, \widehat{\kappa}_{i,N}, x_{i,E}, \widehat{\kappa}_{i,E})$ for $i = 1, \ldots, n$, for the $n = 64$ students, with

$$\begin{aligned} x_{i,N} &= \text{average word-length for 100 Norwegian words for student } i, \\ x_{i,E} &= \text{average word-length for 100 English words for student } i, \end{aligned}$$

along with empirical standard deviations $\widehat{\kappa}_{i,1}$ and $\widehat{\kappa}_{i,2}$, say, for these 100 Norwegian and 100 English words, for student i.

(a) Construct a version of Figure ii.3, one panel with $(x_{i,N}, x_{i,E})$, a second panel with $(\widehat{\kappa}_{i,N}, \widehat{\kappa}_{i,E})$. Why and in which sense was it ok for Hjort and Stoltenberg to throw away the individual data samples, with $nm = 64 \cdot 100$ words in each of the two languages, and just keep the empirical means and standard deviations?

(b) Carry out a test to see if the mean word lengths are about the same, for the Norwegian and English books (in these students' bookshelves). For this point, suppose that $X_{i,N} \sim \mathrm{N}(\xi_{i,N}, \kappa_{i,N}^2/100)$ and $X_{i,E} \sim \mathrm{N}(\xi_{i,E}, \kappa_{i,E}^2/100)$. Then perform a second test, to see if the underlying spread in wordlength distributions are the same for the two languages. [xx polish a bit. answers are no for $\bar{x}$, but yes for the $\widehat{\kappa}$. xx]

(c) We then take an in interest in *the correlation* between the two wordlength distributions. But taking the ordinary correlation between the reported averages $x_{i,N}, x_{i,E}$ is less interesting than inference for *the real correlation*, between say $x_{i,\text{real},N}, x_{i,\text{real},E}$, these being the averages over the tens of thousands of Norwegian and English words on the bookshelves of student i. It turns out the ordinary correlation deflates this underlying real correlation, due to the measurement errors involved in sampling merely 100 words for the two corpora.

In general terms, suppose we have observations (x_i, y_i) for $i = 1, \ldots, n$, where these are really proxies for certain underlying $(x_{i,0}, y_{i,0})$, and where the measurement errors involved are normal with known variance levels. We have in mind situations where the correlation $\rho = \text{corr}(x_0, y_0)$ between these underlying quantities is of higher concern than the deflated correlation $\text{corr}(x, y)$ between the directly observed (x_i, y_i). We formalise a version of the setup described as

$$x_i = x_{i,0} + \delta_{i,1}, \quad y_i = y_{i,0} + \delta_{i,2},$$

where the not fully observed $(x_{i,0}, y_{i,0})$ have a binormal distribution, with correlation ρ, and where the measurement errors $\delta_{i,1}$ and $\delta_{i,2}$ are independent zero-mean normals with known or well estimated standard deviations τ_1 and τ_2. With (ξ_1, ξ_2) the means and (σ_1, σ_2) the standard deviations for $(x_{i,0}, y_{i,0})$, show that

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \text{N}_2(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, V), \quad \text{with} \quad V = \Sigma + D = \begin{pmatrix} \sigma_1^2 + \tau_1^2, & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2, & \sigma_2^2 + \tau_2^2 \end{pmatrix},$$

writing D for $\text{diag}(\tau_1^2, \tau_2^2)$.

(d) First we sort out what happens with the traditional empirical correlation coefficient for the observed data, say $R_n = s_{1,2}/(s_1 s_2)$, where s_1 and s_2 are the empirical standard deviations for the x_i and the y_i, and $s_{1,2} = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/(s_1 s_2)$. Show that

$$R_n \to_{\text{pr}} \frac{\rho\sigma_1\sigma_2}{(\sigma_1^2 + \tau_1^2)^{1/2}(\sigma_2^2 + \tau_2^2)^{1/2}},$$

i.e. the default operation R_n actually estimates a deflated version of the real ρ.

(e) Consider first repair operation 1, which is to estimate the σ_j by $\widehat{\sigma}_j^2 = \max(s_j^2 - \tau_j^2, 0)$ for $j = 1, 2$. Show that $\widehat{\rho} = s_{1,2}/(\widehat{\sigma}_1\widehat{\sigma}_2)$ is consistent for ρ. Note however that its limit distribution is more complicated than for the classical case of no measurement error; see Ex. 2.53. (xx then spell out what this means for the bookshelves story. the point is to see τ_1 and τ_2 from the data, as precisely estimated. For the $x_{i,N}$, with individual variances $\kappa_{i,N}^2/m$, argue that $\tau_N = \{(1/n)\sum_{i=1}^n \widehat{\kappa}_{i,N}^2/m\}^{1/2} = 0.2728$ and $\tau_E = \{(1/n)\sum_{i=1}^n \widehat{\kappa}_{i,E}^2/m\}^{1/2} = 0.2370$ are precise estimates of the measurement errors here. From the directly observed standard deviations $s_N = 0.5285$ and $s_E = 0.4193$ show that these are reduced to $\widehat{\sigma}_N = (s_N^2 - \tau_N^2)^{1/2} = 0.4526$ and $\widehat{\sigma}_E = (s_E^2 - \tau_E^2)^{1/2} = 0.3459$. This adjusts the deflated $R_n = 0.2833$ to $\widehat{\rho} = 0.4010$. xx)

(f) Then consider repair operation 2, using likelihood methods. Show that the log-likelihood function for the observed data becomes

$$\ell_n = \sum_{i=1}^n \Big\{ -\tfrac{1}{2}\log|\Sigma + D| - \tfrac{1}{2}\begin{pmatrix} x_i - \xi_1 \\ y_i - \xi_2 \end{pmatrix}^{\text{t}} (\Sigma + D)^{-1} \begin{pmatrix} x_i - \xi_1 \\ y_i - \xi_2 \end{pmatrix} \Big\}.$$

Show that profiling over the means leads to $\ell_{n,\text{prof}}(\sigma_1, \sigma_2, \rho) = -\frac{1}{2} n Q(\sigma_1, \sigma_2, \rho)$, where

$$Q(\sigma_1, \sigma_2, \rho) = \log |\Sigma + D| + \text{Tr}\{(\Sigma + D)^{-1} S_n\},$$

in terms of the empirical variance matrix S_n for the (x_i, y_i) pairs. (xx do this for the bookshelves data. find and display a cc(ρ). point estimate 0.401, 95 percent interval $[0.064, 0.670]$. xx) (xx variations: could actually have different $\tau_{i,1}, \tau_{i,2}$ for the log-likelihood. xx) (xx careful with wording: We learn that a student having Norwegian books with long words tends to have English books with long words too, and vice versa. The reasons for this interesting finding are not clear, but it's interesting to do a bit of speculation – some readers prefer longer-worded books, others might like shorter-worded literature. We're also reminded that the students were not instructed to choose books from their bookshelves in a totally random fashion, so there's a limit to how far we should stretch our imagination here. xx)

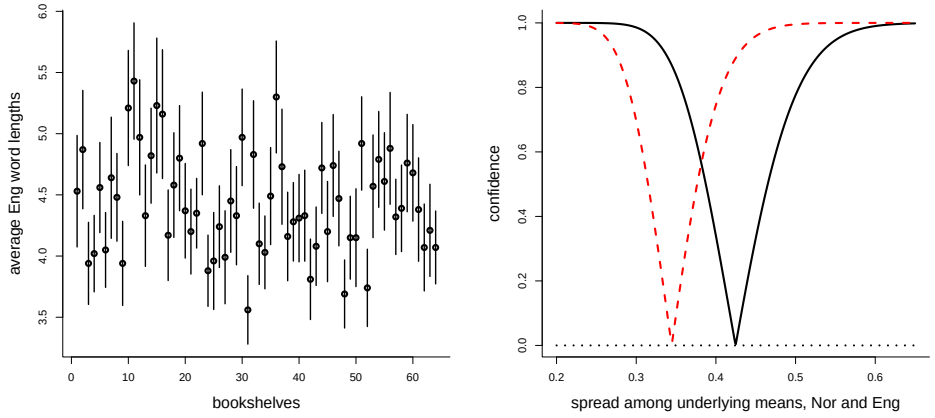


Figure ii.4: *Left panel: for the $n = 64$ students, average word lengths in their English books, with 90 percent confidence intervals. Right panel: confidence curves for the spread parameters ω_N and ω_E, for the models where the averages $x_{i,N}$ and $x_{i,E}$ follow distributions* $\text{N}(\xi_N, \omega_N^2)$ *and* $\text{N}(\xi_E, \omega_E^2)$.

(g) As is already apparent from the correlation analysis, the wordlengths exhibit not merely the obvious variation inside bookshelves, but also between students. Construct a version of Figure ii.4, left panel, displaying English wordlength averages $x_{i,E}$ along with their associated individual 90 percent intervals. To assess the degree of disparity between students, i.e. betweeen their bookshelves, model the $x_{i,E}$ as coming from a $\text{N}(\xi_E, \omega_E^2)$ distribution. Show that marginally, $x_{i,E} \sim \text{N}(\xi_E, \sigma_{i,E}^2 + \omega_E^2)$, with $\sigma_{i,E}^2 = \kappa_{i,E}^2/m$. Since these are well estimated, we take them as nearly known, set equal to $\widehat{\kappa}_{i,E}^2/m$. (xx then calibrate with what is in Ch7. xx) Using

$$Q_E(\omega_E) = \sum_{i=1}^n \frac{\{x_{i,E} - \widehat{\xi}_E(\omega_E)\}^2}{\sigma_{i,E}^2 + \omega_E^2} \sim \chi_{n-1}^2, \quad \text{with } \widehat{\xi}_E(\omega_E) = \frac{\sum_{i=1}^n x_{i,E}/(\sigma_{i,E}^2 + \omega_E^2)}{\sum_{i=1}^n 1/(\sigma_{i,E}^2 + \omega_E^2)},$$

construct the CD $C_E(\omega_E) = 1 - \Gamma_{n-1}(Q_E(\omega_E))$ and its associated confidence curve. Compute cc(ω_E) and cc(ω_N) and display these in a diagram, as with Figure ii.4. Find

median confidence estimates and also 95 percent intervals for the spread parameters ω_E and ω_N, and comment on their sizes.

Story ii.3 *Tirant lo Blanc: When did Author B take over for Author A?*
Data: `tirant-data`
Code: `(xx to come com106* com111* com61* nilswork23 xx)`

"I swear to you, my friend, this [Tirant lo Blanch] is the best book of its kind in the world", writes Cervantes

Full of adventures, battles and love stories, the chivalry romance *Tirant lo Blanch* is a masterpiece of medieval literature. The novel, written in Catalan around c. 1465 and later published in 1490, is arguably the world's first, preceding Cervantes' more famous *Don Quixote* with about 150 years. Its chief author was the Valencian nobleman Joanot Martorell (1410–1465), but his court intrigue related death happened before the novel was finished, after which his friend Martí Joan de Galba (c. 1445-1490) somehow took over, completing the manuscript. The novel hence carries its own posthumous literary mystery; when did Author B take over for Author A? We treat this here as a statistical changepoint challenge.

The book has as many as 487 chapters, varying in size. Chapter i has m_i words, of lengths $w_{i,1}, \ldots, w_{i,m_i}$. From these we compute the relative proportions $\widehat{p}_{i,1}, \ldots, \widehat{p}_{1,10}$ of words of lengths $1, \ldots, 10$, with '10' meaning '10 or longer'. These relative proportions constitute our data, from which we then attempt to estimate a changepoint. (A bureaucratic footnote here is that our dataset actually uses these relative proportions only for a subset of $n = 425$ of the 487 chapters, the remaining chapters judged being too small.)

	1	2	3	4	5	6	7	8	9	10
1	0.082	0.231	0.173	0.075	0.129	0.078	0.063	0.067	0.035	0.067
2	0.111	0.237	0.168	0.103	0.109	0.069	0.059	0.076	0.034	0.034
3	0.093	0.233	0.204	0.109	0.095	0.094	0.065	0.043	0.037	0.027
4	0.103	0.224	0.188	0.106	0.090	0.106	0.070	0.048	0.034	0.031
5	0.109	0.190	0.212	0.113	0.118	0.094	0.056	0.051	0.027	0.031
6	0.112	0.221	0.205	0.112	0.098	0.099	0.060	0.044	0.024	0.024
7	0.113	0.223	0.180	0.064	0.102	0.099	0.053	0.053	0.067	0.046
8	0.110	0.219	0.173	0.080	0.114	0.122	0.046	0.068	0.021	0.046
9	0.103	0.188	0.215	0.072	0.067	0.126	0.054	0.067	0.063	0.045
10	0.106	0.220	0.219	0.107	0.097	0.083	0.054	0.054	0.031	0.028
.....										
485	0.130	0.235	0.143	0.077	0.104	0.116	0.057	0.073	0.032	0.034
486	0.112	0.219	0.226	0.114	0.100	0.070	0.032	0.075	0.027	0.025
487	0.138	0.141	0.178	0.152	0.118	0.103	0.060	0.026	0.046	0.037

(a) The relative proportions lead to means $y_i = \sum_{j=1}^{10} j\widehat{p}_{i,j}$ and variances $s_i^2 = \sum_{j=1}^{10}(j - y_i)^2\widehat{p}_{i,j}$, which can be tracked through the book's 487 chapters. We wiew the means as $y_i \sim \mathrm{N}(\xi_i, \sigma_i^2/m_i)$, with y_i and s_i estimating ξ_i and σ_i. We first attempt to assess whether there indeed has been a change in literary style, in these wordlength distributions, via these means and standard deviations. Compute for each candidate changepoint τ the left and right means $\bar{y}_L = \sum_{i\le\tau} m_i y_i / \sum_{i\le\tau} m_i$ and $\bar{y}_R = \sum_{i\ge\tau+1} m_i y_i / \sum_{i\ge\tau+1} m_i$, along with left and right variances $s_L^2 = \sum_{i\le\tau} m_i(y_i - \bar{y}_L)^2/\tau$ and $s_R^2 = \sum_{i\ge\tau+1} m_i(y_i - \bar{y}_R)^2/(n-\tau)$. Use methods from Ex. 9.40-9.41 to construct, plot, and analyse the monitoring functions

$$H_{n,1}(\tau) = \frac{\bar{y}_R - \bar{y}_L}{\{s_L^2/\sum_{i\le\tau} m_i + s_R^2/\sum_{i\ge\tau+1} m_i\}^{1/2}},$$
$$H_{n,2}(\tau) = \frac{\bar{s}_R - \bar{s}_L}{\{\frac{1}{2}s_L^2/\tau + \frac{1}{2}s_R^2/(n-\tau)\}^{1/2}},$$

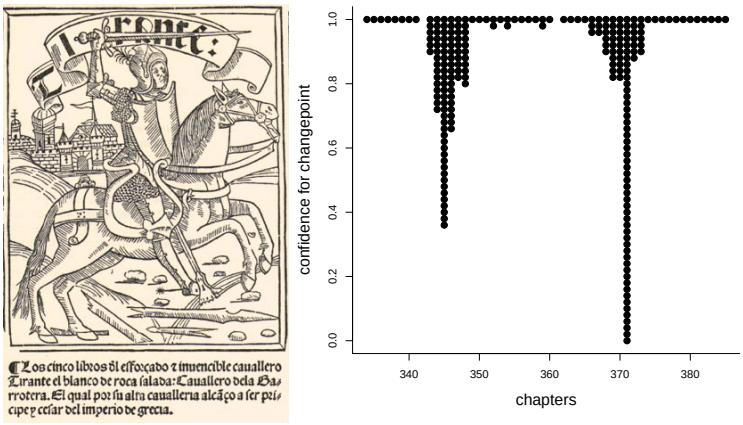


Figure ii.5: *Left panel: Tirant lo Blanch, from a 1511 edition. Right panel: estimated changepoint, chapter 371, with confidence; this is when using wordlengths 5, 6, 7, 8, 9, 10, i.e. six-dimensional vectors.*

see Figure ii.7, left panel. Explain that these should behave as normalised Browian bridges, say $T(s) = W^0(s)/\{s(1-s)\}^{1/2}$, if there has been no change in word length distributions, for Authors A and B. With plots monitored for $\tau = [0.05\,n], \ldots, [0.95\,n]$, show via Ex. 9.40 that these should be inside ± 3.645 with probability 0.99, under the null hypothesis of no change. We see from these monitoring plots that the no-change hypothesis is very firmly rejected regarding the means but less clearly regarding the standard deviations. Going on to estimate the changepoint positions, for means and standard deviations, via the maximum relative change for $\bar{y}_R - \bar{y}_L$ and $\bar{s}_R - \bar{s}_L$, find $\widehat{\tau}_1 = 295$ (which is Chapter 345) and $\widehat{\tau}_2 = 319$ (which is Chapter 370), and construct versions of Figure ii.6. Tentatively, words become longer after about Chapter 345.

(b) We learn from monitoring the wordlength means y_i that there is indeed a clear break with the null hypothesis of no change, and the near triangular shape of the normalised $\bar{y}_R - \bar{y}_l$ plot of Figure ii.7 (left panel), in particular, indicates a changepoint, with slightly longer words to the right and slightly shorter words to the left, seen also in Figure ii.6. Now carry out similar monitoring for each word length $j = 1, \ldots, 10$. This involves computing, for each of the ten $p = \Pr(w = j)$, and for each τ, left and right estimated probabilities $\bar{p}_L = \sum_{i \le \tau} m_i \widehat{p}_i / \sum_{i \le \tau} m_i$ and $\bar{p}_R = \sum_{i \ge \tau+1} m_i \widehat{p}_i / \sum_{i \ge \tau+1} m_i$, along with left and right estimated variances $\widehat{\sigma}^2_L = \sum_{i \le \tau} (\widehat{p}_i - \bar{p}_L)^2/\tau$ and $\widehat{\sigma}^2_R = \sum_{i \ge \tau+1} (\widehat{p}_i - \bar{p}_R)^2/(n-\tau)$. Show via Ex. 9.40-9.41 again that the plots

$$M_n(\tau) = \frac{\bar{p}_R - \bar{p}_L}{(\widehat{\sigma}^2_L / \sum_{i \le \tau} m_i + \widehat{\sigma}^2_R / \sum_{i \ge \tau+1} m_i)^{1/2}}$$

approach normalised Brownian bridges, under the no change hypothesis. Read off minima and maxima for these plots, shown in Figure ii.7, right panel, to tentatively state that *longer words become more frequent* (particularly those of length 7, 9, 10) whereas shorter words become less frequent, with these tentative changes taking place inside the range from Chapter 300 to Chapter 400.

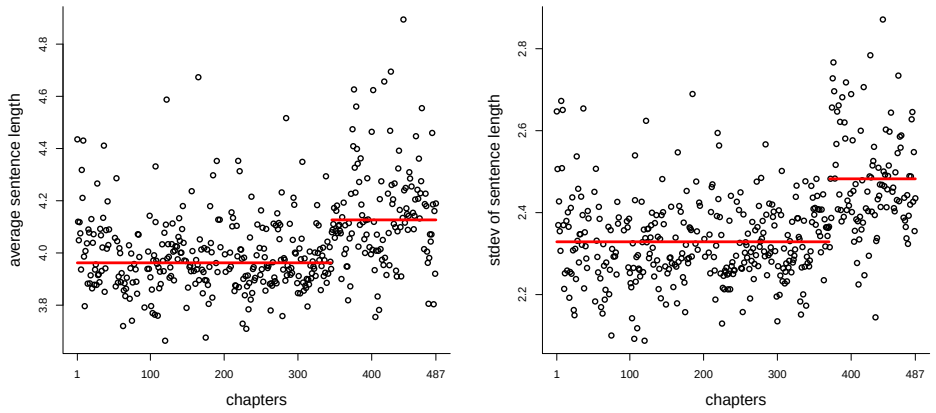


Figure ii.6: *Left panel: average word length, per chapter, with estimated changepoint at chapter 345; overall average changes from 3.962 to 4.126. Right panel: standard deviation of word lengths, per chapter, with estimated changepoint at chapter 370; overall standard deviation changes from 2.328 to 2.482.*

(c) In addition to the already useful monitoring plots above, better statistical precision will come out of examining models and log-likelihoods. Consider then any relevant $p = \Pr(A)$, with A a subset of $\{1, \ldots, 10\}$, where we estimate $p_1, \ldots, p_n$ by summing the basic relative frequencies over elements of A. Consider the model which takes $\widehat{p}_j$ to stem from $\mathrm{N}(p_L, \sigma_L^2/m_j)$ for $j \le \tau$ and then from $\mathrm{N}(p_R, \sigma_R^2/m_j)$ for $j \ge \tau + 1$. Show first that the log-likelihood $\ell(\tau, p_L, \sigma_L, p_R, \sigma_R)$ becomes

$$\sum_{j \le \tau} \{-\log \sigma_L - \tfrac{1}{2} m_j (\widehat{p}_j - p_L)^2/\sigma_L^2\} + \sum_{j \ge \tau+1} \{-\log \sigma_R - \tfrac{1}{2} m_j (\widehat{p}_j - p_R)^2/\sigma_R^2\}$$

plus the for the present purposes immaterial constant $\sum_{j=1}^n \{\frac{1}{2} \log m_j - \frac{1}{2} \log(2\pi)\}$. For fixed τ, show that this is maximised for

$$\widehat{p}_L = \sum_{j \le \tau} (m_j/m_L) \widehat{p}_j, \quad \widehat{\sigma}_L^2 = (1/\tau) \sum_{j \le \tau} m_j (\widehat{p}_j - \widehat{p}_L)^2,$$

with similar expressions for the maximisers $\widehat{p}_R, \widehat{\sigma}_R$. Deduce that the profiled log-likelihood function is

$$\ell_{\mathrm{prof}}(\tau) = -\tau \log \widehat{\sigma}_L - (n - \tau) \log \widehat{\sigma}_R.$$

As a parallel to the monitoring plots $M_n(\tau)$ above, one for each wordlength $j = 1, \ldots, 10$, construct and plot the ten deviance functions $D_n(\tau) = 2\{\ell_{\mathrm{prof,max}} - \ell_{\mathrm{prof}}(\tau)\}$. Do also this for the subset $A = \{7, 8, 9, 10\}$, monitoring the longer words across the chapters.

(d) Then consider a full vector $\widehat{p}_j$ of a subset of size q of the relative frequencies $(\widehat{p}_{j,1}, \ldots, \widehat{p}_{j,10})$. Work with the model taking $\widehat{p}_j \sim \mathrm{N}_q(p_L, \Sigma_L/m_j)$ for $j \le \tau$ and $\widehat{p}_j \sim \mathrm{N}_q(p_R, \Sigma_R/m_j)$ for $j \ge \tau + 1$. Show in generalisation of the above that for a fixed candidate value τ, the ML estimators become

$$\widehat{p}_L = \sum_{j \le \tau} (m_j/m_L) \widehat{p}_j, \quad \widehat{\Sigma}_L = (1/\tau) \sum_{j \le \tau} m_j (\widehat{p}_j - \widehat{p}_L)(\widehat{p}_j - \widehat{p}_L)^{\mathrm{t}},$$

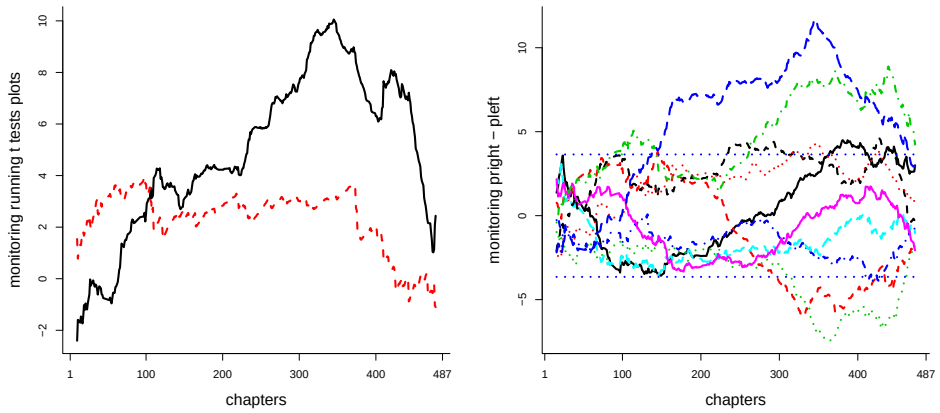


Figure ii.7: *Left panel: monitoring plots for the means ξ and standard deviations σ of wordlengths. Right panel: monitoring plots for each $p_j = \Pr(w = j)$, for word lengths $j = 1, \ldots, 10$. The horizonal lines at ± 3.645 incidate the 99 percent band for such plots, under the hypothesis of there being no change between authors A and B.*

with similar expressions for the parameters to the right. Show then that

$$\ell_{\text{prof}}(\tau) = -\tfrac{1}{2}\tau \log|\widehat{\Sigma}_L| - \tfrac{1}{2}(n-\tau)\log|\widehat{\Sigma}_R|.$$

Plot this profile function, along with the deviance function $D(\tau) = 2\{\ell_{\text{prof,max}} - \ell_{\text{prof}}(\tau)\}$, for a few choices of A, subset of $\{1, \ldots, 10\}$. For the full 9-vector $\{1, \ldots, 9\}$, one finds $\widehat{\tau} = 320$, corresponding to Chapter 371, and with some other choices the plot favours $\widehat{\tau} = 295$, making Chapter 345 the more likely changepoint.

(e) A demanding task is then to supplement the changepoint estimate with confidence intervals. A full confidence curve, as per Ch. 7, can be constructed as

$$\text{cc}(\tau) = \Pr_\tau\{D(\tau, Y^*) < D(\tau, y_{\text{obs}})\} + \tfrac{1}{2}\Pr_\tau\{D(\tau, Y^*) = D_{\text{obs}}(\tau, y_{\text{obs}})\},$$

with Y^* a full simulated dataset, for the given candidate position τ, drawn from the multinormal model with estimated position for p_L, Σ_L to the left of τ and p_R, Σ_R to the right. This requires somewhat laborious bookkeeping and simulation, with say 10^3 such full paths of multinormal data, for each τ. Carry out such a scheme, for the case of wordlengths $A = \{5, 6, 7, 8, 9, 10\}$, leading to a version of Figure ii.5, right panel. The point estimate is at Chapter 371, but still with some probability around Chapter 345, leading to confidence sets being unions of disjoint likely sets. You may similarly run such a programme for other wordlength subsets. Check also with Story iii.3, for building a similarly structured confidence curve for a changepoint, but in a rather easier Poisson model for a simpler dataset.

Story ii.4 *The children of Odin*
Data: `in text below`
Code: `(xx to come com120* nilswork22 xx)`
As we know, Odin had six male offspring – Thor, Balder, Vitharr, Váli, Heimdallr, Bragi – with the sources saying nothing about daughters. So how many children is it likely that

he had, in total? With N the number of children, and y the number of boys, we assume $y \,|\, N \sim \text{binom}(N, p)$, with $p = 0.514$ (a good overall point estimate human reproduction; see Story i.2). So the data is that $y = 6$, and we can attempt confidence inference for N. The themes and details below expand on those given in Schweder and Hjort (2016, Example 3.11).

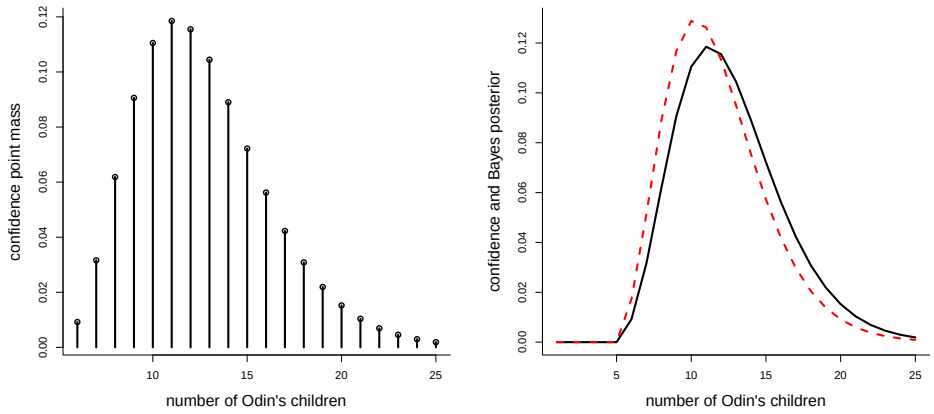


Figure ii.8: *Left panel: confidence point masses $c(N, y)$, for $N \geq 6$. Right panel: The confidence point masses (full line) alongside the Bayesian posterior, with the prior* $1/(N+1)$.

(a) A natural construction for a CD is

$$C(N, y) = \Pr_N(Y > y) + \tfrac{1}{2}\Pr_N(Y = y),$$

with the half-correction for discreteness, as in the partly parallel situation of Ex. 7.34. Compute and display this CD, and take differences to compute also the confidence point masses, $c(N, y)$. Construct a version of Figure ii.8, left panel.

(b) A CD $C(\theta, y)$, for a parameter θ based on data y, should ideally have the uniformity property that $U = C(\theta_0, Y)$ has the uniform distribution, for any fixed θ_0, with Y a random dataset drawn from the model at that position in the parameter space. This is not quite possible here, since the situation is discrete, with not many values to attain for y. For a given $N_0 = 14$, simulate say 10^4 realisations of $U = C(N_0, Y)$, then compute and display the empirical distribution function $\Pr(U \leq u)$. Comment on your findings.

(c) Carry out also a Bayesian analysis, using the prior proportional to $1/(N + 1)$ for $N \geq 0$. Compare the posterior distribution to the CD, as with Figure ii.8, right panel.

(d) Invent your own prior for N (formed before you learn in school that $y = 6$), and compare the posterior distribution wit that that found above.

(e) The frequentist CD $C(N, y)$ above should be trusted as a good and neutral statistical summary function for the unknown N. Find and display the Bayesian prior that would give the same result.

(f) In some of the Snorri kennings there are also references to Týr and Höd as sons of Odin (and yet other names are mentioned in the somewhat apocryphical Skáldskaparmál). Adjust the calculations above to this revised case, with $y = 8$, and comment on your findings.

(g) Find or dream up another situation (not necessarily with full data) where the model above might be used, i.e. p is known, but the binomial N is unknown.

Story ii.5 *How many Abel envelopes from 1902?*
Data: `in text below`
Code: `(xx to come com335* nilswork22 xx)`
(xx needs finetuning with the figures; see com335* of nilswork22. xx) Hundred years after the death of Niels Henrik Abel (1802–1829), the Norwegian postal office issued a certain stamp and a 'first-day cover' envelope commemorating him; this was only the second time such an honour had been bestowed upon a person outside royalty (in 1928, a similar first-day cover had been issued for Henrik Ibsen, hundred years after his being born). As the facsimile of Figure ii.9 (left panel) indicates, these carry 'R numbers' (as in 'rekommandert post'), and R numbers from five such Abel 1929 envelopes, from various philatelic sales lists and auctions in the 2003–2008 period, were 280, 304, 308, 310, 328. The operating assumption is that the Abelian first-day covers with stamps were produced in a running noninterrupted sequence, but one does not know when it started and neither when it ended. So how many were there? An answer to this curiosity question also enters the realm of philatelic market prices and speculations (one such specimen might fetch 5000 kroner, in 2025).

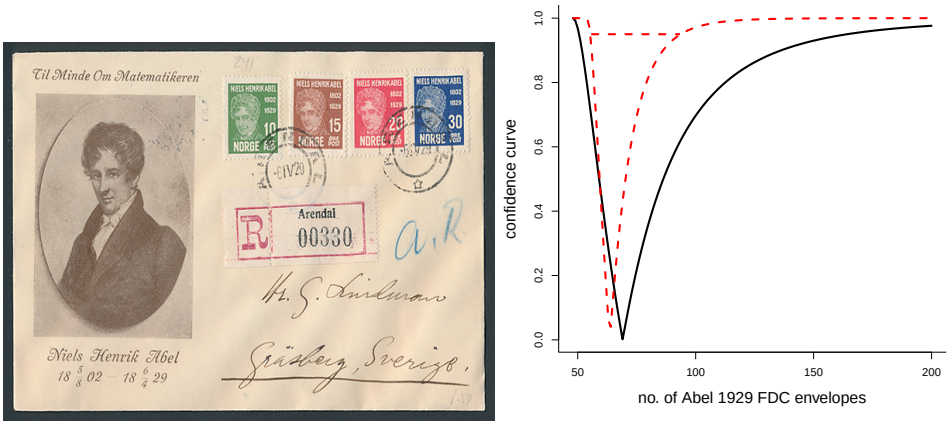


Figure ii.9: *Left panel: a philatelic rarity: an Abel first-day envelope from 1929. Right panel: confidence curves for the unknown number N of such envelopes, based on the first piece of information (black curve), with 5 envelopes, and after the updated information (red slanted), now with 10 known envelopes. 95 percent intervals for the unknown N are from 51 to 164 (with the first 5 envelopes), then narrowed down to 55 to 94 (with the updated information).*

(a) We allow ourselves a statistical detour, discussing natural setups and solutions for range estimation for the case of continuous data, before returning to Abel. So, for a concrete illustration, consider the numbers

```
4.712 6.412 7.043 7.141 7.245 7.379 7.602 8.417 8.671 8.702
```

We've simulated these $n = 10$ points, from a uniform distribution over $[a, b]$, and ordered them, for simplicity. But we won't tell you the values we used for a or b, or indeed the range $\gamma = b - a$. Your task will be to make inference about this γ. We come back to Bayesian solutions below, but now approach the problem using frequentist confidence distributions. With $Y_1, \ldots, Y_n$ from the uniform on $[a, b]$, explain that one may write $Y_i = a + (b-a)U_i$, with the U_i from the standard uniform over the unit interval. Deduce that $R_n = Y_{(n)} - Y_{(1)} = \gamma R_{n,0}$, with $R_{n,0} = U_{(n)} - U_{(1)}$, relating the range of data naturally to the range of a uniform sample. Explain that R_n/γ is a pivot, as defined in Ex. 7.9.

(b) With H_n the c.d.f. of the uniform range $R_{n,0}$ distribution, show that the canonical confidence distribution for γ becomes $C_n(\gamma, \text{data}) = \Pr_{a,b}(R_n \geq R_{n,\text{obs}}) = 1 - H_n(R_{n,\text{obs}}/\gamma)$, for $\gamma \geq R_{n,\text{obs}}$ (here observed to be 3.990). Simulate say 10^4 realisations of $R_{n,0}$ in your computer, and use these to compute and display the CD $C_n(\gamma, \text{data})$, as well as the confidence curve $\text{cc}_n(\gamma, \text{data}) = |1 - 2\,C_n(\gamma, \text{data})|$. Use also the explicit knowledge from Ex. 3.17, that H_n actually is a $\text{Be}(n-1, 2)$, to show that the confidence distribution and its confidence density become

$$\begin{aligned} C_n(\gamma, \text{data}) &= 1 - n(R_{n,\text{obs}}/\gamma)^{n-1} + (n-1)(R_{n,\text{obs}}/\gamma)^n, \\ c_n(\gamma, \text{data}) &= n(n-1)R_{n,\text{obs}}^{n-1}(\gamma - R_{n,\text{obs}})/\gamma^{n+1}, \end{aligned}$$

for $\gamma \geq R_{n,\text{obs}}$. Compute the median confidence and maximum confidence estimates.

(c) We now approach the inference problem with Bayesian means. Starting with the likelihood function, show that it can be written as follows, expressed as a function of (a, γ) rather than of (a, b):

$$L_n(a, \gamma) = (1/\gamma)^n \, I(a \leq y_{(1)} \text{ and } y_{(n)} \leq a + \gamma),$$

in particular taking the value zero if $a > y_{(1)}$ or $y_{(n)} > a + \gamma$. Find the ML estimates for a and for γ. Then, with a flat prior on a, independently of a prior $p(\gamma)$ for γ, show that the posterior distribution of γ is

$$p(\gamma \,|\, \text{data}) \propto p(\gamma)(\gamma - R_{n,\text{obs}})(1/\gamma)^n \quad \text{for } \gamma \geq R_{n,\text{obs}}.$$

(d) Without clear prior knowledge concerning the range a natural prior is proportional to $1/\gamma$. Show that this leads to the Bayesian posterior distribution agreeing precisely with the frequentist CD above. In particular, explain that the Bayes machine, starting from the $1/\gamma$ prior, leads to credibility intervals with perfect frequentist coverage. The 95 percent interval becomes $[4.094, 7.189]$, for example. As an alternative, consider also using a flat prior for γ, and show that this leads to a posterior with density and cumulative equal to (xx check all details here xx)

$$\begin{aligned} g_n(\gamma \,|\, \text{data}) &= (n-1)(n-2)R_{n,\text{obs}}^{n-2}(\gamma - R_{n,\text{obs}})/\gamma^n, \\ G_n(\gamma \,|\, \text{data}) &= 1 - (n-1)(R_{n,\text{obs}}/\gamma)^{n-2} + (n-2)(R_{n,\text{obs}}/\gamma)^{n-1}, \end{aligned}$$

for $\gamma \geq R_{n,\text{obs}}$. Plot both posteriors (with one of these equal to the CD) for the dataset above.

(e) We now return to the Abel numbers 280, 304, 308, 310, 328, first with a natural CD approach. Take these to be a random sample $X_1, \ldots, X_n$ (without replacement) of size $n = 5$ from $\{a+1, \ldots, a+N\}$, with both a and N unknown. It is natural to base the inference on the range $R_n = V_n - U_n$, where $U_n = \min_{i \le n} X_i$ and $V_n = \max_{i \le n} X_i$. Show that its distribution is independent of a. Argue that this leads to the confidence distribution $C(N) = \Pr_N(R_n > 48) + \frac{1}{2}\Pr_N(R_n = 48)$; as usual, $\Pr_N$ signals probability calculations under the value N of this parameter.

(f) It remains to find expressions for the distribution of R_n. Consider first the joint distribution of (U_n^0, V_n^0), where U_n^0 and V_n^0 are as U_n and V_n, but in the situation where $a = 0$. Show that their joint probability distribution can be expressed as

$$f(u,v) = \binom{v-u-1}{n-2} \Big/ \binom{N}{n} \quad \text{for } 1 \le u,\ u+n-1 \le v \le N.$$

Deduce that the distribution of $Z_n = V_n - U_n$ can be written

$$\Pr_N(Z_n = z) = \sum_{v-u=z} f(u,v) = (N-z)\binom{z-1}{n-2} \Big/ \binom{N}{n}$$

for $z = n-1, \ldots, N-1$. Compute and display the CD and the confidence curve $\mathrm{cc}(N)$ for N.

(g) We then work towards a Bayesian solution, based on data $X_1, \ldots, X_n$ as above, a random draw from $\{a+1, \ldots, a+N\}$. With independent priors $p_0(a)$ and $p(N)$ for the start-point a and sequence length N, show that

$$p(a, N \,|\, \text{data}) \propto p_0(a)p(N)\, I(a+1 \le U_n < V_n \le a+N) / \binom{N}{n}.$$

With a flat prior on the starting point a, show that this under some conditions leads to

$$p(N \,|\, \text{data}) \propto p(N) \frac{(N-R_n)}{N(N-1)\cdots(N-n+1)};$$

note the partial similarity to the posterior $p(\gamma \,|\, \text{data})$ in the continuous case above. Work out the posterior distribution for N, over a suitable range of N values, starting with a flat prior. Find posterior median and a 95 percent interval. – One ought to be careful here, since the prior for a should be flat on $1, 2, 3, \ldots$, not including negative numbers. Show that the associated refinement of the direct result above becomes $p(N \,|\, \text{data}) \propto p(N)q(N)/\{N(N-1)\cdots(N-n+1)\}$, where $q(N)$ counts the number of $a \ge 1$ satisfying $V_n - N \le a \le U_n - 1$. Show that this means either $U_n - 1 - (V_n - N - 1) = N - R_n$, provided $V_n \ge N+1$, or $U_n - 1$, in the case of $V_n \le N$. In other words and symbols, $q(N) = (N - R_n)\, I(V_n > N) + (U_n - 1)\, I(V_n \le N)$. Show however that for the present occasion, the relevant values of N are smaller than $V_n = 328$, so this additional layer of care turns out not to be needed.

(h) In addition to the five R numbers 280, 304, 308, 310, 328 known as of 2008, five more such first-day Abel envelopes have been unearthed, the latest in 2022: 285, 314, 317, 327, 334. Update your inference, for the CD and the Bayesian posterior, and construct versions of Figure ii.10; CDs in the left panel and Bayesian cumulatives in the right. Translate also the CDs to confidence curves, as with Figure ii.10, right panel. Compute

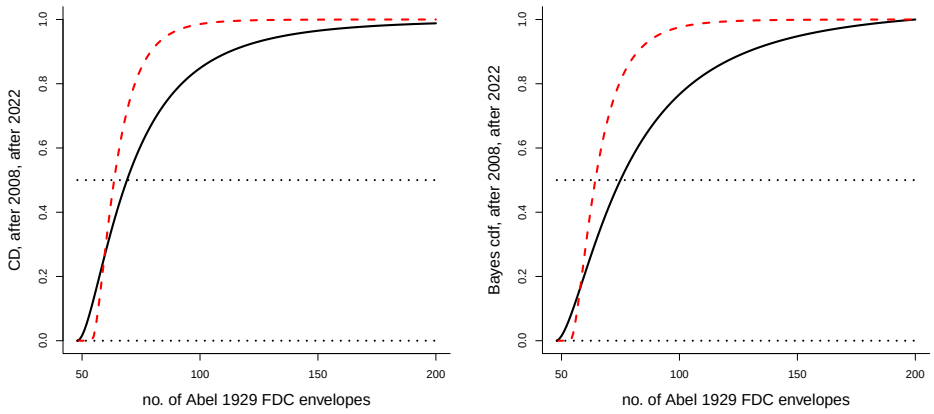


Figure ii.10: *Left panel: confidence distribution for N, the number of Abel 1929 first-day cover envelopes, based on the 5 known numbers by 2008 (full curve), and on the now 10 known numbers by 2022 (dashed curve). Right panel: for the same information, the Bayesian posterior c.d.f.*

also the median confidence and median Bayes estimates, along with 95 percent intervals. (Answers: for data up to 2008, point estimates are 69 and 75, with intervals $[51, 164]$ and $[51, 170]$, for the CD and the Bayes. With extended data up to 2022, point estimates are 64 and 64, with intervals $[55, 94]$ and $[55, 100]$.)

Story ii.6 *Markov and Pushkin*
Data: `onegin-data (xx to be organised xx)`
Code: `(xx to come com321* nilswork23 xx)`
(xx calibrate things with what's in Ch. 12. calibrate with what we describe in dataoverview. xx) A.A. Markov invented Markov chains in 1906, with monumental consequences for probability theory, statistics, dependence models, time series, Bayesian computation and simulation, and for a steadily increasing range of applications in multiple domains, from biology and economics to mobile phones and AI. He should also be remembered for having presented the first ever *data analysis* of a Markov chain. Astoundingly, he went through the first 20,000 letters of Pushkin's classic 1833 epic poem Евгений Онегин, tabling transitions from vowels and consonants, reporting his findings in Markov (1913). In Markov model language, with transition probabilities $p_{a,b}$ on the two-state $\{0, 1\}$, with 0 for согласний and 1 for гласний, and where Markov demonstrated that $\widehat{p}_{0,1}$ (consonant to vowel) is clearly different from $\widehat{p}_{1,1}$ (vowel to vowel); see the analysis below.

will disappear as fast as smoke: for Satan, love's a splendid joke

The crucial point is that consonants and vowels do not flow independently, when we speak or think or recite poems. In statistical terms, this standard 1st order memory Markov model increases the explanatory power enormously, when compared to the too simple independence model. As we shall see, however, the 1st order memory model is not good enough either; checking triples of transitions, with a 2nd order memory model $p_{a,b,c} = \Pr(X_{i+1} = c \,|\, X_{i-1} = a, X_i = b)$ for $a, b, c \in \{0, 1\}$, there are clear signs of departure from the 1st order memory model. The 33 letter Russian alphabet has 10 vowels and 21 consonants (where we include й), along with the soft-sign and hard-sign. We base our analysis below on the $2^3 = 8$ triple counts $N_{a,b,c} = \sum_{i=2}^{n-1} I\{(X_{i-1}, X_i, X_{i+1}) = (a, b, c)\}$,

This has involved taking a Russian online text-file of Pushkin's poem, available at wikisource, cleaning away punctuation and the occasional non-Russian words (in French, English, Italian), then patiently letter by letter find-replace-ing consonants and vowels to 0s and 1s for the first $n = 20000$ letters, where we have disregarded the soft-sign and hard-sign letters. The triple counts thus found are as follows, summing to 19998:

$$N_{0,0,0} = 644,\ \ N_{0,0,1} = 3516,\ N_{0,1,0} = 7018,\ N_{0,1,1} = 593$$
$$N_{1,0,0} = 3516,\ N_{1,0,1} = 4095,\ \ N_{1,1,0} = 593,\ \ \ N_{1,1,1} = 23$$

(a) Before coming to the triples, carry out 1st order Markov chain estimation, via $N_{a,b} = \sum_c N_{a,b,c}$ to find the transition matrix

$$P = \begin{pmatrix} 0.353,\ 0.647 \\ 0.925,\ 0.075 \end{pmatrix}.$$

Find the equilibrium distribution, with 58.9 percent consonants and 41.1 percent vowels.

(b) (xx calibrate and smooth. xx) For counting transitions in a 1st order memory Markov chain, yielding transition probability estimators $\widehat{p}_{a,b} = N_{a,b}/N_{a,\cdot}$ we have learned in Ex. 12.17 that these behave as multinomial ratios, given $N_{a,\cdot}$, with covariances $p_{a,b}(\delta_{b,b'} - p_{a,b'})/N_{a,\cdot}$, and independently for different a. Now we need to generalise this to 2nd order memory Markov chains. With no further structure on these $k^2(k-1)$ parameters $p_{a,b,c}$, show that the log-likelihood function is $\ell_n = \sum_{a,b,c} N_{a,b,c} \log p_{a,b,c}$, with ML estimators $\widehat{p}_{a,b,c} = N_{a,b,c}/N_{a,b,\cdot}$. Show that these ratios with given $N_{a,b,\cdot}$ behave as in a multinomial setup, and independently for (a,b) different from (a',b'). This makes it possible to estimate variances of all smooth functions of the $p_{a,b,c}$ parameters, via the delta metod.

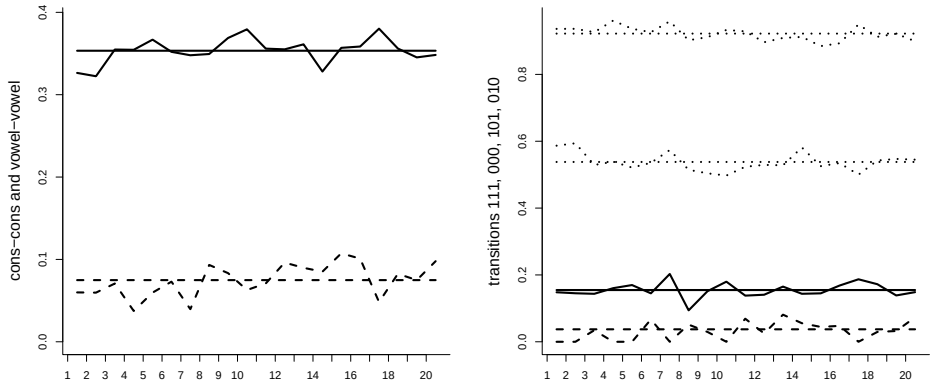


Figure ii.11: *Transition probabilities for vowel-consonant shifts in Pushkin's Onegin, here computed for the $k = 20$ consecutive parts with 1000 letters in each. With 0 for consonants and 1 for vowels: Left panel: 1st order Markov, $p_{1,1}$ (mean 0.075) and $p_{0,0}$ (mean 0.353). Right panel: $p_{1,1,1}$ (mean 0.037), $p_{0,0,0}$ (mean 0.155), $p_{1,0,1}$ (mean 0.538), $p_{0,1,0}$ (mean 0.022).*

(c) Coming back to Pushkin, we model the vowels and consonants as 0s and 1s from a 2nd order memory Markov chain, to learn the extent to which Markov's 1st order memory Markov chain model from 1913 is too simplistic. Consider the ratios

$$\rho_{0,0} = p_{1,0,0}/p_{0,0,0}, \quad \rho_{0,1} = p_{1,0,1}/p_{0,0,1}, \quad \rho_{1,0} = p_{1,1,0}/p_{0,1,0}, \quad \rho_{1,1} = p_{1,1,1}/p_{0,1,1}.$$

Explain that these should all be close to 1 under Markov's 1st order memory assumptions. Show that $\widehat{\rho}_{0,0} = \widehat{p}_{1,0,0}/\widehat{p}_{0,0,0}$ has approximate variance

$$\begin{aligned}\tau_{0,0}^2 &= \frac{1}{(p_{0,0,0})^2}\frac{p_{1,0,0}\,p_{1,0,1}}{N_{1,0,\cdot}} + \frac{(p_{1,0,0})^2}{(p_{0,0,0})^4}\frac{p_{0,0,0}\,p_{0,0,1}}{N_{0,0,\cdot}} \\ &= \frac{1}{(p_{0,0,0})^2}\Big\{\frac{p_{1,0,0}\,p_{1,0,1}}{N_{1,0,\cdot}} + \rho_{0,0}^2\frac{p_{0,0,0}\,p_{0,0,1}}{N_{0,0,\cdot}}\Big\}.\end{aligned}$$

Give estimates and estimated standard deviations for the four $\rho_{a,b}$ ratios. The 1st order probability of coming to vowel from vowel is 0.0749, for example. The more carefully estimated 2nd order probabilities are 0.0373 from (vowel, vowel) to vowel, rather smaller than 0.0779 from (consonant, vowel) to vowel. Explain how this contradicts the 1st order memory Markov model.

	from 1	from 0	rho	tau
00	0.4620	0.1548	2.9841	0.1142
01	0.5380	0.8452	0.6366	0.0080
10	0.9627	0.9221	1.0440	0.0090
11	0.0373	0.0779	0.4792	0.0998

(d) Was Pushkin reasonably consistent, regarding the vowel-consonant interplay, throughout his long work? Divide the long chain of 20000 vowels and consonants into $k = 20$ consecutive chunks, each of length 1000, and for each of these compute the 1st order $\widehat{p}_{a,b}$ and 2nd order $\widehat{p}_{a,b,c}$ transition frequencies, for $a, b = 0, 1$ (again with 0 for consonant and 1 for vowel). Give versions of Figure ii.11, left and right panels. Then carry out simple t-testing of equality of parts 1-10 and 11-20. For (consonant, vowel) to consonant, for instance, show that the $p_{0,1,0}$ have not remained constant over the full work.

(e) (xx this perhaps to be omitted. xx) How would you go about estimation in and testing of the 3rd order memory Markov model here? (xx may ask for log-likelihood maxima for Markov chains of order 0, 1, 2, 3. parameter dimensions are 1, 2, 4, 8. apparently, the 2nd order model, with 4 parameters, is better than the 3rd order model, with 8 parameters. xx)

Story ii.7 *And Quiet Does Not Flow the Don*
Data: `tikh-data`
Code: `(xx to come com5* nilswork23 xx)`
The Nobel Prize in literature for 1965 was awarded Mikhail Sholokhov (1905–1984), for the epic novel Тихий Дон about Cossack life and the birth of a new Soviet society. Sholokhov has been compared to Tolstoĭ and was at least a generation ago called 'the greatest of our writers' in Russia and in the Soviet Union, with thousands of editions of his novels and stories. But in the autumn of 1974 an article was published in Paris, Стремя 'Тихого Дона' (Загадки романа) *('The Rapids of Quiet Don: the Enigmas of the Novel')*, by the author and critic D*. He claimed that Tikhiĭ Don was not at all Sholokhov's work, but that it rather was written by Fiodor Kriukov, a more obscure

author who fought against bolshevism and died in 1920. The article was given credibility and prestige by none other than Aleksandr Solzhenitsyn (a Nobel prize winner five years after Sholokhov, with a history of previous quarrels), who wrote a preface giving full support to D*'s conclusion. Scandals followed, also touching the upper echelons of Soviet society, and Sholokhov's reputation was faltering abroad ('vibrations of dislike instantly flowed between us', writes Lessing (1997), another Nobel Prize winner). Are we in fact faced with one of the most flagrant cases of plagiarism in the history of literature?

Even experts on literature, art, and music are prone to making occasional mistakes, as demonstrated often enough, and it is clear that independent arguments based on quantitative comparisons are of interest. If not taken as 'direct proof', then such comparisons may at least offer independent objective evidence and sometimes additional insights. In such a spirit, an inter-Nordic research team was formed in 1975, captained by Geir Kjetsaa, a professor of Russian literature at the University of Oslo, with the aim of disentangling the Don mystery. In addition to various linguistic analyses and several doses of detective work, quantitative data were gathered and organised, relating to word lengths, sentence lengths, frequencies of certain words and phrases, grammatical characteristics, etc. These data were extracted from three corpora: (i) Ш, or Sh, 4183 sentences, from published work guaranteed to be by Sholokhov; (ii) Кр, or Kr, 3739 sentences, that which with equal trustworthiness came from the hand of the alternative hypothesis Kriukov; and (iii) ТД, or TD, the Nobel winning apple of discord, with 3760 sentences. The `Nx` numbers below, the number of sentences with lengths inside windows 1-5, 6-10, 11-15, etc., have been extracted from tables in Kjetsaa et al. (1984). Our contribution here is to squeeze clearer author discrimination and some deeper statistical insights out of some of Kjetsaa et al.'s data. In particular, with N_j lengths inside window j below, the `expected` numbers $\widehat{N}_j$ and `pearson` residuals $(N_j - \widehat{N}_j)/\widehat{N}_j^{1/2}$ are computed using a certain parametric sentence length distribution developed below, then used to discriminate between authors. See also Hjort (2007) for further background, details, and analyses.

	Sh:			TD:			Kr:		
	Nx	expected	pearson	Nx	expected	pearson	Nx	expected	pearson
1- 5	799	803.38	-0.15	684	690.10	-0.23	714	717.56	-0.13
6-10	1408	1396.97	0.30	1212	1188.52	0.68	1046	1038.89	0.22
11-15	875	884.84	-0.33	826	854.39	-0.97	787	793.32	-0.22
16-20	492	461.25	1.43	480	418.67	3.00	528	504.56	1.04
21-25	285	275.90	0.55	244	248.07	-0.26	317	305.22	0.67
26-30	144	161.52	-1.38	121	151.10	-2.45	165	174.82	-0.74
31-35	78	91.34	-1.40	75	89.66	-1.55	78	96.12	-1.85
36-40	37	50.34	-1.88	48	52.08	-0.56	44	51.29	-1.02
41-45	32	27.21	0.92	31	29.76	0.23	28	26.76	0.24
46-50	13	14.49	-0.39	16	16.80	-0.19	11	13.72	-0.73
51-55	8	7.62	0.14	12	9.38	0.85	8	6.93	0.40
56-60	8	3.97	2.03	3	5.20	-0.96	5	3.46	0.83
61-65	4	2.05	1.36	8	2.86	3.04	5	1.71	2.51
	4183		16.69	3760		30.22	3730		14.41

(a) Before we begin constructing sentence length distributions, let us examine the raw table as such, with its $k = 13$ length windows 1-5 to 61-65. Viewing these as multinomial processes, with probability vectors $p_{\rm Sh}$, $p_{\rm TD}$, $p_{\rm Kr}$ of length k, test the hypothesis H_A, that Sh and TD have the same mechanism, and then H_B, that Kr and TD have the same mechanism. You may use the $\sum_{j=1}^k (\widehat{p}_j - \widehat{q}_j)^2/\widehat{r}_j$ type test developed in Story vii.1. You

should find that H_A is accepted but that H_B is rejected, already pointing to Sholokhov being the rightful author.

(b) Wouldn't it be splendid, to be a very clever statistician and compute clear probabilities

$$\begin{aligned}
\mathrm{pr}_A &= \Pr(\text{Sh is the TD author} \,|\, \text{data}),\\
\mathrm{pr}_B &= \Pr(\text{Kr is the TD author} \,|\, \text{data}),\\
\mathrm{pr}_C &= \Pr(\text{neither of them is the TD author} \,|\, \text{data}),
\end{aligned}$$

via the sentence length data? This is ambitious, conceptually and operationally, but a Bayesian attempt is as follows. For the three scenarios here, called A, B, C, suppose (i) that prior probabilities π_A, π_B, π_C are put up, by an expert, or by yourself (and a simple neutral $(\frac13,\frac13,\frac13)$ can be used); (ii) that appropriate priors are specified for p_A and $p_{\rm Kr}$, in the case of A, for p_B and $p_{\rm Sh}$, in the case of B, and for $p_{\rm Sh}$, $p_{\rm TD}$, $p_{\rm Kr}$ in the case of C; and (iii) that the multinomial setup is judged acceptable. Show then that

$$\mathrm{pr}_A = \pi_A \bar f_A/\bar f, \qquad \mathrm{pr}_B = \pi_B \bar f_B/\bar f, \qquad \mathrm{pr}_C = \pi_C \bar f_C/\bar f, \tag{ii.1}$$

featuring marginal probabiilties

$$\begin{aligned}
\bar f_A &= \int \prod_{j=1}^k p_j^{{\rm Sh}_j+{\rm TD}_j} \pi_A(p_A)\,{\rm d}p_A \int \prod_{j=1}^k p_j^{{\rm Kr}_j} \pi_{\rm Kr}(p_{\rm Kr})\,{\rm d}p_{\rm Kr},\\
\bar f_B &= \int \prod_{j=1}^k p_j^{{\rm Kr}_j+{\rm TD}_j} \pi_B(p_B)\,{\rm d}p_B \int \prod_{j=1}^k p_j^{{\rm Sh}_j} \pi_{\rm Sh}(p_{\rm Sh})\,{\rm d}p_{\rm Sh},\\
\bar f_C &= \int \prod_{j=1}^k p_j^{{\rm Sh}_j} \pi_{\rm Sh}(p_{\rm Sh})\,{\rm d}p_{\rm Sh} \int \prod_{j=1}^k p_j^{{\rm TD}_j} \pi_{\rm Kr}(p_{\rm TD})\,{\rm d}p_{\rm TD} \int \prod_{j=1}^k p_j^{{\rm Kr}_j} \pi_{\rm Kr}(p_{\rm Kr})\,{\rm d}p_{\rm Kr}.
\end{aligned}$$

Also, $\bar f = \pi_A \bar f_A + \pi_B \bar f_B + \pi_C \bar f_C$, so indeed $\mathrm{pr}_A, \mathrm{pr}_B, \mathrm{pr}_C$ sum to one.

(c) This might look forbidding but is actually doable. Argue first that it makes sense to employ the same prior, for the different $p = (p_1, \ldots, p_k)$ probability vectors here, so that differences in the posterior probabilities pr_A, pr_B, pr_C will not be due to differences in prior perceptions of the sentence length distributions; also, these are rather similar across the three corpora. Secondly, it helps to choose this prior as a Dirichlet, say $p \sim \mathrm{Dir}(cp_0)$ with mean $p_0 = (p_{0,1}, \ldots, p_{0,k})$; see Ex. **??**. Show that

$$\begin{aligned}
\bar f_A &= K^2 \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm Sh}_j + {\rm TD}_j)}{\Gamma(c + n_{\rm Sh} + n_{\rm TD})} \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm Kr}_j)}{\Gamma(c + n_{\rm Kr})},\\
\bar f_B &= K^2 \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm Kr}_j + {\rm TD}_j)}{\Gamma(c + n_{\rm Kr} + n_{\rm TD})} \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm Sh}_j)}{\Gamma(c + n_{\rm Sh})},\\
\bar f_C &= K^3 \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm Sh}_j)}{\Gamma(c + n_{\rm Sh})} \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm TD}_j)}{\Gamma(c + n_{\rm TD})} \frac{\prod_{j=1}^k \Gamma(cp_{0,j} + {\rm Kr}_j)}{\Gamma(c + n_{\rm Kr})},
\end{aligned}$$

in which $K = \Gamma(c)/\prod_{j=1}^k \Gamma(cp_{0,j})$. Now implement these formulae. Use p_0 roughly equal to the normalised N_j counts for TD, and try out values e.g. $100, 1000, 4000$ for the prior sample size c. You should find that even if you start with a Solzhenitsyn type prior $(0.05, 0.90, 0.05)$, heavily favouring Kriukov, the result is an overwhelming $\mathrm{pr}_A = 0.999$ or even more.

(d) The conclusion is already rather clear, without even having attempted to model the sentence lengths beyond having a generic $(p_1, \ldots, p_{13})$ for the 13 length windows. There will be benefits from attempting such length modelling, however, because conclusions might be firmer and more informative for the present Quiet Don issue, and because such efforts might lead to fruitful comparison tools for other authors and other corpora. – We do come to more relevant parametric models below, but first we briefly consider the Poisson. Without going into finer likelihood details, compute perhaps crude estimates of means and variances, for the three corpora, using e.g. midpoints of the length windows 1-5, 6-10, etc. Demonstrate that the variances are much bigger than the means, actually with a factor of about 6. Perhaps no serious writers distribute their sentence lengths quite as primitively as via a Poisson.

(e) A sensible model for the sentence lengths needs to have at least two parameters on board. Consider the mixed Poisson, where the rate parameter is not constant but varies in the world of sentences. If Y given λ is Poisson with this parameter, but λ has a $\mathrm{Gam}(a, b)$ distribution, show as with Ex. 1.27 that the marginal takes the begative binomial form

$$f^*(y, a, b) = \frac{b^a}{\Gamma(a)} \frac{1}{y!} \frac{\Gamma(a+y)}{(b+1)^{a+y}} \quad \text{for } y = 0, 1, 2, \ldots.$$

Show that its mean is $\mu = a/b$ and its variance $a/b^2 = \mu(1 + 1/b)$, indicating the level of over-dispersion. Generally speaking, a model $\Pr(L = y) = f(y, \theta)$ for length data y implies window probabilities $q_j(\theta) = \sum_{y \in \text{window } j} f(y, \theta)$. Fit the two-parameter mixed Poisson model to the three corpora, using minimum chi-squared with the grouped data, minimising $Q(\theta) = \sum_{j=1}^k \{N_j - nq_j(\theta)\}^2 / \{nq_j(\theta)\}$. Here N_j is the number of sentences in length window j, and $n = \sum_{j=1}^k N_j$ the full number of sentences in the corpus examined. Fit this two-parameter model to the three corpora, recording also the minimum chi-squared scores and checking the Pearson type residuals $(N_j - n\widehat{q}_j)/\widehat{q}_j^{1/2}$. You will find minimum values 118.145, 130.002, 19.951 for Sh, TD, Kr, which is too large for a good model, and with several residuals out of normal range.

(f) The two-parameter mixed Poisson found to be too simple invites further probing into the sentence length mechanisms. An acceptable model, as it turns out, is the following mixture of a pure Poisson and another mixed Poisson, with a modification stemming from the fact that sentences containing zero words do not really count among Nobel literature laureates (with the notable exception of a 1958 story by Heinrich Böll):

$$f(y, p, \xi, a, b) = p \frac{\exp(-\xi)\xi^y / y!}{1 - \exp(\xi)} + (1-p) \frac{f^*(y, a, b)}{1 - f^*(0, a, b)} \quad \text{for } y = 1, 2, \ldots.$$

It would have been easier and statistically more informative to analyse the full empirical distributions, for lengths 1 to 65, but unfortunately the research team only kept the summary tables for windows 1-5, 6-10, etc. (G. Kjetsaa, personal communication to N.L.H., 1995). Explain that the log-likelihood function for these binned multinomial data becomes $\ell(\theta) = \sum_{j=1}^k N_j \log q_j(\theta)$. For the three text corpora, and for this four-parameter model, find both minimum chi-squared and maximum likelihood estimates; these are indeed quite close, see the table below. Construct versions of Figures ii.12 and ii.13, along with computing `expected` $\widehat{N}_j = nq_j(\widehat{\theta})$ and `pearson` residuals $(N_j - \widehat{N}_j)/\widehat{N}_j^{1/2}$, as in the table above. We learn that the sentence length distributions are fairly similar,

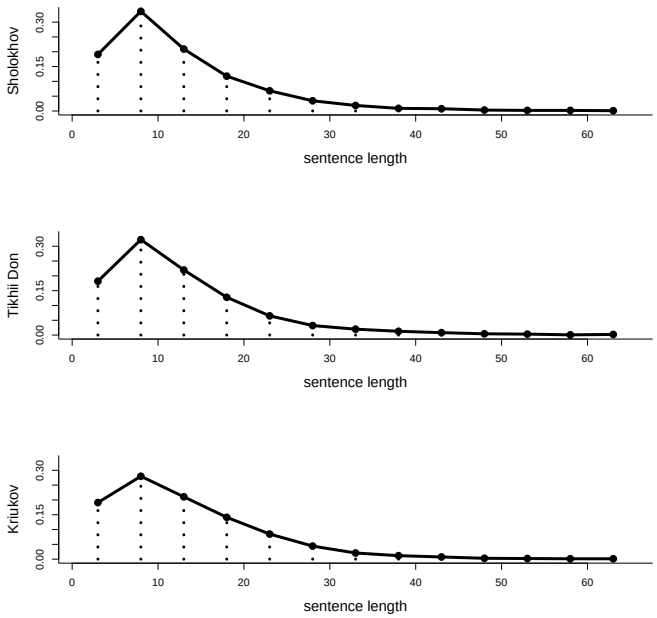


Figure ii.12: *Sentence length distributions, observed (full lines) and fitted to the four-parameter model (dotted lines), for the three text corpora Sholokhov, Tikhii Don, and Kriukov. The distributions are quite similar, and a statistical magnifying glass is needed to see which of Sh and Kr is closest to TD.*

for Sh, TD, Kr, making it a challenging statistical task to disentangle them. (xx need to sort out and to point to large-sample similarity of ML and minimum chi-squared. check Ferguson (1996). xx)

	Sh:			TD:			Kr:		
	ML	MQ	se	ML	MQ	se	ML	MQ	se
p	0.211	0.220	0.022	0.214	0.230	0.024	0.042	0.051	0.025
xi	8.530	8.678	0.407	9.476	9.649	0.389	7.802	8.707	2.280
a	2.260	2.161	0.126	2.097	1.962	0.115	2.458	2.352	0.140
b	0.172	0.163	0.009	0.157	0.145	0.009	0.186	0.177	0.009

(g) Translating the hypotheses H_A and H_B of question (a) to the present setting, with parameter vectors $\theta = (p, \xi, a, b)^{\mathrm{t}}$ driving the models, carry out natural tests for $\theta_{\mathrm{Sh}} = \theta_{\mathrm{TD}}$ and for $\theta_{\mathrm{Kr}} = \theta_{\mathrm{TD}}$, using methods of Ex. 4.43. You should find that the first is fully accepted but the second firmly rejected, in line with what was found for the probability vectors p_{Sh}, p_{TD}, p_{Kr} above. Again, this lends support to the Stalin Prize 1941 winner being the rightful author of Tikhiĭ Don.

(h) The disputed authorship question can also be approached via model selection methodology. The two main theories or claims correspond to Model A: Sh and TD stem from the same source, with common parameter θ_A, whereas Kr comes from a different θ_{Kr}; and Model B: Kr and TD come from the same source, with common parameter θ_B, and Sh comes from a different θ_{Sh}. Let us also include Model C: the three corpora stem from different sources with three different parameter vectors. Explain that Models A

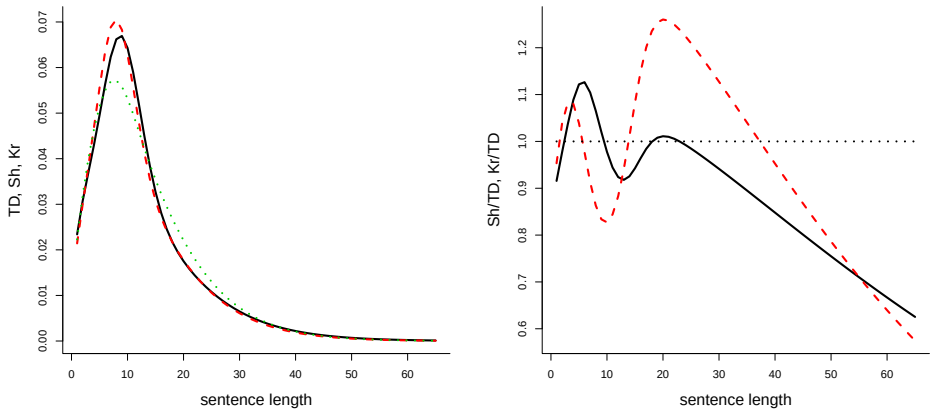


Figure ii.13: *Left panel: sentence length distributions, estimated via the four-parameter model, for the three text corpora Sholokhov, Tikhii Don, and Kriukov; Sh is rather closer to TD than Kr is. Right panel: the estimated ratios* $f_{\mathrm{Sh}}/f_{\mathrm{TD}}$ *and* $f_{\mathrm{Kr}}/f_{\mathrm{TD}}$*, the former much closer to 1 than the latter.*

and B have 8 parameters, whereas Model C has 12 parameters. Fit the three different models, via maximum likelihood, recording the implied log-likelihood maxima for the $\ell_A(\theta)$, $\ell_B(\theta)$, etc. Compute first AIC scores, and comment. For BIC scores, argue that we should have

$$\begin{aligned}
\mathrm{bic}_A &= 2(\ell_{A,\max} + \ell_{\mathrm{Kr},\max}) - 4\log(n_{\mathrm{Sh}} + n_{\mathrm{TD}}) - 4\log n_{\mathrm{Kr}},\\
\mathrm{bic}_B &= 2(\ell_{B,\max} + \ell_{\mathrm{Sh},\max}) - 4\log(n_{\mathrm{Kr}} + n_{\mathrm{TD}}) - 4\log n_{\mathrm{Sh}},\\
\mathrm{bic}_C &= 2(\ell_{\mathrm{Sh},\max} + \ell_{\mathrm{TD},\max} + \ell_{\mathrm{Kr},\max}) - 4(\log n_{\mathrm{Kr}} + \log n_{\mathrm{TD}} + \log n_{\mathrm{Kr}}).
\end{aligned}$$

How do you conclude, based on this?

(i) We computed posterior probabilities pr_A, pr_B, pr_C in the course of question (b) above, but then only used the direct window counts $N_1, \ldots, N_{13}$. With the parametric model developed above we should be able to compute more precise estimates. Argue that formulae (ii.1) can still be used, provided we have priors p_A and p_{Kr} under Model A, priors p_B and p_{Sh} under Model B, and priors p_{Sh}, p_{TD}, p_{Kr} under Model C. These involve marginal probabilities $\bar f_A, \bar f_B, \bar f_C$. Show that

$$\bar f_A = \int \Big\{\prod_{j=1}^k q_j(\theta_A)^{\mathrm{Sh}_j+\mathrm{TD}_j}\Big\} p_A(\theta_A)\,\mathrm{d}\theta_A \int \Big\{\prod_{j=1}^k q_j(\theta_{\mathrm{Kr}})^{\mathrm{Kr}_j}\Big\} p_{\mathrm{Kr}}(\theta_{\mathrm{Kr}})\,\mathrm{d}\theta_{\mathrm{Kr}},$$

involving 4-dimensional integrals over (p, ξ, a, b), and put up similar expressions for $\bar f_B$ and $\bar f_C$. Then use Laplace approximation methods of Ex. 6.26 to establish that

$$\begin{aligned}
\log \bar f_A &\doteq \tfrac12 \mathrm{bic}_A - \tfrac12 \log|J_A| - \tfrac12 \log|J_{\mathrm{Kr}}| + \log p_A(\widehat\theta_A) + \log p_{\mathrm{Kr}}(\widehat\theta_{\mathrm{Kr}}) + 4\log(2\pi),\\
\log \bar f_B &\doteq \tfrac12 \mathrm{bic}_B - \tfrac12 \log|J_B| - \tfrac12 \log|J_{\mathrm{Sh}}| + \log p_B(\widehat\theta_B) + \log p_{\mathrm{Sh}}(\widehat\theta_{\mathrm{Sh}}) + 4\log(2\pi),\\
\log \bar f_C &\doteq \tfrac12 \mathrm{bic}_C - \tfrac12 \log|J_{\mathrm{Sh}}| - \tfrac12 \log|J_{\mathrm{TD}}| - \tfrac12 \log|J_{\mathrm{Kr}}|\\
&\qquad + \log p_{\mathrm{Sh}}(\widehat\theta_{\mathrm{Sh}}) + \log p_{\mathrm{Sh}}(\widehat\theta_{\mathrm{TD}}) + \log p_{\mathrm{Sh}}(\widehat\theta_{\mathrm{Kr}}) + 6\log(2\pi).
\end{aligned}$$

Here $\ell_{A,\max}$ is the maximum of $\ell_A(\theta)$, with $J_A = -\partial^{\mathrm{t}}\ell_A(\widehat{\theta}_A)/\partial\theta\partial\theta^{\mathrm{t}}/(n_{\mathrm{Sh}}+n_{\mathrm{TD}})$ the associated normalised Hessian matrix, etc. Several paths may be considered hre, but a natural simplification is the use of Jeffreys type neutral priors, of the type $p_A(\theta) \propto |J_A|^{1/2}$. Show that this causes several terms to cancel in these approximations, and that it all leads to posterior probabilities for A, B, C, as per formulae (ii.1), with

$$\mathrm{pr}_A = \pi_A \exp(\tfrac{1}{2}\mathrm{bic}_A)/d, \quad \mathrm{pr}_B = \pi_B \exp(\tfrac{1}{2}\mathrm{bic}_B)/d, \quad \mathrm{pr}_c = \pi_C \exp(\tfrac{1}{2}\mathrm{bic}_C + 2\log(2\pi))/d.$$

Here d is the normalising constant needed to have the probabilities summing to 1. Compute these probabilities, based on neutral start prior $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$, on a Solzhenitsyn type prior $(0.01, 0.98, 0.01)$ strongly negative to Sholokhov, and perhaps your own. Show that the pro-Sholohhov probability is at least 0.999.

Story ii.8 *Republic, Laws, Critias, Philebus, Politicus, Sophist, Timaeus*
Data: `platon-data`
Code: `(xx to come com64* nilswork24 xx)`
(xx needs polish, but i'm getting there. i mention the Markov type models for $p_{i_1,\dots,i_5}$ at the end, but don't really pursue this. xx) Perhaps you, like Plato, or $\Pi\lambda\alpha\tau\omega\nu\omega\varsigma$, for stylistic or rhetorical reasons, are very careful with your sentence endings. In this case your clausula, the five last syllables, is an instance of S L S L S, for 'short' and 'long' (or 'light' and 'stressed'). There are $2^5 = 32$ variations. Corpora can be carefully read and analysed, with the different types of clausulae tabulated and compared. In probability modelling language, we then get estimates of each work's 32-dimensional clausula probability vector, say $p = (p_1, \dots, p_{32})$.

(xx to be done. data from Cox and Brandwood (1959), but here we do more. Scholars agree that A: Republic (Politeia) comes several years before B: Laws (Nomoi). There is no clear consensus regarding the correct placing of the five Socratic dialogues Critias (Kritias), Philebus (Filebos), Politicus (Politikos), Sophist (Sofistis), Timaeus (Timaios), inside the time window from Republic to Laws, however. Here we make a statistical attempt at solving this puzzle. data collected in 2.B. point briefly to Story vii.1. Sample sizes, i.e. the number of sentences or phrases from which the clausulae have been lifted and sifted, are quite big for Rep and Laws, with $n_A = 3778$ and $n_B = 3783$, but lower for the five intermediate dialogues; they are 150, 958, 770, 919, 762 for Crit, Phil, Pol, Soph, Tim. xx)

(a) Let $N = (N_1, \dots, N_k)$ be a multinomial $(n, p_1, \dots, p_k)$, see Ex. 1.5. We write $p_k = 1 - p_1 - \cdots - p_{k-1}$, making $(p_1, \dots, p_k)$ a probability vector. Show that with no further constraints on the p_j, the maximum likelihood estimators are $\widehat{p}_j = N_j/n$, with log-likelihood maximum $\ell_{\max} = n\sum_{j=1}^k \widehat{p}_j \log \widehat{p}_j + c_n$, where $c_n = \log(n!) - \sum_{j=1}^k \log(N_j!)$ is a constant depending on the data but not the parameters.

(b) Suppose $N_A = (N_{A,1}, \dots, N_{A,k})$ and $N_B = (N_{B,1}, \dots, N_{B,k})$ are two independent multinomials, with total counts n_A and n_B, and probability vectors $p_A = (p_{A,1}, \dots, p_{A,k})$ and $p_B = (p_{B,1}, \dots, p_{B,k})$, and that these count vectors can be meaningfully compared, in the sense that $p_{A,j}$ and $p_{B,j}$ relate to the same category j. Show that the full log-likelihood can be written $\sum_{j=1}^k (N_{A,j}\log p_{A,j} + N_{B,j}\log p_{B,j}) + c_{A,B}$, with constant $c_{A,B} = \log n_A + \log n_B - \sum_{j=1}^k \{\log(N_{A,j}!) + \log(N_{B,j}!)\}$.

(c) Consider testing the null hypothesis that $p_A = p_B$, against the alternative that these two vectors are not equal. Show that maximising the log-likelihood under the null and

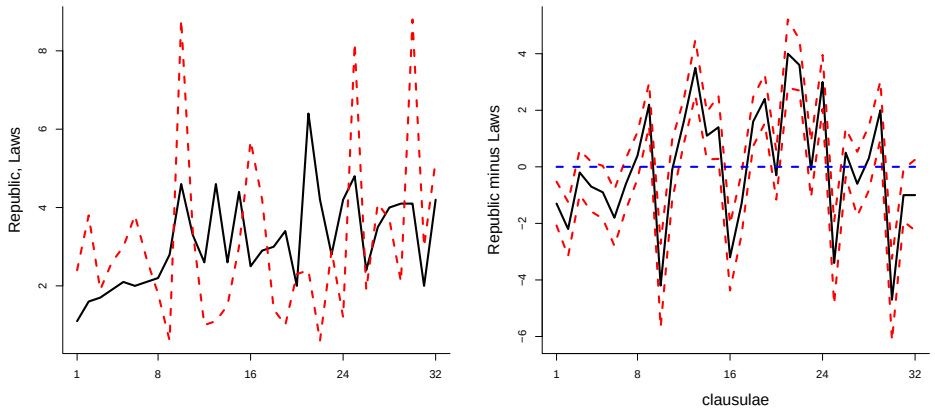


Figure ii.14: *Left panel: Republic (black, full curve), Laws (red, dashed), frequencies* $\widehat{p}_{A,j}$ *and* $\widehat{p}_{B,j}$ *for the* $2^5 = 32$ *clausulae, in percent. Right panel: Republic minus Laws: frequency differences* $\widehat{p}_{A,j} - \widehat{p}_{B,j}$*, in percent, with 99 percent confidence intervals.*

under the wider alternative leads to

$$\ell_{\max,0} = (n_A + n_B) \sum_{j=1}^{k} \widehat{p}_j \log \widehat{p}_j + c_{A,B},$$

$$\ell_{\max,\text{wide}} = n_A \sum_{j=1}^{k} \widehat{p}_{A,j} \log \widehat{p}_{A,j} + n_B \sum_{j=1}^{k} \widehat{p}_{B,j} \log \widehat{p}_{B,j} + c_{A,B},$$

with $\widehat{p}_j = (N_{A,j} + N_{B,j})/(n_A + n_B) = (n_A \widehat{p}_A + n_N \widehat{p}_B)/(n_A + n_B)$. Show from the Wilks theorems that

$$D = 2 \sum_{j=1}^{k} \{ n_A \widehat{p}_{A,j} \log \widehat{p}_{A,j} + n_B \widehat{p}_{B,j} \log \widehat{p}_{B,j} - (n_A + n_B) \widehat{p}_j \log \widehat{p}_j \}$$

is approximately a χ^2_{k-1}, under the $p_A = p_B$ hypothesis assumption. Compute D for the case of comparing Platon's Republic (A) with Laws (B), and establish that these are firmly different.

(d) Construct a version of Figure ii.14. It plots the frequencies $\widehat{p}_{A,j}$ and $\widehat{p}_{B,j}$ (left panel), in percent, and then the differences $\widehat{d}_j = \widehat{p}_{A,j} - \widehat{p}_{B,j}$ (right panel), also in percent, with 99 percent confidence intervals; these are pretty narrow, with precise frequencies, due to the high sample sizes n_A, n_B. Construct also a table with these differences, their estimated standard deviations sd_j, and Wald ratios $\widehat{d}_j/\text{sd}_j$ (the lines below are from such a table). Verify that the three cases with the strongest $p_{A,j} > p_{B,j}$ behaviour are '10010' (or LSSLS), '01001' (or SLSSL), '10011' (or LSSLL); and similarly that the three cases exhibiting the strongest $p_{A,j} < p_{B,j}$ are '10001' (or LSSSL), '00011' (or SSSLL), '11101' (or LLLSL).

```
# cases where pA is much bigger than pB:
  9  1 0 0 1 0       2.8    0.6    2.2  0.296  7.424
 13  0 1 0 0 1       4.6    1.1    3.5  0.381  9.194
```

```
 22  1 0 0 1 1       4.2    0.6    3.6  0.350 10.296
# cases where pA is much smaller than pB:
 10  1 0 0 0 1       4.6    8.8   -4.2  0.573 -7.330
 16  0 0 0 1 1       2.5    5.7   -3.2  0.455 -7.040
 30  1 1 1 0 1       4.1    8.8   -4.7  0.562 -8.358
```

(e) In the Platonic context, consider one of the five dialogues, giving rise to a multinomial $N = (N_1, \ldots, N_k)$, with total count n and probability vector $p = (p_1, \ldots, p_k)$. We attempt to somehow place this p on a probabilistic bridge from p_A to p_B. For nonnegative weights $w_1, \ldots, w_k$ given to the 32 rhythmic clausulae, consider

$$d_w(p, p_A) = \sum_{j=1}^{k} w_j\{p_j \log(p_j/p_{A,j}) - (p_j - p_{A,j})\},$$
$$d_w(p, p_B) = \sum_{j=1}^{k} w_j\{p_j \log(p_j/p_{B,j}) - (p_j - p_{B,j})\},$$

seen as weighted distances, from p to p_A, and from p to p_B. Show that $r\log(r/r_A) - (r - r_A)$ is always nonnegative, for r and r_A in $(0, 1)$, implying that the two distances are indeed nonnegative. Taking all w_j equal to 1 corresponds to placing equal importance weight to all k cases; show that one then has the simplified expressions $d(p, p_A) = \sum_{j=1}^{k} p_j \log(p_j/p_{A,j})$ and $d(p, p_B) = \sum_{j=1}^{k} p_j \log(p_j/p_{B,j})$ (where individual terms might be negative, though their sums are guaranteed to be nonnegative, by the above). Show that these are Kullback–Leibler distances, as per Ex. 5.6.

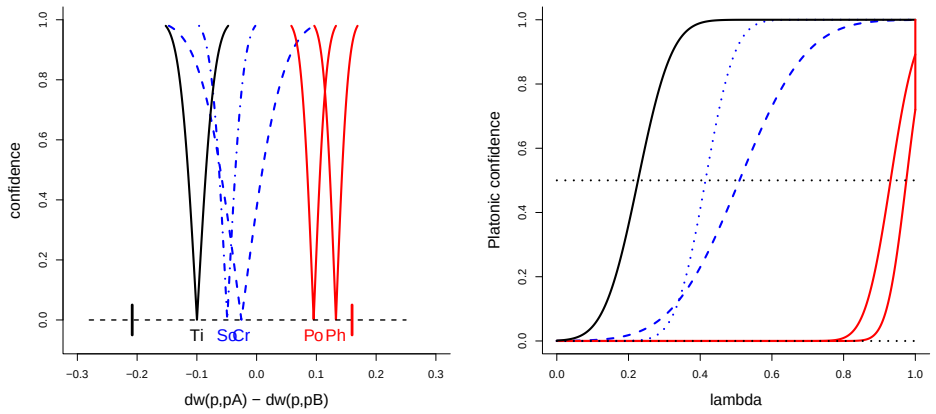


Figure ii.15: *Left panel: confidence curves for* γ *of (ii.2), in the order of Timaeus, Sophist, Critias, Politicus, Philebus, from left position of A, Republic, to B, Laws. Right panel: confidence distributions for the* λ *of (ii.3), appearing in the precise same order. Here we have used* $w_j = 1$ *for the 10 clausulae with the most prominent changes, and* $w_j = 0$ *for the rest.*

(f) The idea is now to estimate

$$\gamma = d_w(p, p_A) - d_w(p, p_B) = \sum_{j=1}^{k} w_j\Big\{p_j \log \frac{p_{B,j}}{p_{A,j}} - (p_{B,j} - p_{A,j})\Big\} \qquad \text{(ii.2)}$$

for each of the five works between A and B. It should be negative for those composed just after A and positive for those written close to B. With $\widehat{p}_j = N_j/n$ the proportion of clausula j, for the work considered, show that

$$\widehat{\gamma} = d_w(\widehat{p}, p_A) - d_w(\widehat{p}, p_B) = \sum_{j=1}^{k} w_j \Big\{ \widehat{p}_j \log \frac{p_{B,j}}{p_{A,j}} - (p_{B,j} - p_{A,j}) \Big\}$$

is unbiased, approximately normal, with variance $\tau^2 = (1/n)\{\sum_{j=1}^k c_j^2 p_j - (\sum_{j=1}^k c_j p_j)^2\}$, where $c_j = w_j \log(p_{B,j}/p_{A,j})$; see also Story vii.1. Estimate the γ, in this fashion, along with estimated variances, and construct a version of Figure ii.15 (left panel), with confidence curves $\mathrm{cc}(\gamma_j)$. For this use $w_j = 1$ for the 10 clausulae with the most prominent changes, and $w_j = 0$ for the rest. Verify that they occur in the order of Timaeus, Sophist, Critias, Politicus, Philebus.

(g) One model for Plato's envisaged transition in style, from A, Republic, to B, Laws, considering p_A and p_B as given, or estimated with good precision, takes $p_j = (1-\lambda)p_{A,j} + \lambda p_{B,j}$, with $\lambda \in [0,1]$ indicating the Platonic path from A to B. We should now fit this model, for each of the five intermediate dialogues. A method for achieving this is to minimise the empirical version of the weighted KL distance

$$d_w(p, p_\lambda) = \sum_{j=1}^{k} w_j [p_j \log\{p_j/p_j(\lambda)\} - \{p_j - p_j(\lambda)\}]. \tag{ii.3}$$

Show that this leads to maximising the weighted log-likelihood function

$$\ell_w(\lambda) = n \sum_{j=1}^{k} w_j \{\widehat{p}_j \log p_j(\lambda) - p_j(\lambda)\}.$$

(xx more. approximate standard error, as with mwl estimators Ch5. Make a version of Figure ii.15 (right panel), with confidence distributions $C_j(\lambda)$, and verify that their order is the same as that found above. xx)

(h) Another model for this transition from A to B takes $p_j = p_{A,j}^{1-\kappa} p_{B,j}^{\kappa}/R(\kappa)$, with $R(\kappa) = \sum_{j'=1}^k p_{A,j'}^{1-\kappa} p_{B,j'}^{\kappa}$, with one appropriate κ for each of the works composed between A and B. On this bridge, endpoints p_A and p_B correspond to $\kappa = 0$ and $\kappa = 1$. Show that the log-likelihood function becomes

$$\ell_n(\kappa) = \sum_{j=1}^{k} N_j \{\kappa(\log p_{B,j} - \log p_{A,j}) - \log R(\kappa)\} = \kappa U_n - n \log R(\kappa),$$

where $U_n = \sum_{j=1}^k N_j(\log p_{B,j} - \log p_{A,j})$. (xx more. fit κ for each of the five. this model, even though it's not perfect, generates the score from the $d(p, p_A)$ and $d(p, p_B)$ analyses. xx)

(i) (xx somewhat briefly. making models for $p = (p_1, \ldots, p_k)$ with fewer parameters than $k - 1 = 31$. can attempt Markov models. it works but is laborious. xx)

Story ii.9 *Presidents of the First Republic*
Data: `frenchA-data, frenchB-data (to be organised)`
Code: `(xx to come com* of nilswork23 xx)`

LIBERTÉ, EGALITÉ, FRATERNITÉ is the famous motto of the French Republic. At the time of its origination, during the French revolution, the motto often came with a darker addition: – OU LA MORT. It is this aspect we will examine in this story. Revolutionary days are a time for experimentation, and a number of new systems were tested out during the First Republic (1792–1804), for example a new calendar system, several new state religions (among others the Cult of Reason and the Cult of the Supreme Being), and, naturally, new political systems. One such system was the National Convention (of 749 elected members), whose president could then be considered France's legitimate Head of State in this period. The presidents were elected for 14 day terms, and this gives us an interesting dataset of $n = 73$ different French presidents for the full National Convention period (September 1792 to November 1795). The dataset comprises id (identity label, 1 to 73); birth (date); death (date); presistart (start of presidency); presiend (end of presidency); v, indicator for having experienced a violent death; giro, indicator for belonging to the Gironde party; vip, a proxy for fame, taken here to be the number of languages in which there is a wikipedia page for the president in question.

In the following, we take an interest the time t_i it took president i to die, from the end of his presidency, for the 73 presidents. To analyse such data we may operate with two viewpoints, so to speak. Perspective A is that 'a life is a life and you die when you die', and since we know the t_i for each president there is no statistical censoring. Perspective B is different, and holds that the life of a guillotined man 'should' have been longer, so we then consider the imagined what-if lifetime t_i^* to have been unpleasantly censored. In yet other words, with perspective B, the survival data are of the form (t_i, δ_i), with $\delta_i = 0$ for those having met a violent death and $\delta_i = 1$ for those lucky enough not to have been killed.

(a) With viewpoint A, just described, compute and display the Nelson–Aalen estimator for the implied cumulative hazard function. Use a couple of lines to explain what the estimator is telling us here. What is the estimated median time, say $\widehat{m}_A$, from end of presidency to death?

(b) We now leave viewpoint A and for the rest of this exercise operate under perspective B. Compute and display the Nelson–Aalen estimator $\widehat{A}_B$ for the underlying cumulative hazard rate, along with an approximate 90% pointwise confidence band $\widehat{A}_B(t) \pm 1.645\,\widehat{\sigma}_B(t)$, say (xx pointer here to formula and properties from Ch10 xx). Compute and display also the Kaplan–Meier estimator, and read off the estimated median time from end of presidency to death, say $\widehat{m}_B$, from this plot. Comment on the two median estimates $\widehat{m}_A$ and $\widehat{m}_B$.

(c) To test whether $A_B(t_0) = a_0$, for some given time t_0 and level a_0, explain that a test with approximate level 0.10 is to accept provided $|\widehat{A}_B(t_0) - a_0| \le 1.645\,\widehat{\sigma}_B(t_0)$. Use this to construct a nonparametric approximate 90 percent confidence interval for the median m_B, collecting together all values t_0 for which $|\widehat{A}_B(t_0) - \log 2| \le 1.645\,\widehat{\sigma}_B(t_0)$. (xx pointer here to nonparametric ci for quantiles for iid data in Ch3. cross-check here if this is inside Ch10 or not. thus no need to get into hazard rate estimation etc. xx)

(d) (xx pointer to general loglik expression for survival data, inside Ch10. xx) We shall now fit a parametric Weibull model to the survival data (t_i, δ_i), with cumulative hazard rate function of the form $A(t) = (t/a)^b$, i.e. hazard rate $\alpha(s) = bs^{b-1}/a^b$ for $s > 0$. Show

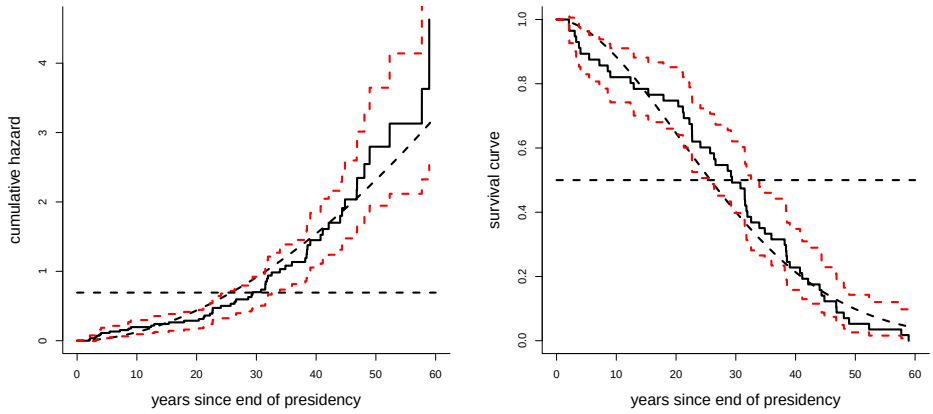


Figure ii.16: *Left panel: nonparametric Nelson–Aalen estimate for the cumulative hazard (ragged curve), for the time to death since end of presidency, along with 90 percent confidence band. The smooth curve is the estimated Weibull model based cumulative hazard. The median time for the distribution corresponds to crossing* $\log 2$. *Right panel: corresponding figure with the Kaplan–Meier survival curve (ragged), along with 90 percent confidence bands, and the smooth Weibull based survival curve.*

that the log-likelihood function may be expressed as

$$\ell_n(a,b) = \sum_{\delta_{B,i}=1} \{\log b + (b-1)\log t_i - b\log a\} - \sum_{i=1}^{n} (t_i/a)^b.$$

Fit this two-parameter model by numerically finding the maximum likelihood estimates (c,b) in question (you should find $(31.513, 1.821)$); you may e.g. use the R algorithm `nlm` to minimise the negative log-likelihood function. Compute also approximate standard errors (estimated standard deviations) for the two parameter estimates. Construct a version of Figure ii.16, and comment on what might be learned from this.

(e) Using the Weibull model, estimate the probability p_B that a newly retired president will live for at least twenty more years, supposing and praying he is not beheaded. Also give a 90% confidence interval for this p_B. How different is your estimate of p_A, the corresponding probability under perspective A?

(f) So 18 out of the 73 presidents were executed. How long would their lives have been, in the what-if nation of France, sans guilliotine? With T a lifetime known to have come to at least t_0, show in general terms that $\Pr(T \geq t \,|\, T \geq t_0) = \exp[-\{A(t) - A(t_0)\}]$ for $t \geq t_0$. with A the cumulative hazard function. Show then that a person who has reached age t_0 has median lifetime $t^* = A^{-1}(A(t_0) + \log 2)$. If your lifetime is governed by the Weibull model, and your age today is t_0, deduce that your median lifetime is $t^* = a\{(t_0/a)^b + \log 2\}^{1/b} = (t_0^b + a^b \log 2)^{1/b}$. For presidents 1, 2, 3, 8, 9, 12, 15, 16, 21, 22, 25, 28, 30, 34, 35, 41, 44, 58, estimate their median lifetimes, had they not been executed, (i) using the Weibull model, (ii) nonparametrically. For methods (i) and (ii), supply also approximate 90 percent confidence intervals for these median what-if lifetimes.

(g) Now push the covariate `x = giro` into the analysis. There are several options here, corresponding to different Weibull regression setups. Try out $A_i(t) = (t/a_i)^{b_i}$, with the choices (i) $a_i = a\exp(\beta_1 x_i)$ and b constant; (ii) a constant but $b_i = b\exp(\beta_2 x_i)$; (iii) $a_i = a\exp(\beta_1 x_i)$ and $b_i = b\exp(\beta_2 x_i)$. Give standard errors for the regression coefficients and test $\beta_1 = 0$ and $\beta_2 = 0$. Summarise your findings.

(h) (xx one more thing. could ask to have the analyses redone with one or two other models, comparing to the Weibull. point to a bit of common themes for story GoT-WoR, where we drive Gompertz instead. xx)

Story ii.10 *Dangerous job assignment: Roman Emperor*
Data: `roman-emperors-data`
Code: `(xx to come com10* nilswork23 xx)`
How often do you think about the Roman Empire? Pontifex maximus, princeps senatus, augustus, basieleus – whatever the title used, being the official ruler of the Senatus Populusque Romanus was not an easy job. Lists can be compiled pertaining to the different Roman emperors, their reigns, how and when they were elected, and how and when they died. A source for such information, and more, is the encyclopaedic website *De Imperatoribus Romanis*. Here we use the dataset given in Saleh (2019, Appendix), for the 69 legitimate emperors, from Augustus (reign 31 B.C. to 14 A.D.) to Theodosius I (reign 379 to 395 A.D.), with start-date, end-date, and an indicator v for whether he met a violent death (by murder, suicide, or during combat with a foreign enemy) or not.

There is a certain existential and statistical resemblance here to the drama of the 73 presidents of the French Republic 1792–1795, studied in Story ii.9. To examine, model, analyse the violent-death-or-not data for the Roman emperors below we add more components, however, and ask other questions. We formulate the time T to death, measured from the start-of-reign date, as $\min(T_0, T_1)$, where T_0 is time to non-violent death and T_1 time to violent death, whatever comes first. This is an instance of the *competing risks* setup from survival analysis. Methods of Ch. 10 may be used to model and analyse the distributions of T_0 and T_1 separately, where $v = 1$ means censoring for T_0 and $v = 0$ means censoring for T_1.

(a) Access and organise the data to have a file of (t_i, x_i, v_i), with t_i the length of the reign, starting in year x_i, and with $v_i = 1$ for violent death and $v_i = 0$ for non-violent death. Compute and display Nelson–Aalen estimates $\widehat{A}_0$ and $\widehat{A}_1$ for the cumulative hazard rates, and Kaplan–Meier estimates $\widehat{S}_0$ and $\widehat{S}_1$ for the survival curves; see the left panel of Figure ii.17. Read off median time-to-death estimates for the two distributions; you should find $\widehat{m}_0 = 19.0$ and $\widehat{m}_1 = 7.0$. How do you interpret these median estimates?

(b) With the apparatus of survival analysis and censored data we can model the distributions involved, associated with T_0 and T_1, even when only one of T_0, T_1 is actually observed. To attempt parametric modelling, start out fitting the survival data to Gamma distributions, say $\mathrm{Gam}(a_0, b_0)$ and $\mathrm{Gam}(a_1, b_1)$. Plot the estimated gamma survival curves along with the Kaplan–Meier curves, and form an opinion of how well they fit.

(c) To check if the core mechanisms might have been changing, for T_0 or T_1, over the four hundred years, fit gamma regression survival models, of the type

$$T_{0,i} \sim \mathrm{Gam}(a_0\exp(\gamma_0 x_i), b_0) \quad \text{and} \quad T_{1,i} \sim \mathrm{Gam}(a_1\exp(\gamma_1 x_i), b_1).$$

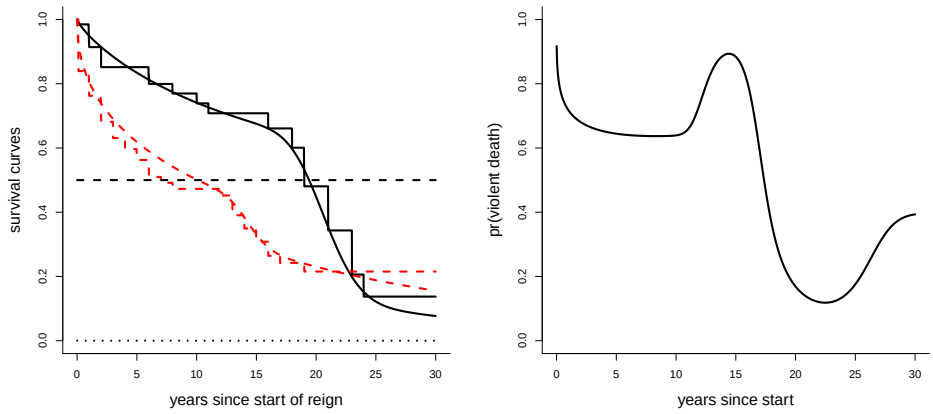


Figure ii.17: *Left panel: Kaplan–Meier survival curves, with gamma mixture models. The higher survival is for non-violent death and the lower survival, with more rapid-coming death, is with violence. Estimated medians* $\widehat{m}_0 = 19.0$ *and* $\widehat{m}_1 = 7.0$ *correspond to crossing the 0.50 line. Right panel: probability of a death having been a violent one, as a function of time since start of reign.*

Estimate γ_0 and γ_1, and decide if these are significantly different from zero or not. Construct also versions of Figure ii.18, for the T_0 and T_1 data, with model checking monitoring plots $Z_n(v) = \sqrt{n}\{G_n(v) - v\}$, with the empirical c.d.f.s for (i) $\widehat{v}_i = F(t_i, \widehat{a}, \widehat{b})$, for the plain gamma model, and for (ii) $\widehat{v}_i = F(t_i, x_i, \widehat{a}, \widehat{b}, \widehat{\gamma})$, with the gamma regression structure. (xx pointer to Ch9 things. we learn that the simple gamma model without the x_i covariate is not good enough for the T_1 data. xx)

(d) We now add on one more statistical assumption, namely that T_0 and T_1 are independent. Discuss briefly what this means in the present context. Show that $T = \min(T_0, T_1)$ has survival function $S(t) = S_0(t)S_1(t)$, cumulative hazard rate $A(t) = A_0(t) + A_1(t)$, and hazard rate $\alpha(t) = \alpha_0(t) + \alpha_1(t)$, in terms of the individual hazard rates α_0 and α_1.

(e) Show also in general terms that with parametric models $f_0(t, \theta_0)$ and $f_1(t, \theta_1)$ for the T_0 and T_1 distributions, and data (T_i, δ_i) with δ_i being 0, 1 for the $T_{i,0}$ smallest or $T_{i,1}$ smallest, then the likelihood function can be expressed as (xx check this with care xx)

$$L(\theta_0, \theta_1) = \prod_{\delta_i=0} f_0(t_i, \theta_0) S_1(t_i, \theta_1) \prod_{\delta_i=1} f_1(t_i, \theta_1) S_0(t_i, \theta_0) = L_0(\theta_0) L_1(\theta_1).$$

This invites modelling and analysing the two distributions separately.

(f) Then fit gamma distribution mixtures to T_0 and T_1,

$$\begin{aligned} f_0 &= p_0 \,\mathrm{Gam}(a_{0,1}, b_{0,1}) + (1 - p_0)\,\mathrm{Gam}(a_{0,2}, b_{0,2}), \\ f_1 &= p_1 \,\mathrm{Gam}(a_{1,1}, b_{1,1}) + (1 - p_1)\,\mathrm{Gam}(a_{1,2}, b_{1,2}), \end{aligned}$$

by maximising the two log-likelihood functions numerically. Carry out Wilks type testing to assess the increase in maximised log-likelihood, and comment. Construct a version of Figure ii.17, left panel.

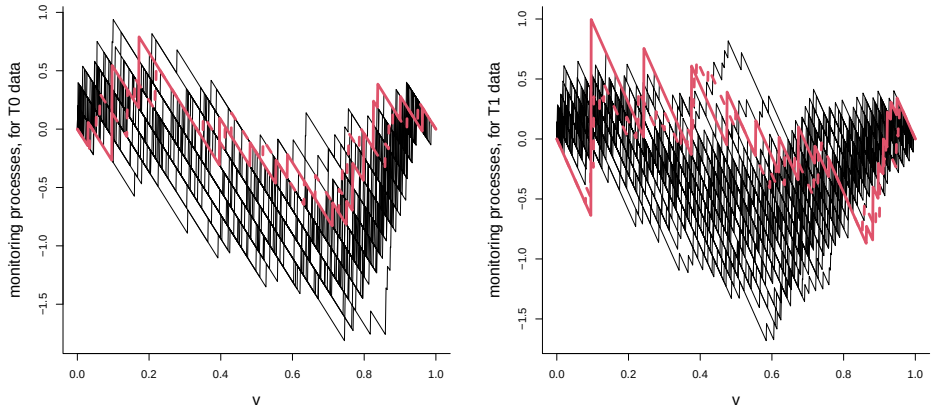


Figure ii.18: *The monitoring plots* $Z_n = \sqrt{n}\{G_n(v) - v\}$, *with* G_n *the empirical c.d.f. for (i)* $\widehat{v}_i = F(t_i, \widehat{a}, \widehat{b})$ *for the plain gamma model (fat smooth curve) and (ii)* $\widehat{v}_i = F(t_i, x_i, \widehat{a}, \widehat{b}, \widehat{\gamma})$ *(fat dashed curve) for the gamma regression model, along with 50 simulated realisations of these from the plain gammal model, constructed at the ML position* $(\widehat{a}, \widehat{b})$. *Left and right panels for the* T_0 *and* T_1 *data.*

(g) Suppose you learn that an emperor has died, at time t after start-of-reign. What is the probability that his death was a violent one? Compute this probability, as a function of t, and construct a version of Figure ii.17, right panel.

Story ii.11 *Lifelengths in Roman Era Egypt, 2100 years ago*
Data: `egypt-data`
Code: `(xx to come com11* nilswork24 more xx)`
Intriguingly, archeologists have been able to learn the ages at death of 141 mummified individuals living in Roman Era Egypt, some 2100 years ago, see Spiegelberg (1901). These lifelengths, varying from 1 to 96 years, for 82 men and 59 women, were discussed and analysed by Karl Pearson in the very first volume of *Biometrika*, see Pearson (1902). We treat them here as a random sample of lifelengths from the upper social class of Roman Era Egypt, during a period of relatove societal stability; more details are in Claeskens and Hjort (2008b, Ch. 2). (xx with data and details in Ch. B.2.B. when polishing, do the right pointing to Ch10 and to ML machinery of Ch5. xx)

Despite Pearson's not unreasonable comment that "in dealing with [these data] I have not ventured to separate the men from the women mortality, the numbers are far too insignificant" we shall work with parametric modelling of the men's and women's survival functions and hazard rates, and in that process illustrate the main practical uses of maximum likelihood machineries, both for model parameters and for natural parameter functions of these, and for model comparison and model selection.

(a) Go through as many as eight candidate models for these data, given below. For each model, estimate the parameters via maximum likelihood, along with estimated standard deviations for these. Here we use the general versatile machinery partly showcased and summed up in Stories iv.7 and vii.4, involving programming the log-likelihood functions, finding their optima, and inverting the Fisher information matrices. Part of the learning experience here is that handling rather different parametric models does not take

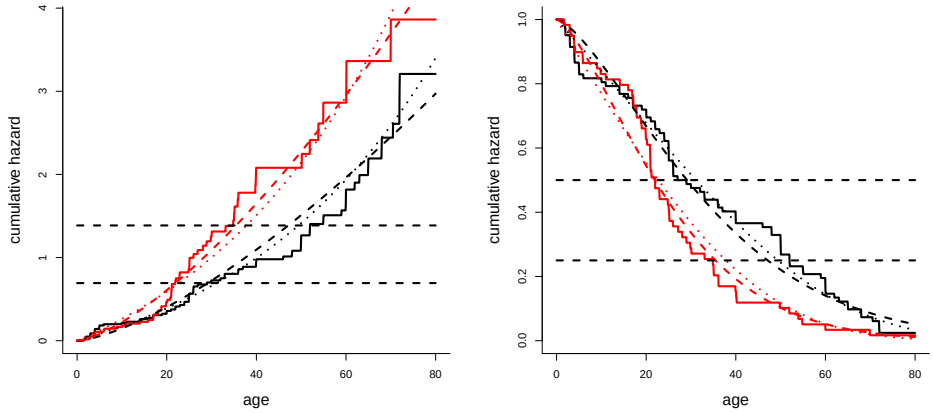


Figure ii.19: *Lifelengths in Roman Era Egypt, a century B.C., with men tending to have longer lives than women. Left panel: nonparametric Nelson–Aalen plots for the cumulative hazards, for men (lower curves) and for women (upper curves), along with parametric fits from Models 2b (Weibull with equal 2nd parameter) and 3b (Gompertz with equal 2nd parameter) as per Ex. ii.11. Estimated 50 and 75 percent survival times can be read off from the two horizontal lines at* $\log 2$ *and* $\log 4$. *Right panel: survival curves for the men (upper curves) and women (lower curves), nonparametric (rugged) along with parametric fits. Estimated 0.50 and 0.75 percent survival times are read off from where the curves cross the 0.50 and 0.25 lines.*

many extra forces, but may involve relatively small changes from script to script. (i) Use gamma distributions $\text{Gam}(a_m, b_m)$ and $\text{Gam}(a_w, b_w)$ for the men and the women, with parametrisation as in Ex. 1.9. Then use gamma distributions again, but take a common b for the shape parameter. (ii) Then use Weibull distributions (a_m, b_m) and (a_w, b_w), with parametrisation as in Ex. 1.58. Similarly, use a common shape parameter b for the two groups, but separate a_m and a_w. (iii) Then use Gompertz distributions with parameters (a_m, b_m) and (a_w, b_w), with parametrisation as in Ex. 1.59. Again allow the variation taking a common shape parameter b but different a_m, a_w for the two Gompertz distributions. (iv) Throw in also the log-normal distributions, first with four free parameters (ξ_m, σ_m), (ξ_w, σ_w), then with a common σ for the log-normals. For each of these $4 + 4$ models, graph the estimated cumulative hazard functions $A(t, \widehat{\theta})$ for men and women, plotted alongside the nonparametric Nelson–Aalen curves, and also the estimated survival curves $S(t, \widehat{\theta})$ for men and women, alongside the nonparametric Kaplan–Meier curves. In other words, construct versions of Figure ii.19, left and right panels.

(b) After having fitted all the candidate models, and computed the log-likelihood maxima in question, it is a small extra step to count parameters and compute the AIC scores. Do this, organising your results into a table with the three first columns here, with 'dim' denoting the number of parameters in the model. Conclude that model 3B is the best (so far), the Gompertz model with parameters (a_m, b) and (a_w, b), as judged by the AIC; see the AIC ranks in column 4. Incidentally, show that the log-normal models are decidely worse. We include them here for the sake of exercising the general maximum likelihood machinery, and since we could not have known a priori which models are good and which

are not.

```
       dim logLmax   aic     rank men    women  delta sd     low   up
model 1A 4 -612.064 -1232.129  5  26.655 21.877 4.778 3.371 -0.766 10.323 gamma
model 1B 3 -614.922 -1235.844  6  26.055 22.725 3.330 3.439 -2.327  8.986
model 2A 4 -609.954 -1227.909  4  28.262 22.728 5.534 3.502 -0.226 11.294 weib
model 2B 3 -610.387 -1226.774  3  29.120 21.910 7.209 2.997  2.279 12.139
model 3A 4 -608.388 -1224.776  2  31.783 22.140 9.644 4.246  2.659 16.629 gomp
model 3B 3 -608.520 -1223.040  1  31.124 22.723 8.401 3.447  2.730 14.072
model 4A 4 -627.397 -1262.794  7  23.237 19.958 3.279 3.447 -2.391  8.949 logN
model 4B 3 -629.511 -1265.023  8  23.237 19.958 3.279 3.521 -2.514  9.071
```

(c) For the three best of the fitted models, compute and graph the estimated densities $\widehat{f}(t)$, survival curves $\widehat{S}(t)$, and cumulative hazard rates $\widehat{A}(t)$. Complement these with the nonparametric Nelson–Aalen estimators. Present also the estimated hazard rates $\widehat{\alpha}(t)$. Construct a version of Figure ii.19. Explain how estimated median survival time in Roman era Egypt can be read off from the horizontal $\log 2$ line, and similarly the estimated 75 percent quantile survival time via the $\log 4$ line.

(d) It appears clear that in Roman era Egypt, men tended to have longer lives than women. The direct nonparametric median lifetime estimates are 28 for men and 22 for women. For each of the eight candidate models, compute the implied median-life difference estimate, i.e. of $\delta = F_m^{-1}(0.50) - F_w^{-1}(0.50)$. Also use the maximum likelihood theory as partly summarised in Story vii.4, specifically the use of the delta method for any smooth function of the model parameters, to compute the approximate standard deviation for these $\widehat{\delta}$ estimates, and give 90 percent confidence intervals. These estimates, with lower and upper confidence points, are given in the table above. Your code should be flexible enough to carry out similar analyses for e.g. the upper quartile difference $F_m^{-1}(0.75) - F_w^{-1}(0.75)$, a parameter of high interest for the five million Egyptians two thousand years ago. Attempt to pinpoint where the men and women of Roman Era Egypt started having different lifelength expectancies.

(e) (xx find an easy reference to the fact that a high proportion of women died in childbirth, in many socities. xx) The models worked through above are generic in character and do not take on board why or in which ways the lives of men and women might have been dfferent in old Egypt. The flexibility and versality of the maximum likelihood machinery should inspire building other models. Consider random lifetimes

$$T_m = \min\{t \geq 0\colon Z_m(t) \geq c\}, \quad T_w = \min\{t \geq 0\colon Z_w(t) \geq c\},$$

defined via cumulative risk processes Z_m and Z_w for men and women; when these cross threshold c, the individual dies. A natural class of such processes, amenable to further survival analysis for their threshold crossing times, is that of independent increment gamma processes, see Cunen and Hjort (2024). For the present purposes we take $Z_m(t)$ having mean function at whereas $Z_w(t)$ has mean function $at + d\,\mathrm{ex}(t)$, with an extra risk function $\mathrm{ex}(t)$ here taken to be the c.d.f. of a uniform distribution on $[15, 40]$. Show that this leads to survival functions $S_m(t) = G(c, at, 1)$ and $S_w(t) = G(c, at + d\,\mathrm{ex}(t), 1)$ for men and women, where $G(\cdot, u, 1)$ is the c.d.f. for $\mathrm{Gam}(u, 1)$. Show that the log-likelihood function becomes

$$\ell(a, c, d) = \sum_{i=1}^{n_m} \log f_m(t_{m,i}) + \sum_{i=1}^{n_w} \log f_w(t_{w,i}),$$

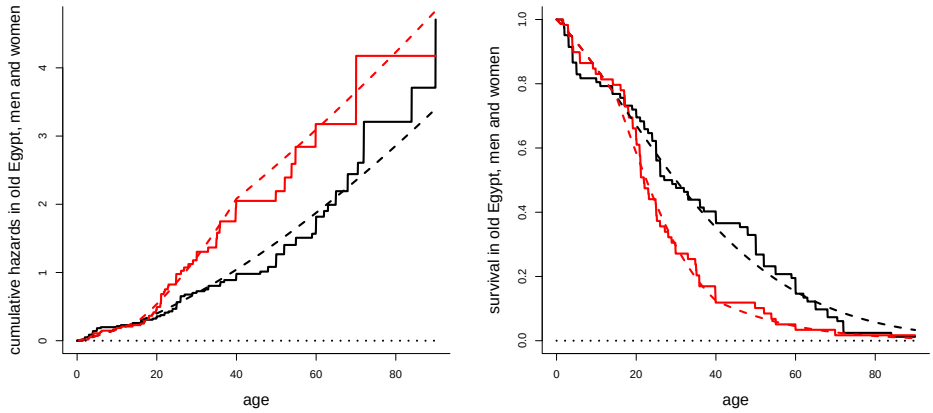


Figure ii.20: *(xx text to be coordinated and polished. xx) Left panel: Nelson–Aalen cumulative hazards for men and women of Roman Era Egypt, along with those fitted to the gamma process threshold crossing model; these are better than the 4 + 4 models worked with initially. Right panel: associated survival curves, nonparametric and parametric.*

with the n_m and n_w lifelengths for men and women, and with densities $f_m(t) = -S'_m(t)$ and $f_w(t) = -S'_w(t)$ implied by the survival functions.

(f) Now programme and optimise the log-likelihood. You should find $(\widehat{a}, \widehat{c}, \widehat{d}) = (0.033, 0.687, 0.810)$, and with a much higher log-likelihood maximum -604.368 than for the eight models worked with above. Show also that this leads to an AIC score very clearly better than for the competitors. Display nonparametric and parametrically fitted cumulative hazard rates and survival curvess, as in Figure ii.20, left and right panels. The gamma process models provide much better fits than for models portrayed in Figure ii.19.

(g) To illustrate how the gamma process models work, simulate e.g. 25 Z_m processes, with mean function at, and Z_w processes, with mean function $at + d\,\mathrm{ex}(t)$, using the estimated $(\widehat{a}, \widehat{c}, \widehat{d})$. Death occurs when the process reaches c. More men than women survive the age of forty. Construct a version of Figure ii.21 (male processes in left panel, female in right panel). (xx put in somewhere: with our gamma process model, women and men have the same longer-time survival chances after the age of forty. xx)

Story ii.12 *Bach, Reger, organ fugues, and Wohltemperierte I und II*
Data: `fuga-data, bach-wtk-data (to be organised) xx`
Code: `to come com1* of kioskvelterwork22 xx`
A fugue, whether for a piano, an organ, a choir, or an ensemble of instruments, starts with the principal fugue theme itself, before it is imitated and varied, perhaps in complex ways, in other voices; typical Bach fugues have from three to five voices. Rydén (2020) has studied such fugue themes from the organ works of Bach and other composers. He has accurately defined certain features, for quantitative analysis and comparisons. These can be identified and counted for each given fugue theme. In brief, these are

x_1, the length, number of notes, range 7 to 64;

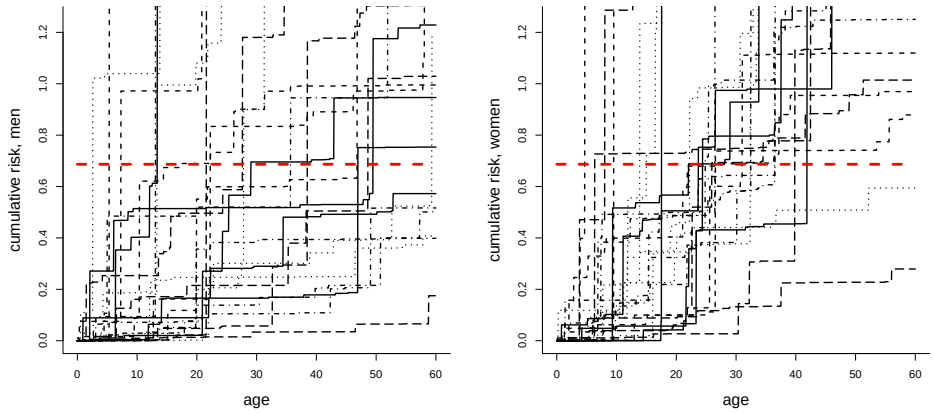


Figure ii.21: *(xx text to be coordinated and polished. xx) 25 simulated gamma processes, for mean functions at for the men (left panel) and* $at+d\,\mathrm{ex}(t)$ *for the women (right panel); an individual dies when his or her process crosses the threshold* $c = 0.687$*, the horizontal line.*

x_2, the compass, range (in semitones), range 5 to 20;

x_3, the number of unique notes, range 4 to 12;

x_4, the initial interval (in semitones), range 0 to 12;

x_5, the number of unique intervals between successive notes, range 2 to 11;

x_6, the max interval (in semitones), range 2 to 12.

Further aspects of the data are briefly described in (xx data overview 2.B xx). (xx could mention Prout, 1891, Tovey, 1924. xx)

The musical here is to statistically describe and compare the fugues of J.S. Bach (1685–1750) and Max Reger (1873–1916). Figure ii.22 (left panel) shows (x_1, x_6) for the $n_B = 47$ Bach fugues and $n_R = 45$ Reger fugues, indicating also that the distributions are not very different. (xx mention the Händel concerto gross, is it no. 7, with only a single note for the fugue theme, so $x_3 = 1$; for Bach and Reger the range is from 4 to 12, though. xx)

(a) For an initial check of the data, take the $n_B + n_R = 92$ fugues together. Go through each of the fugue features $x_1, \ldots, x_6$, and give brief statistical descriptions. Identify also pairs of features with strong correlation, if any. Construct a version of Figure ii.22, left panel, which has (x_1, x_6) for Bach and Reger; construct a similar one for (x_1, x_3).

(b) For each of the fugue features $x_1, \ldots, x_6$, compute means and standard deviations, for the 47 Bach fugues and 45 Reger fugues. Then for each feature, test equality of means, say $\xi_{B,j} = \xi_{R,j}$, using t testing; see Ex. 3.11. Comment both on the use of t testing for these data and on your findings. (You should find that for feature x_3, Reger has higher mean than Bach, whereas they are more or less equal, for the other five features.)

```
    mean B mean R  sd B   sd R    kurt B kurt R
x1  25.766 21.111  16.240 11.924  -0.336  2.571
```

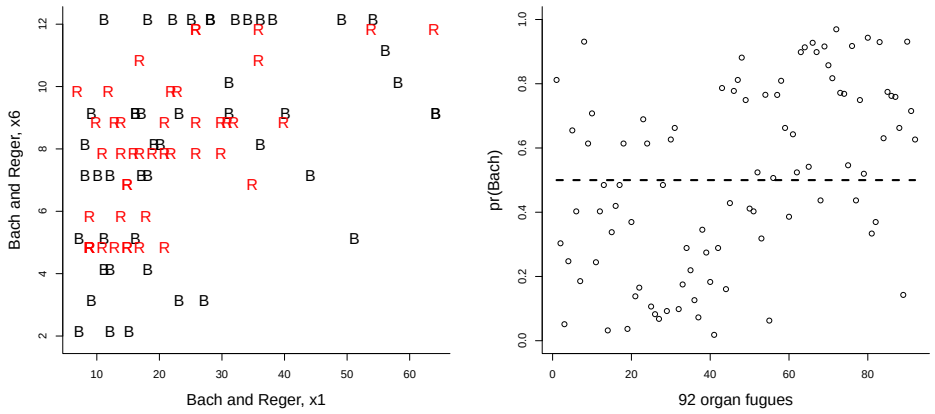


Figure ii.22: Left panel: for the chief organ fugue themes of Bach (B, 47 fugues) and Reger (R, 45 fugues), the plot gives features x_1, the length, and x_6, the max interval. Right panel: using logistic regression on the basis of fugue features x_1, x_3, the figure shows the estimated $\Pr(\text{Bach} \,|\, x_1, x_3)$, for the 92 fugues, listed with the 45 Reger ones first, the 47 Bach ones afterwards.

```
x2  11.043 11.911   3.520  2.636  -0.642 -0.821
x3   6.872  8.556   1.541  1.791   0.729 -1.055
x4   2.617  2.222   2.524  1.894   2.323 13.201
x5   6.000  5.978   2.467  2.072  -0.579 -0.901
x6   8.021  8.089   3.267  2.275  -1.114 -1.053
```

(c) Then go on to testing equality of standard deviations, say $\sigma_{B,j} = \sigma_{R,j}$. Do this first by applying a traditional F test, as from Ex. 4.39, even though the data are not normal. This should give an indication that Bach intriguingly exhibits greater variability than Reger, for features x_1, x_2, x_4, x_6, with the means being about the same. Also carry out the somewhat more elaborate testing regime, for equality of standard deviations, from Ex. 4.39, which does not rely on normal data. Does this change the previous tentative findings?

(d) The fugue features $x_1, \ldots, x_6$ devised by Rydén (2020) are meant as useful musicological descriptors, but as they concern merely the fugue theme itself, not the further compositional development, they cannot be expected to and do not pretend to discriminate between e.g. Bach and Reger to any high degree. Even amateur musicians are able to see or hear the difference between a Bach page and a Reger page, by looking through or playing the music, though it would be hard to translate such knowledge into algorithms. Leaving these musical considerations aside, we look here into the degree of discrimination afforded by the fugue theme features. This can clearly be done in several ways, but here we attempt to build a formula for the probability that the piece is by Bach, via logistic regression,

$$\Pr(\text{Bach} \,|\, x_1, \ldots, x_6) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6)}.$$

Carry out such an analysis, and check how well it works, when we tentatively sort fugues into Bach, if the probability is at least 0.50, and Reger, if the probability is less than

0.50. In this analysis, which features x_j are significantly present, according to the logistic regression? (xx nils will check whether this is an ok illustration or not: given the hope of expressing Pr(Bach) in this way, what's the uncertainty, how wide are confidence intervals? but we should have another illustration of this somewhere. xx)

(e) Search through submodels, where some but not all of the six features are being used, and check their AIC scores. You should find that including x_1, x_3, but excluding the other four, gives the best AIC score. Construct a version of Figure ii.22, right panel, which uses logistic regression for x_1, x_3. What is the apparent success rate, for the ensuing algorithm, which sorts fugues into Bach and Reger?

(f) The direct counting of how many of the 92 fugues are correctly sorted into Bach and Reger suffers from a certain bias (xx point to things in Ch. 11 xx), since the data are used both to construct a formula and to test that formula. To form a clearer picture, carry out leave-one-out cross validation (xx pointer xx), and estimate the success rate.

(g) Rydén (2020) concentrated on the organ fugues of Bach, Reger, and others, as discussed above. We recommend playing through also the 24 fugues of Wohltemperiertes Klavier I (from the Köthen period, c. 1722) and the 24 fugues of Wohltemperiertes Klavier II (from Leipzig, c. 1742). Has Bach stayed about the same, as a fugue theme composer, for the clavier? (xx might point to Hindemith 1950. xx) One of us has actually played all 24 + 24 fugues and carefully recorded a table of $x_1, \ldots, x_6$; see 2.B. Use this to construct a version of the table below, of means, standard deviations, skewness, kurtoses, for the six characteristics, for WTK I and WTK II:

	xi		sigma		skewness		kurtosis	
	I	II	I	II	I	II	I	II
x1	18.042	21.333	7.932	9.342	0.653	0.429	-0.098	-0.241
x2	11.083	11.000	2.749	2.874	-0.264	-0.158	-0.947	-0.767
x3	7.750	7.333	2.069	1.685	0.547	0.120	-0.466	-0.615
x4	2.667	2.750	1.903	1.800	1.149	0.745	-0.004	0.226
x5	5.083	5.500	1.767	1.445	0.921	-0.373	0.350	-0.833
x6	8.167	7.750	2.220	2.625	-0.201	0.316	0.107	-1.023

(h) To assess the grand hypothesis that Bach did not change much, as a fugue theme composer from 1722 to 1742, carry out tests for the hypotheses $\xi_{I,j} = \xi_{II,j}$ and $\sigma_{I,j} = \sigma_{II,j}$, for the means and standard deviations, for features $x_1, \ldots, x_6$.

(i) Then compute empirical correlations, say $r_{I,j,k}$ and $r_{II,j,k}$, for the two datasets, for $j < k$. To compare these, test equality, using the machinery of Ex. 2.53. Argue that since kurtoses values are relatively small, these simpler methods will suffice, without bringing in the somewhat heavier machinery of Ex. 2.54.

(j) Take the 48 WTK clavier fugues together, and compare these with the organ fugues. What might be notable differences?

(k) (xx briefly, other themes, other questions to briefly explore. distance between two distributions for (x, y), when these take on integer values. xx)

Story ii.13 *How many piano tuners in Oslo?*
Data: `in text below`

Code: `to come com48* nilswork21 xx)` Applications of Bayesian methods, in complicated and perhaps nonstandard cases, particularly in situations with limited data and many model parameters, involve the challenging task of *setting up the priors*. This could often involve translation of different pieces of information into probability distributions, perhaps followed by nontrivial combinations of these. This exercise is meant as a perhaps rough illustration of such themes. – The task is to guess the number N of piano tuners in Oslo, based on perhaps rough approximations and calculations, along with uncertainty assessments. (xx note, 7 june 2021, from nils piano tuner: $N = 23$ in oslo, 62 in Norway, but there are various half-timers and charlatans. xx)

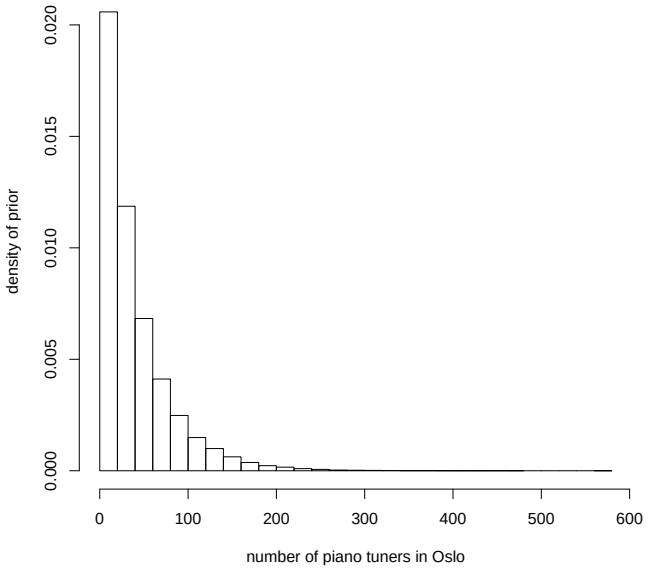


Figure ii.23: *Histogram of* 10^5 *simulated values of* $N = np_1p_2/k$ *from the prior indicated, for the number of piano tuners in Oslo.*

(a) Such a moderately rough setup takes $N = np_1p_2/k$, with n the population of Oslo; p_1 the 'piano fraction' of these, so that np_1 is roughly the number of pianos in Oslo; p_2 the fraction of these again who tune their pianos at least once a year; and k the average number of piano tuning jobs carried about in the course of a year, by a piano tuner in Oslo. – Of course there are other ways to set up such things, with other components, and a clearer chain of arguments could be forced to form more precise definitions and assumptions (one could think in terms of families and their sizes, etc., to get a better grip on p_1, etc.). Show that with independent priors for n, p_1, p_2, k, then a prior estimate for N is $\mathrm{E}\, n\, \mathrm{E}\, p_1\, \mathrm{E}\, p_2\, \mathrm{E}\, 1/k$.

(b) An attempt is as follows: let

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

One then first finds good prior means, for the four components in the setup, and then finetune their precision via σ_0 for n, c_1 for p_1, and c_2 for p_2. – Try this setup, with prior guesses $0.60 \cdot 10^6$, 0.07, 0.40, $k_0 = 500$, and then variability determined by $\sigma_0 = 0.05 \cdot 10^6$, $c_1 = 15$, $c_2 = 10$. Simulate 10^5 realisations from each of the four distributions, inspect their histograms, and present a final histogram for $N = np_1p_2/k$. Also compute the 0.10, 0.50, 0.90 points in this distribution (our simulations for this particular prior lead to 4.03, 26.2, 91.7), and compare in particular the median with the mean.

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

(d) Suppose we use the four prior components as above, but that further information is gathered regarding the two components n and p_2. For n, the population size of Oslo, along with the perhaps not fully defined suburbs and surroundings (since piano tuners from Oslo might travel to Ski and Ås), assume there is a point estimate $\widehat{n} = 0.64 \cdot 10^6$, with standard deviation $0.04 \cdot 10^6$. For p_2, the fraction of pianos tuned at least once a year, a quick Facebook check with friends and acquantainces (and their friends and acquantainces again) indicates that $y_2 = 55$ of $m_2 = 100$ piano owners tune their instruments at least once a year. Use these extra pieces of information to update the priors of n and p_2, and then revise the full distribution of $N = np_1p_2/k$. Simulate again 10^5 realisations from the N distributio, find out how the 0.10, 0.50, 0.90 quantiles have changed.

(e) (xx rounding this off. mention briefly Fermi. also the number of solar systems in a galaxy with intelligent life. illustrate the machinery of changing the final $N = np_1 \cdots p_k$ with more information on a single component. xx)

Notes and pointers

(xx notes and follow-up things for the stories in this chapter. xx)

(xx For Story ii.9, mention Cunen, and point to yet other questions to raise. xx)

(xx for Story ii.6, mention Markov (1906, 1913); Hjort and Varin (2008), language models, more. xx)

II.v

Sports

(xx WELL: lots of things to clean, as of 16-January-2025. a partial nils todo list includes: (i) set up the iiccff things in Ch7 properly, then do median cc to log-likelihoods and produce cc for x_0^*. (iii) CD for shock probability $p(a, \sigma)$. (iv) round off inner-outer, with CD for τ, variation among the d_j. (v) could throw in the handball match wathing poisson things, mostly probability, not much statistics. (vi) the golf story. xx)

Story v.1 *Bolt from heaven*
Data: `sprint100-data`
Code: `(xx to come, com94* of nilswork23 xx)`

the previous WR was Asafa Powell, 0.74, September 2007

On 31 May 2008, Usain Bolt burst upon us, with his first world record, 9.72. How surprised were we? To approach that question, along with those which followed as the Bolt From Heaven did 9.69 (August 2008) and then 9.58 (August 2009), we compare the 9.72 performance with the $n = 195$ sub-10.00 races of 2000–2007; (xx these are given in data description 2.B, check that data description says these are bona fide races; dopers pushed out of dataset xx).

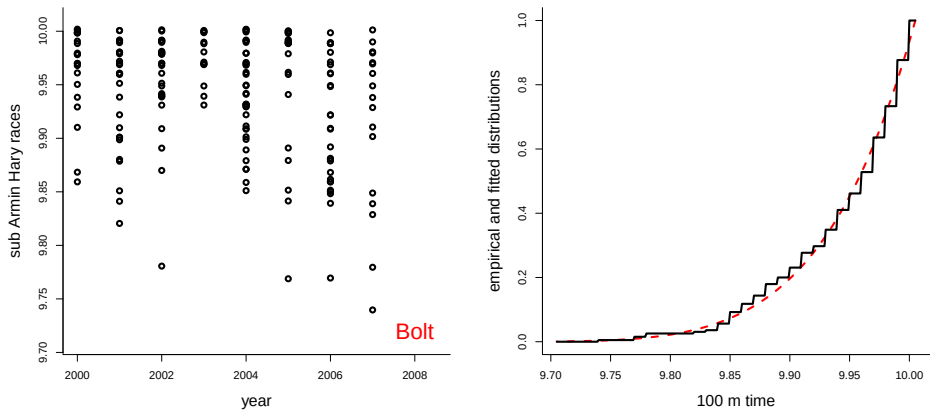


Figure v.1: *Left panel: all the 195 sub-Hary races achieved during the eight seasons 2000 to 2007, along with the new word record 9.72 ran by Bolt in May 2018. Right panel: the empirical distribution function (black, rugged) for these 195 races, along with the fitted two-parameter distribution from extreme values theory.*

(a) To readily access a body of literature on extreme values theory, see e.g. Embrechts et al. (1997), we transform these race times r_i to $y_i = 10.005 - r_i$. Such theory predicts that the y_i should follow the distribution

$$G(y, a, \sigma) = 1 - (1 - ay/\sigma)^{1/a} \quad \text{for } y > 0,$$

for parameters (a, σ). This assumes that the phenomenon studied has remained reasonably stationary during the period of observation, here that the top level of sprinting has not undergone serious changes in the period 2000–2007. Show that the log-likelihood function takes the form

$$\ell(a, \sigma) = \sum_{i=1}^{n} \{-\log \sigma + (1/a - 1) \log(1 - ay_i/\sigma)\}.$$

Fit the model, which should give ML estimates $(\widehat{a}, \widehat{\sigma}) = (0.1821, 0.0701)$, and produce a version of Figure v.1. As we see, the model works very well. Via log-likelihood profiling, produce also confidence curves $\mathrm{cc}(a)$ and $\mathrm{cc}(\sigma)$ for the two parameters; see Ex. 7.11.

(b) For a season with N top races, below the Hary threshold 10.00, consider $p = p(a, \sigma, N) = \Pr(\max(Y_1', \ldots, Y_N') \geq y_0)$. With N being a $\mathrm{Pois}(\lambda)$, show that the probability of seeing a race r with $y = 10.005 - r \geq y_0$, in the course of a new season, is

$$p = p(a, \sigma) = 1 - \exp\{-\lambda(1 - ay_0/\sigma)^{1/a}\}.$$

(c) Use $\lambda = 195/8 = 24.375$, the rate of top races per year. For each threshold y_0 we may estimate $p(a, \sigma)$. With $y_0 = 10.005 - 9.72 = 0.285$, for 31 May 2008, compute $\widehat{p} = 0.035$; the estimated probabiity of seeing a 9.72 or better in the course of 2008, as judged from 1 January 2008, was 3.5 percent.

(d) It turns out that the delta metod for assessing the variability of $\widehat{p} = p(\widehat{a}, \widehat{\sigma})$ here does not work so well, even though $(\widehat{a}, \widehat{\sigma})$ is approximately binormally distributed. In spite of the sample size $n = 195$, the function $p(a, \sigma)$ is not well approximated by a linear function around the ML position. What works better is the Wilks theorem and the associated CD methods, as with Ex. 7.11. This requires computation of the log-likelihood profile function

$$\ell_{\mathrm{prof}}(p_0) = \max\{\ell(a, \sigma) \colon p(a, \sigma) = p_0\}$$

for a grid of p_0 values. To this end, show that $p(a, \sigma) = p_0$ entails $\sigma = aw/(1 - (\alpha_0/\lambda)^a)$, with $\alpha_0 = -\log(1 - p_0)$, leading to an easier one-dimensional optimisation problem, for each p_0. Compute the log-likelihood profile and use the Wilks theorem recipe to produce the confidence curve $\mathrm{cc}(p_0)$, as with Figure v.2, left panel. On the percentage scale, the point estimate is 3.4 and the 90 percent interval is $[0, 18.9]$; note the skewness. Carry out similar analysis for Bolt's 2008 Olympics race of 9.69, and transform estimates and confidence to the shock barometer scale of $100(1 - p)$, as with the figure's right panel. His 9.58 in Berlin August 2009 really shattered the scale, being very close to being unbelievable, as seen from January 2008 (but then we had shifted our scales of expectation).

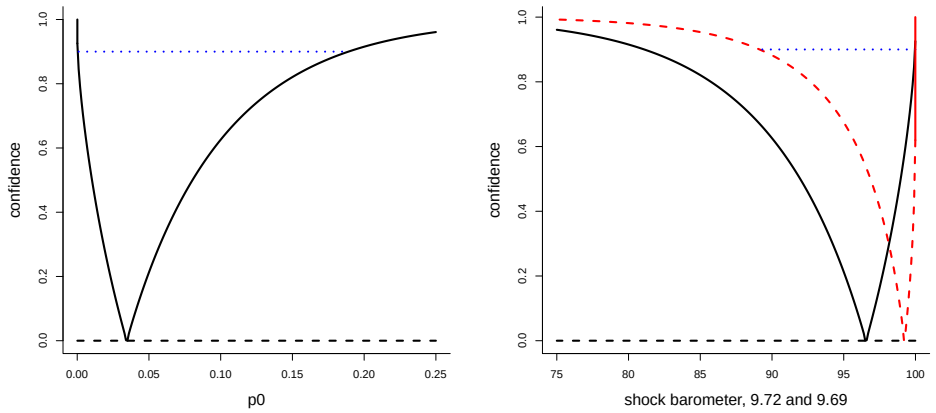


Figure v.2: *Left panel: confidence curve for the probability p of seeing a 100 m race with 9.72 or better, as judged by the start of the 2008 season. The point estimate is 3.4 percent, but the distribution is rather skewed; the 90 percent interval for p is $[0, 18.9]$ on the percentage scale. Right panel: the p probability transformed to the shock barometer scale $100(1-p)$, with confidence curves both for 9.72 and 9.69; we were even more shocked by his Olympics 2008 race, with p estimated at 0.7 percent, 90 percent confidence interval $[0, 10.8]$ percent, and shock barometer 99.2. His Berlin 2009 race of 9.58 is measured to be perfectly Beamonesque, on this scale.*

Story v.2 *Will there be a new World Record this season?*
Data: `fivekmA-data, fivekmB-data`
Code: `(xx to come, com235* of nilswork24 xx)`

by Swedish Olympic and World Champion Nils van der Poel

Each speedskating season, the classical gruelling twelve-and-a-half lap 5000 m is skated by several thousands athletes from about forty nations. The very best of these are world-top athletes, skating for medals and glory, and for setting personal bests. Over the calender years 2005–2024, there have been $n = 109$ splendid races below the threshold 6:10 (by a total of 33 skaters), with the world record as of 2024 being 6:01.56. Will there be a new world record, in the coming season, and will finally the ultimate holy grail barrier 6:00.00, an almost incomprehensible 50 km an hour speed for the 5k (200 m opening 18:00 and then twelve Olympic stadion laps with 28.5 seconds for each), be broken (and by whom)?

(a) We work with race results r, taken here as the amount in seconds above the 6:00.00 mark. As with the Bolt Story v.1, we translate race results r_i to $y_i = 10.00 - r_i$; the world record hence corresponds to the result to beat $y_0 = 10.00 - 1.56 = 8.44$. Fit the y_i to the two-parameter model with c.d.f. $G(y, a, \sigma) = 1 - (1 - ay/\sigma)^{1/a}$, via maximum likelihood; the result will be $(\widehat{a}, \widehat{\sigma}) = (0.207, 2.669)$. For the race results r_i, compute and display the empirical c.d.f. H_n, along with the parametrically fitted $H(y, \widehat{a}, \widehat{b})$, on the $r_i = 10.00 - y_i$ scale. Construct a version of Figure v.3, left panel.

(b) Assume that a new season will give us races $r'_1, \ldots, r'_N$ below 6:10, translated to nonnegative $y'_i = 10.00 - r'_i$. With N a Poisson, with rate λ, show as with the Bolt Story v.1 that the probability of having the best of these below some r_0, i.e. $\max y'_i$ above threshold $y_0 = 10.00 - r_0$, is $p_0(a, \sigma) = 1 - \exp\{-\lambda(1 - ay_0/\sigma)^{1/a}\}$. Taking $\lambda = 10$,

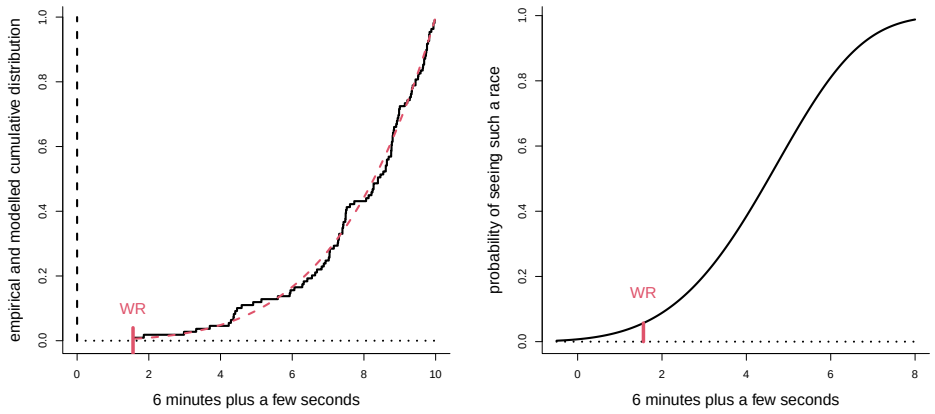


Figure v.3: *Left panel: the empirical c.d.f. H_n for the 109 sub-6:10 race results r_i, along with the fitted two-parameter c.d.f., associated with the points-above-threshold c.d.f. $G(y,\widehat{a},\widehat{b})$ for $y = 10.00 - r$. The world record 6:01.56 (as of 2024) is indicated to the left. Right panel: the estimated probability of having the best-in-season result below r, as a function of this r. The probability is 0.057 for seeing van der Poel's record beaten, in the course of a season, and a more tiny 0.007 for seeing the Six Minutes Barrier broken.*

plot the estimated probability curve, as with Figure v.3, right panel; in particular, the probability that a new season will see van der Poel's world record beaten is estimated at 5.7 percent. Show also that the probability of seeing the Six Minutes Barrier is estimated at 0.7 percent, tiny, but not unbelievable.

(c) For the probability p_0 of seeing the world record 6:01.56 broken in a season, compute the log-likelihood profile function $\ell_{\rm prof}(p_0) = \max\{\ell(a,\sigma)\colon p_0(a,\sigma) = p_0\}$, then use the Wilks theorem to produce the confidence curve $\text{cc}(p_0)$. Construct a version of Figure v.4, left panel, for the race thresholds 6:01.56 and the Dutch record 6:02.98.

by Patrick Roest, World Champion 2024 on the 5k

(d) The above analyses have assumed that the top level of the speedskating world has remained essentially stationary during the period 2005–2024. Followers of the sport agree that the appears reasonable, with no significant development in terms of skates and gear. To actually test this assumption, show first that the median of the $G(y,a,\sigma)$ distribution may be expressed as $y_0(a,\sigma) = (\sigma/a)\{1-(\frac12)^a\}$. This motivates a three-parameter regression model, where we allow x_i, the calendar year minus 2000, to show its possible influence on top races, with

$$G_i(y,a,\sigma,\gamma) = 1-(1-ay/\sigma_i)^{1/a} \quad \text{with } \sigma_i = \sigma\exp(\gamma x_i),$$

for race results obtained in that calendar year. Carry out the required log-likelihood regression analysis; show that $(\widehat{a},\widehat{\sigma},\widehat{\gamma}) = (0.207, 2.114, 0.013)$, with a non-signicant $\widehat{\gamma}$. Compute also the deviance statistic function $D(\gamma)$, and produce the Wilks theorem based confidence curve $\text{cc}(\gamma) = \Gamma_1(D(\gamma))$, shown in Figure v.4, right panel. (xx in Notes and Pointers, mention that this apparatus can be used for other sports with annual best lists, for predicting new records, and comparing them. the women's marathon record 2:09.56 of October 2024 is statistically ridiculously out-of-planet strong. xx)

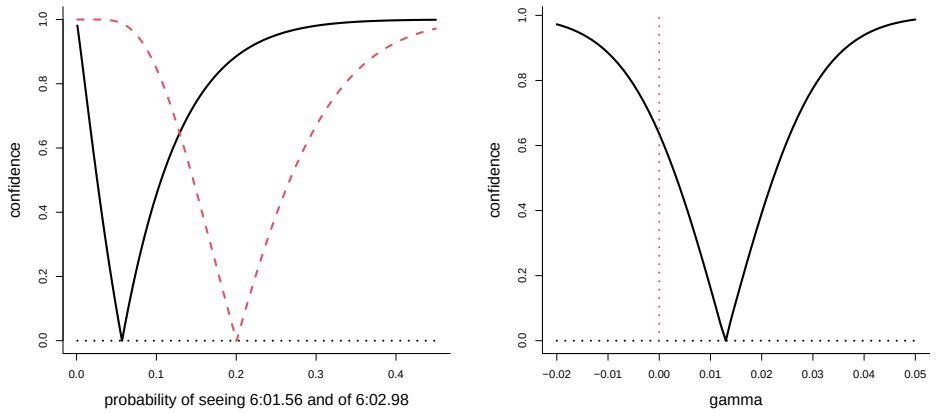


Figure v.4: *Left panel: confidence curves for the probabilities p_0 of seeing races better than respectively the van der Poel record 6:01.56 and the Dutch record 6:02.98. Right panel: the confidence curve* $\mathrm{cc}(\gamma)$ *for the season parameter in the regression model that takes* $\sigma_i = \sigma \exp(\gamma x_i)$*, with* x_i *the calendar year minus 2000. The top level has remained essentially stationary over the period 2005–2024.*

Story v.3 *Modern evolution of the 500 m sprint*
Data: `fivehundredA-data, fivehundedB-data`
Code: `(xx com236* of nilswork24 xx)`
(xx ranting first. could be taken as a part of previous story, if simplified. but there are nice plots and methods with regression for points-above-threshold. xx)

note Kulizhnikov 33.61 and Stolz 33.69

(a) plot the data, as in a version of Figure v.5, left panel. comment on what can be reasonbly seen from the plot. translating races r to points above threshold $y = 34.665 - r$, fit the y data to the two-parameter model $G(y, a, \sigma) = 1 - (1 - ay/\sigma)^{1/a}$. Do the monitoring plot $Z_n(y) = \sqrt{n}\{G_n(y) - G(y, \widehat{a}, \widehat{\sigma})\}$, with G_n the empirical c.d.f. for the $n = 1002$ races. Simulate a number of paths from the model at the ML position. Construct a version of Figure v.5, right panel, and comment on the model fit.

(b) this model is adequate provided the top level for speedskating sprinters has remained reasonably stationary over the time period 2000 to 20024. then try the model with $G(y_j, a, \sigma, \gamma) = 1 - (1 - a_j y/\sigma)^{1/a_j}$, where $a_j = a \exp\{\gamma(x_j - \bar{x})\}$, writing x_j for the calendar year. show that γ negative entails higher y, i.e. better performance. fit via likelihood and show that $\widehat{\gamma}$ is significantly negative.

(c) will there be a new world record the coming season (and by whom)? note λ increases with the season, possibly more so in Olympic seasons. compare probabilities with and without the γ. (xx) stranno: it looks like there is a negative impact in the Poisson model for the number of top taces, with $\lambda_j = \exp(\beta_0 + \beta_1 x_j + \beta_2 o_j)$. see Figure v.5, left panel. xx)

(d)

Story v.4 *Golf putting probabilities*
Data: `golf-data`

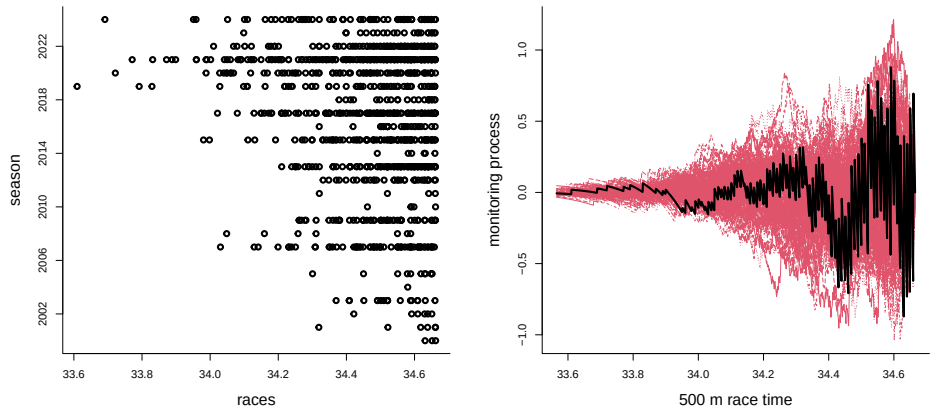


Figure v.5: *Left panel: all 1002 races r at 34.66 or better, obtained through the seasons 2000 to 2024. Right panel: model monitoring plot for the $G(y, a, \sigma) = 1 - (1 - ay/\sigma)^{1/a}$ model, for $y = 34.665 - r$, along with 100 model based simulations.*

Code: `(xx to come, com85* of nilswork24 xx)`

You're golfing, and when closer to the hole than some twenty feet need to focus on your putting. Drawn from databases of several hundreds of professional tournaments, the data below, from Gelman and Nolan (2002) with further discussion in Schweder and Hjort (2016, Ch. 14), give the number m_j of attempts and the number y_j of successful ones from these, at distances x_j, in feet, for say $j = 1, \ldots, k$. Our story concerns estimating the success probability $p(x_j)$, and also modelling the inherent variability at work. See Figure v.6, left panel, which in particular also displays the raw estimates $\widetilde{p}_j = y_j/m_j$, with small vertical 90 binomial confidence intervals around them. It will be of relevance for a few of these models to factor in the radii R for the hole and r for the ball, which are respectively 4.252/2 inches and 1.680/2 inches, or 0.0177 and 0.070 on the foot scale.

```
feet away; number of tries; number of successes
                  11  237   75
 2 1443 1346      12  202   52
 3  694  577      13  192   46
 4  455  337      14  174   54
 5  353  208      15  167   28
 6  272  149      16  201   27
 7  256  136      17  195   31
 8  240  111      18  191   33
 9  217   69      19  147   20
10  200   67      20  152   24
```

(a) We start out viewing the data as a sequence of independent binomial experiments, with $Y_j \sim \text{binom}(m_j, p_j)$ for $j = 1, \ldots, k$. The task is to model $p_1, \ldots, p_k$, as functions of the distances $x_1, \ldots, x_k$ to the hole. Show that with any such model, say $p_j = p(x_j, \theta)$, the log-likelihood function becomes $\sum_{j=1}^{k} [y_j \log p_j(\theta) + (m_j - y_j) \log\{1 - p_j(\theta)\}]$. Carry out logistic regressions, in x (order one); in $x, (x - \bar{x})^2$ (order two); in $x, (x - \bar{x})^2, (x - \bar{x})^3$

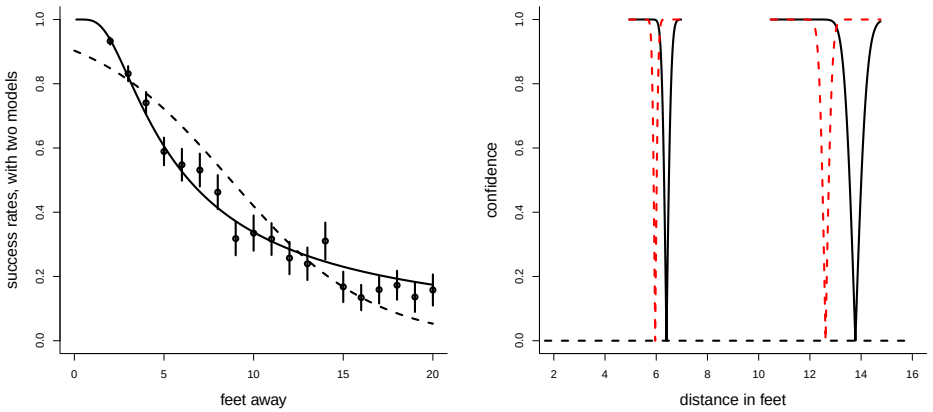


Figure v.6: *Left panel: the raw success estimates* $\widetilde{p}_j = y_j/m_j$ *at distances* $2, 3, \ldots, 20$ *feet from the hole, along with small vertical 90 percent binomial intervals. The full curve is the fitted* $p(x_j, a, b)$ *with the geometric model, and the dashed curve is the simple logistic regression curve. Right panel: confidence curves* $\mathrm{cc}(x_0)$*, the distance at which the putting probability is* $p_0 = 0.50, 0.25$ *(from left to right), with the good two-parameter geometric* $x_0(a, b)$ *model (full curves, black) and the not so good one-parameter* $x_0(\sigma)$ *model (slanted, red).*

(order three); in $x, (x-\bar{x})^2, (x-\bar{x})^3, (x-\bar{x})^4$ (order four). As usual, $\bar{x}$ is the mean of the x_j. For each of these models, estimate and plot the curves

$$p_1(x) = H(a+bx) \quad \text{up to} \quad p_4(x) = H(a+bx+c(x-\bar{x})^2+d(x-\bar{x})^3+e(x-\bar{x})^4).$$

This can be achieved in `R` via `glm(cbind(y,m-y) ~ x + x2 + x3, family=binomial)`, and so on; for this standard type of model there is then no need to programme the log-likelihood function etc. For the four models, find the log-likelihood maxima and AIC scores, as per Chapter 11. In particular, you should find that the most traditional order one model does not work well here, and that AIC prefers the order four model among these.

(b) Considering the population of good golfers, and disregarding other geometric aspects of these thousands of putting situations, let Z be *the angle of the put*, from putting position to the hole. Not all attempts are perfect (Z close to zero), so we can translate uncertainty and variability to a distribution of the random angle Z. In terms of such a distribution, show that to a good geometric approximation,

$$p(x) = \Pr(\sin Z \in (-(R-r)/x, (R-r)/x)) \quad \text{for } x \geq R-r.$$

A natural model for the random angles is a normal $(0, \sigma^2)$. Fit the resulting model

$$p(x_j, \sigma) = \Pr(\sigma N \in (-d_j, d_j)) \quad \text{for } j = 1, \ldots, k,$$

with N denoting a standard normal, and where we write $d_j = \arcsin((R-r)/x_j)$ for the bounds inside which successful putting angles must land at distance x_j. Compute the log-likelihood maximum, and compare with the logistic regressions above, using the

AIC scores. In particular, demonstrate that this simple geometric one-parameter model works better than logistic regressions of order one and two.

(c) (xx a little bit more with simple $p(x_i, \sigma)$ model before we go to variable σ. show that it works much better than standard first order two-parameter logistic regression; see also Figure v.6. check $\widehat{\sigma}_1, \ldots, \widehat{\sigma}_k$, fitted at the individual x_j. Can the $\widehat{\sigma}_j$ reasonably be taken as constant, across putting distances? can have a simple figure. xx)

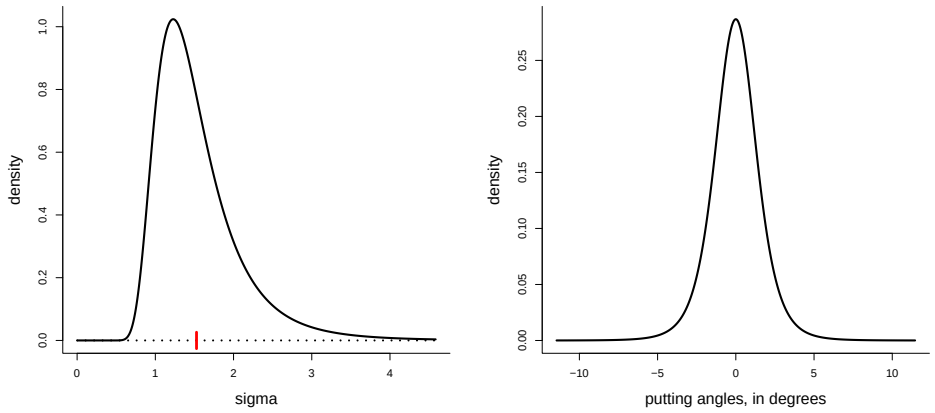


Figure v.7: *Left panel: the estimated density for σ, on the scale of ordinary degrees (i.e. $90/(\pi/2)$ times radians). The point estimate 1.53 from the no-variability model is shown on the horizontal axis. Right panel: the estimated density of putting angles, again in ordinary degrees.*

(d) The simple model above somehow puts all angular uncertainty into one common σ. It might be better and more informative to view these σ as coming from a distribution, across golfers. There are several such possibilities, starting with $Z \,|\, \sigma \sim \mathrm{N}(0, \sigma^2)$, but here we take σ^2 to have an inverse gamma distribution, i.e. $\lambda = 1/\sigma^2 \sim \mathrm{Gam}(a, b)$; the model is flexible, and we can get through the mathematics to an explicit distribution for Z. Writing $g(\cdot, a, b)$ for that gamma density, show that this leads to a density for the random Z of the form

$$\bar{f}(z) = \int_0^\infty \phi(\lambda^{1/2} z) \lambda^{1/2} g(\lambda, a, b)\, \mathrm{d}\lambda = \frac{1}{(2\pi)^{1/2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} \frac{b^a}{(b + \frac{1}{2} z^2)^{a+1/2}}.$$

While this may be worked with directly, it is useful to transform the density to a member of the well-known distributions, to facilitate computations of probabilities etc. Show therefore that

$$V = (\tfrac{1}{2} Z^2/b)/(1 + \tfrac{1}{2} Z^2/b) \sim \mathrm{Beta}(\tfrac{1}{2}, a),$$

and express the c.d.f. of Z in terms of the c.d.f. Be of this Beta distribution. Demonstrate that all this leads to the model

$$p(x_j, a, b) = \Pr(|Z| \le d_j) = \mathrm{Be}((\tfrac{1}{2} d_j^2/b)/(1 + \tfrac{1}{2} d_j^2/b), \tfrac{1}{2}, a) \quad \text{for } j = 1, \ldots, k.$$

Fit this model numerically, maximising the log-likelihood function; you should find $(\widehat{a}, \widehat{b}) = (2.8498, 0.00154)$. Compute also the log-likelihood maximum, and demonstrate that this two-parameter model has the best AIC score of the four plus two models considered (so far).

(e) Compute and display the estimated densities for σ, the variable normal scale for the putting angle, and for Z, the putting angle itself. Since most golfers prefer standard angular degrees to radians, present these densities on the degree scale $z' = 90/(\pi/2)\, z = (180/\pi)\, z$. Construct versions of the plots in Figure v.7. With Z the random angle, use these fitted densities to demonstrate that 95 percent of all putting angles are inside ± 3.30 degrees, and 99 percent are inside ± 5.05 degrees. Note that this signifies rather heavier tails than for the normal; in a fair proportion of cases, the shot is off with more than say 4 degrees, which is enough to not hit the hole.

(f) It does perhaps not appear likely, but we may check statistically whether this population of players might have some systematic angular bias in their putting. The simplest check on this is to use the two-parameter normal $Z \sim (\xi, \sigma^2)$. Compute and display the profiled log-likelihood $\ell_{\text{prof}}(\xi)$. It will indeed be seen to be very flat at the top, around zero, with no indication of such a bias. (xx point to previous stuff perhaps in Ch1 regarding noise in ξ being picked up in the σ. xx)

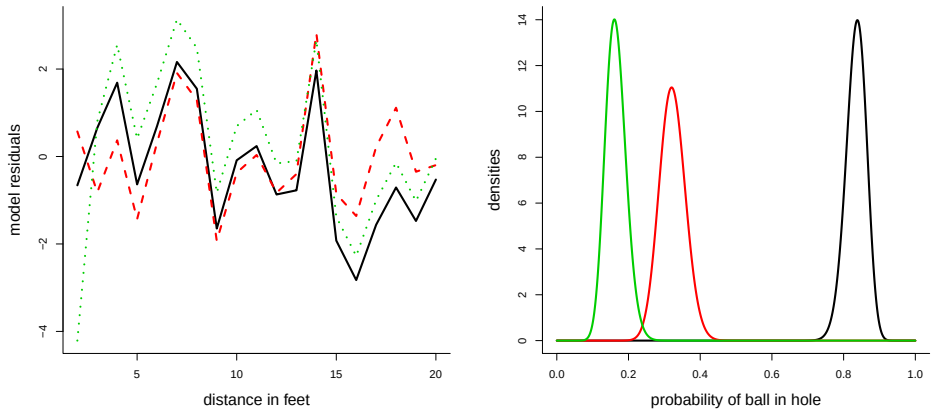


Figure v.8: *Left panel: model residuals* $(y_j - m_j\widehat{p}_j)/\{m_j\widehat{p}_j(1-\widehat{p}_j)\}^{1/2}$, *for two-parameter geometric model (full black), simple geometric model (dotted), and for the third-order logistic model (dashed). These indicate good fit, but overdispersion. Right panel: at distances* x *equal to 3, 10, 20 feet, the Beta-binomial model leads to probability densities, centred around respectiely 0.701, 0.322, 0.165. The strict binomial models take these probabilities as fixed, for all golfers at a fixed distance from the holw.*

(g) We have models for $p(x)$, the putting probability at distance x feet. This may be turned around to estimate the x_0 distance at which $p(x_0)$ equals some fixed p_0, along with its uncertainty; for the right panel of Figure v.6, we've used $p_0 = 0.25, 0.50$. For the one-parameter geometric model, show that $p(x, \sigma) = \Gamma_1(d(x)^2/\sigma^2)$, with $d(x) = \arcsin((R - r)/x)$, and derive the expression $x_0(\sigma) = (R - r)/\sin(a_0\sigma)$, where $a_0 = \Gamma_1^{-1}(p_0)^{1/2}$.

The task is then to compute $\ell_{n,\text{prof}}(x_0)$, for conversion to a confidence curve $\text{cc}(x_0)$, via the Wilks theorem based recipe of Ex. 7.11. Carry out this. The two-parameter geometric model is more involved, but should be better, by the discussion above. Solve $p(x,a,b)=p_0$ to find

$$x_0(a,b)=\frac{R-r}{\sin(\omega^{1/2})},\quad \text{with } \omega=2b\frac{\text{Be}^{-1}(p_0,\frac12,a)}{1-\text{Be}^{-1}(p_0,\frac12,a)}.$$

To compute the confidence curve, via the Wilks based recipe, we need the log-likelihood profile function, $\ell_{n,\text{prof}}(x_0)=\max\{\ell_n(a,b)\colon x_0(a,b)=x_0\}$. Show how this can be done my reducing the maximisation to a one-dimensional task, by expressing b in terms of a:

$$b(a)=\tfrac12\frac{\{\arcsin((R-r)/x_0)\}^2}{\text{Be}^{-1}(p_0,\frac12,a)/(1-\text{Be}^{-1}(p_0,\frac12,a))}.$$

Carry out all this, for $p_0=0.25,0.50$, to arrive at a version of Figure v.6, right panel, and comment on in which ways the two-parameter model gives different results from the one-parameter one. With your code, you may also experiment with lower probabilities, like 0.10 or 0.05, to see how far away the professional golfers are, where only one in ten or one in twenty succeed.

(h) (xx we do one more thing. binomial overdispersion. point to Story i.2. xx) There is another tool for assessing and comparing model adequacy, with such data for a collection of tables, which is to monitor the model residuals $\widehat{r}_j=(y_j-\widehat{p}_j)/\{m_j\widehat{p}_j(1-\widehat{p}_j)\}^{1/2}$ for $j=1,\ldots,k$, with $\widehat{p}_j$ that model's implied estimates of the p_j. If the model is adequate, explain why these should be distributed approximately as standard normals; also, the residual sum of squares $Q=\sum_{j=1}^k\widehat{r}_j^2$ should roughly have a χ^2_{df} distribution, with $\text{df}=n-p$, where p is the number of parameters estimated. Compute and display these residuals, for the models entertained so far; see the left panel of Figure v.8. It will be seen that even when the fitted $\widehat{p}_j$ manage to be close to the raw estimates y_j/m_j, there is binomial overdispersion; the $(y_j-m_j\widehat{p}_j)^2$ tend to be bigger than $m_j\widehat{p}_j(1-\widehat{p}_j)$.

(i) The modelling and analyses above rest on the binomial assumption, for each y_j and position x_j, which means relying on all shots having the very same success probability p_j. This is not entirely realistic, as seen via the model residuals in the previous point. This invites modelling an extra little layer of uncertainty in p_j around some central value $p_{j,0}$. A natural way for this is the Beta-binomial setup, see Ex. 1.22, with $y_j\,|\,p_j\sim\text{binom}(m_j,p_j)$, but $p_j\sim\text{Beta}(cp_{j,0},c(1-p_{j,0}))$. In other words, we use one of the parametric models for $p_{j,0}$, but then estimate the additional variability via c. Show that the log-likelihood becomes

$$\sum_{j=1}^k\log\Big[\frac{\Gamma(c)}{\Gamma(cp_{0,j}(\theta))\,\Gamma(c(1-p_{0,j}(\theta)))}\frac{\Gamma(cp_{0,j}(\theta)+y_j)\,\Gamma(c(1-p_{0,j}(\theta))+m_j-y_j)}{\Gamma(m_j+c)}\Big].$$

Analyse this binomial overdispersion model, for the case of the simple geometric $p_j(\sigma)=2\,\Phi(d_j/\sigma)-1$ model, by maximising the log-likelihood over (σ,c). Show that the log-likelihood maximum increases very significantly, from the one-parameter binomial based to the two-oarameter overdispersion model. This does not necessarily influence the estimated overall curve $p(x)$, but aims at describing the probability mechanisms much better, e.g. for prediction. (xx can ask for a figure to complent the first. instead of

binomial based 90 percent intervals around y_j/m_j, give 90 percent intervals induced by the estimated beta-binomial model. xx) (xx we might give a little table, for models 1, 2, 3, 4, then 5, 6, 7, with logLmax, AIC, Q. xx)

```
  dim   logLmax    aic       Q
1  2  -3020.155  -6044.309 258.968  logistic order 1
2  3  -2929.041  -5864.082  74.356  logistic order 2
3  4  -2912.873  -5833.745  40.743  logistic order 3
4  5  -2904.889  -5819.778  25.288  logistic order 4
5  6  -2904.365  -5820.731  24.337  logistic order 5
6  1  -2922.639  -5847.279  62.436  one-para geometric
7  2  -2911.589  -5827.178  36.741  geometric with extra
8  2  -2910.926  -5825.852  71.528  two-para Beta-binomial
```

Story v.5 *NBA three point shooting averages.* The last few years have seen a revolution in basketball, where teams depend more and more on the three point shot. This has been due to an analytics revolution in professional basketball, and also the impact of Stephen Curry, the three point expert and point guard for the Golden State Warriors of the National Basketball Association (NBA). NBA-players are regularly traded between different teams, both during the season (up until the trade deadline), and during the off-season. If you are looking to add a player to your team that can deliver from behind the arc, as they say, you would naturally base your decision on the past shooting of various players. In this story we investigate various ways of predicting future shooting based on past shooting. In particular, we are going to consider the 2018-2019 season, and use the shooting percentages from the first half of the season, to predict shooting percentages in the second half of the season. The dataset `NBAthrees20182019wdate.txt` contains the three point attempts, the three points made, and the minutes played, in each game, for each of the 530 players active during the 2018-2019 NBA season. Many thanks to the excellent website basketball-reference.com/ for making these data available!

An NBA regular season (usually) consists of each team playing 82 games. We split the season in the first 41 and the last 41 games of the season, and refer to these as the first and second half of the season. For player i, let $H_{1,i}$ and $H_{2,i}$ be the total number of three point makes in the first and second half of the season, respectively, and similarly, let $N_{1,i}$ and $N_{2,i}$ be the total number of three point attempts in the two halves of the season.

We restrict our analysis to the players with 100 or more three point attempts in the two halves of the season. Therefore, let $S_j = \{i\colon N_{j,i} \geq 50\}$ for $j = 1, 2$.

Conditionally on $N_{j,i} \geq 1$, we take $H_{j,i} \sim \text{binom}(N_{j,i}, p_i)$. This means that for the ith player the true three point hit rate p_i is taken as constant over the entire season.

(a) The estimators we develop are going to be based on theory for normally distributed data, as worked with in Ex. **??**, for example. Our first task is therefore to find a decent variance stabilising transformation. Show that $2 \arcsin \sqrt{H_{j,i}/N_{j,i}}$ is one such.

(b) A class of variance stabilising transformations is given by

$$Y^{(c)} = 2 \arcsin \big(\frac{H + c}{N + 2c}\big)^{1/2}, \quad \text{for } c > 0.$$

Show that $\text{Var}\, Y^{(c)} = 1/N + O(1/N^2)$ and that

$$\text{E}\, Y^{(c)} = 2 \arcsin \sqrt{p} + \frac{1 - 2p}{\sqrt{p(1-p)}} \frac{N^2(4c-1)}{4(N+2c)^3} + O(1/N^2).$$

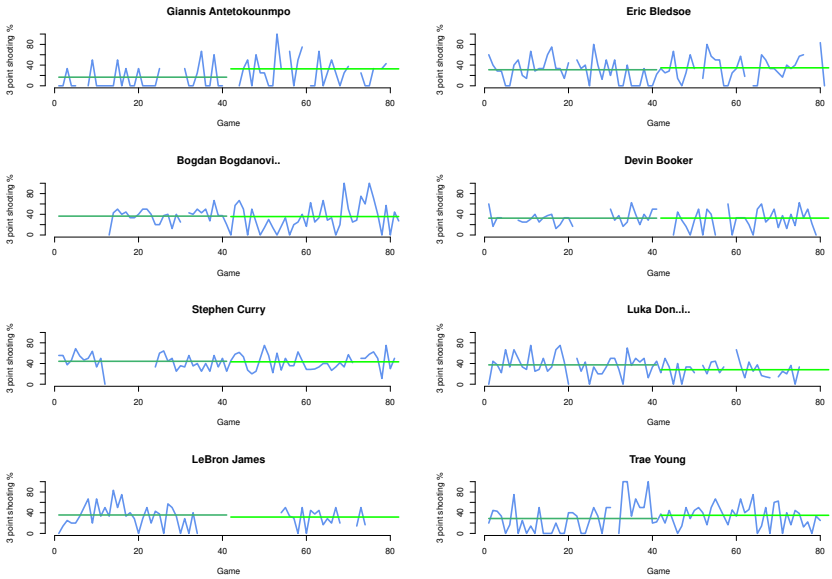


Figure v.9: Game-to-game three point shooting percentages of eight NBA players during the 2018–2019 season. The dark green lines indicate the three point percentage of the 41 first games of the season, while the bright green lines indicates the three point percentage of the 41 latter games of the season. The playoffs are not included. The games at which there is no blue line are games in which the player did not play, or did not attempt a single three point field goal.

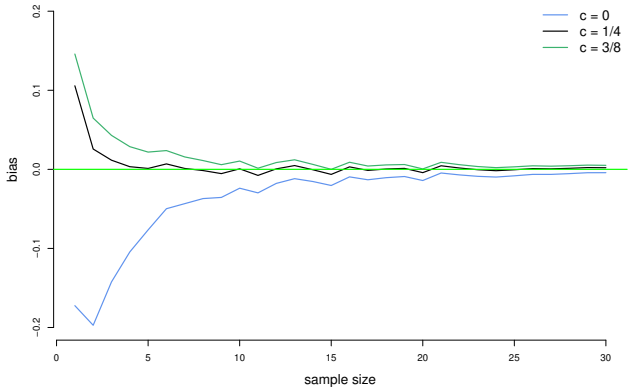


Figure v.10: Simulation estimates of the bias $\mathrm{E}\, Y^{(c)} - 2\arcsin\sqrt{p}$ for $c0$, $c = 1/4$, and $c = 3/8$.

From this expression we see that $c = 1/4$ gives us good control of the bias, in that $\mathrm{E}\, Y^{(1/4)} = 2\arcsin\sqrt{p} + O(1/N^2)$.

(c) Show that with $c = 3/8$, the variance is $\mathrm{Var}\, Y^{(3/8)} = 2/(4N+2) + O(1/N^3)$. [xx point to Anscombe 1948 xx].

(d) Reproduce the plot in Figure v.5. [xx more text xx]

(e) In view of Figure v.5 we choose $c = 1/4$ as our transformation, write $Y = Y^{(1/4)}$, and proceed as if $Y_{j,i} \sim \mathrm{N}(\theta_i, \sigma^2_{j,i})$ are independent random variables, where $\theta_i = 2\arcsin\sqrt{p_i}$, and $\sigma^2_{j,i} = 1/N_{j,i}$. The sum of squared prediction error of the estimator $\delta = \{\delta_i \colon i \in S_1\}$ is defined by

$$\mathrm{sspe}(\delta) = \frac{1}{|S_1 \cap S_2|} \sum_{i \in S_1 \cap S_2} (\delta_i - Y_{2,i})^2.$$

where $|S_1 \cap S_2|$ is the number of i in $S_1 \cap S_2$. Show that

$$\widehat{\mathrm{risk}}(\delta) = \mathrm{sspe}(\delta) - \mathrm{E}\,\mathrm{sspe}(\theta),$$

is an unbiased estimator of $\mathrm{risk}(\delta) = \mathrm{E} \sum_{i \in S_1 \cap S_2} (\delta_i - \theta_i)^2$. In the following, all estimators are compared with the naïve estimator $\delta_0(Y_1) = Y_1$, using

$$\widehat{\mathrm{rr}}(\delta) = \frac{\widehat{\mathrm{risk}}(\delta)}{\widehat{\mathrm{risk}}(\delta_0)} \tag{v.1}$$

(f) An estimator that is even more simple than δ_0 is the overall mean $\bar{Y}_1 = |S_1|^{-1} \sum_{i \in S_1} Y_{1,i}$. This is also the estimator that, in a sense, gives the maximal shrinkage. Find an expression for $\mathrm{risk}(\bar{Y}_1, \theta)$, and compute $\widehat{\mathrm{rr}}(\bar{Y}_1)$.

(g) A parametric Bayes model is the one where $\theta_i \sim \mathrm{N}(\mu, \tau^2)$ are independent, and

$$Y_{1,i} \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma^2_{1,i}), \quad \text{for } i = 1, \ldots, |S_1|,$$

are independent. Show that the Bayes solution under squared error loss is

$$\theta_{i,\text{Bayes}} = \mu + \frac{\tau^2}{\tau^2 + \sigma_{1,i}^2}(Y_{1,i} - \mu). \tag{v.2}$$

This is the pure Bayes solution in the sense that the hyper-parameters μ and τ^2 are subjectively chosen, hopefully by a basketball connoisseur. If you are no such, an empirical Bayes approach might be a more viable option.

(h) The empirical Bayes approach consists of estimating μ and τ^2 from the data, then replacing μ and τ in (v.2) with these estimates. There are several sensible estimators for these parameters. In the following we consider three such. Show that the marginal of $Y_{1,i}$ is $\mathrm{N}(\mu, \tau^2 + \sigma_{1,i}^2)$, and solve the system of equations,

$$\mathrm{E}\, Y_{1,i} = \frac{1}{|S_1|}\sum_{i\in S_1} Y_{1,i}, \qquad \frac{1}{|S_1|}\sum_{i\in S_1} \mathrm{E}\, Y_{1,i}^2 = \frac{1}{|S_1|}\sum_{i\in S_1} Y_{1,i}^2,$$

to find the method of moments estimators $\widetilde{\mu}$ and $\widetilde{\tau}^2$, say. Compute $\widehat{\mathrm{rr}}(\delta_{\text{mom}})$, where

$$\delta_{\text{mom},i} = \widetilde{\mu} + \frac{\widetilde{\tau}^2}{\widetilde{\tau}^2 + \sigma_{1,i}^2}(Y_{1,i} - \widetilde{\mu}).$$

(i) The method of moments type estimators $\widetilde{\mu}$ and $\widetilde{\tau}^2$ that you derived above might not be the most clever ones. First, $\widetilde{\mu}$ weighs the data points equally, even though their variances differ. Second, the estimator $\widetilde{\tau}^2$ might be negative. Another set of estimators are those that solves the system of equations given by

$$\breve{\mu} = \frac{\sum_{i\in S_1} Y_{1,i}/(\breve{\tau}^2 + \sigma_{1,i}^2)}{\sum_{i\in S_1} 1/(\breve{\tau}^2 + \sigma_{1,i}^2)}, \qquad \breve{\tau}^2 = \frac{1}{|S_1|}\sum_{i\in S_1}(Y_{1,i} - \breve{\mu})^2 - \frac{1}{|S_1|}\sum_{i\in S_1}\sigma_{1,i}^2.$$

You can find the estimates by implementing an iterative procedure in you programming language. Implement such a procedure, and compute $\mathrm{rr}(\breve{\delta})$, where $\breve{\delta}$ is the estimator in (v.2) with $\breve{\mu}$ and $\breve{\tau}^2$ inserted.

(j) The final estimator we consider is also an empirical Bayes estimator, but this time starting with a non-parametric prior for the θ_i. Let $\theta_i \sim F$ be independent, where F is a distribution on the real line. Show that the Bayes solution under squared error loss is

$$\theta_{F,i} = \frac{\int \theta_i \phi((Y_{1,i} - \theta_i)/\sigma_{1,i}) F(\mathrm{d}\theta)}{\int \phi((Y_{1,i} - \theta_i)/\sigma_{1,i}) F(\mathrm{d}\theta)},$$

where $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$ is the standard normal density. Show also that the Bayes solution can be expressed as

$$\theta_{F,i} = Y_{1,i} + \sigma_{1,i}^2 \frac{\partial g(y)/\partial y\big|_{y=Y_{1,i}}}{g(Y_{1,i})},$$

where $g(y)$ is a density function, a fact of which you should convince yourself.

(k) [xx something about estimating g using Ch. 12 nonparametric density estimation methods xx]

(l) [xx Simulations. xx]

Estimator	$\widehat{\text{rr}}(\delta)$
deltabar	0.4820
MoM0	0.3020
MoM1	0.3286
MoM2	0.3285
ML	0.3169
JS	0.3276
JShet	0.2957

Table v.1: Risk reduction relative to the estimator δ_0 for xyz estimators. The statistic $\widehat{\text{rr}}(\delta)$ is defined in (v.1).

(m) [xx predict three point shooting in the second half of the season. Fix Table v.1 below xx]

Story v.6 *Olympic Unfairness I: Inner and outer for 1000 m speedskating*
Data: `wsprint1000-01men to wsprint1000-24men`
Code: `(xx to come, com9* of sporskating18 xx)`
(xx again, nils-emil need to finetune balance between what is in dataoverview and what counts as intro here. Data in 2.B. analysis of the olympic unfairness parameter d, in a sequence of World Championships. cc(d_j), and combined, using Schweder and Hjort (2016); Cunen and Hjort (2022). nils ranting on before polish and editing. aiming towards Figure v.12, left and right panels. from sporskating18 com9* and com10*. Obviously the xxxx seconds estimated advantage is an overall figure across both events and skaters, and different skaters handle the challenges differently. Among those having already commented on the potential difference between inner and outer conditions are four Olympic gold medallists in their autobiographies (Holum, 1984; Jansen, 1994; Le May Doan, 2002; Friesinger, 2004). Dianne Holum, op. cit. page 225, rather clearly blames starting in the outer on her Sunday 1000 m for losing the world championship to Monika Pflug. xx) This story concerns the 1000-m speedskating race, raced in Winter Olympics since 1976, as part of the annual World Sprint Championships since 1970, and in the annual World Single Distances Championships since 1996. (xx point to: only one race in Olympics, but two races in World Sprint. medals easily change necks, so $\widehat{d} = 0.12$ is a significant figure. travelling 15 m per second means 1.5 m difference if your opponent has a 0.1 second advantage. xx) As one either knows, via correct cultural upbringing, or by studying the geometry of Figure v.11, this two-and-a-half-lap race is more asymmetric than for the other Olympic distances (500-m, 1500-m, 5000-m, 3000-m for ladies, 10000-m for men). If you start in inner, you'll have *in in out out in* before you cross the finishing line, whereas your compatriot in the race will have *out out in in out*. Ideally, of course, there should be no noticeable difference between starting in inner or outer, as 1000 m is 1000 m. There are two reasons casting some doubt on this Fairness Hypothesis, however: (i) a start in inner gives a longer straight stretch when building up to maximum speed of above, 55 km/h; and (ii) the last outer is a heavier burden on an athlete's aching body, after nearly a minute at such top speed.

So, though an Olympic or World Championship 1k certainly can be won by a top skater starting in the outer lane, many prefer and hope for being blessed by starting in the inner lane. That's the Olympic Question we'll be addressing below, whether there is

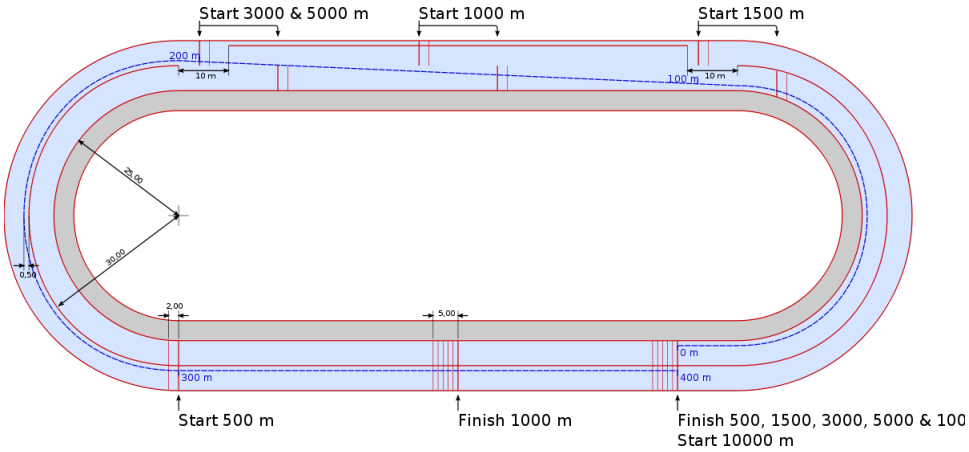


Figure v.11: The rink, where we in this story focus on the two-and-a-half lap 1000-m race.

a perhaps slight but statistically and Olympically significant difference at work. In that case, it should not and cannot be expected to work in the same way, for all skaters, from occasion to occasion. Rather, we're after a statistical average-across-skaters parameter, say

d = average time lost by starting outer compared to if you could have started inner,

which by its nature is *small*, i.e. close to zero; if it had not been small, it would have been detected and agreed upon since the 1970ies. To make such a d parameter well defined, and identifiable and estimable from data, we need relevant good-quality data, along with a proper statistical model. Again, the d looks a bit counterfactual, and can't easily be assessed from results for one given race; also, its value (when we find it, below) doesn't apply to all skaters; rather, it's an average-across-top-skaters value (in well executed races without falls or mishaps).

(a) (xx nils decides on one particular WCh Sprint dataset to work through first, before becoming meta. before we come to the full mixed-effects model, we do simple things. simple regression y_1 on z_1 and y_2 on z_2, show that this is about the same as t-testing the inner vs. the outer group; not powerful. better with the pretty informative differences, $D_i = y_{2,1} - y_{1,i}$, regress on z_i, or on easy covariates like $u_{i,2} - u_{i,1}$, $v_{i,2} - v_{i,1}$, and inner-outer. this works, but is less precise than using the seven-parameter model. also do wilcoxon, or even wilcoxon with parameter. xx) i think i for WCh Sprint 2020, Hamar, Feb 28-29:

	day one				day two			
		200	600	finish		200	600	finish
1	T Shinhama	o	16.12	41.30	68.28	i	16.12	41.46	68.71
2	L Dubreuil	o	16.42	41.33	68.79	i	16.29	41.21	68.39
3	MK Cha	i	16.44	41.60	69.26	o	16.24	41.09	68.73
4	K Verbij	o	16.71	42.04	68.96	i	16.81	42.09	68.75
5	HH Lorentzen	o	16.79	42.18	68.95	i	16.74	42.05	68.84
6	Y Matsui	i	16.44	41.77	68.69	o	16.51	41.98	69.89
	and onwards								

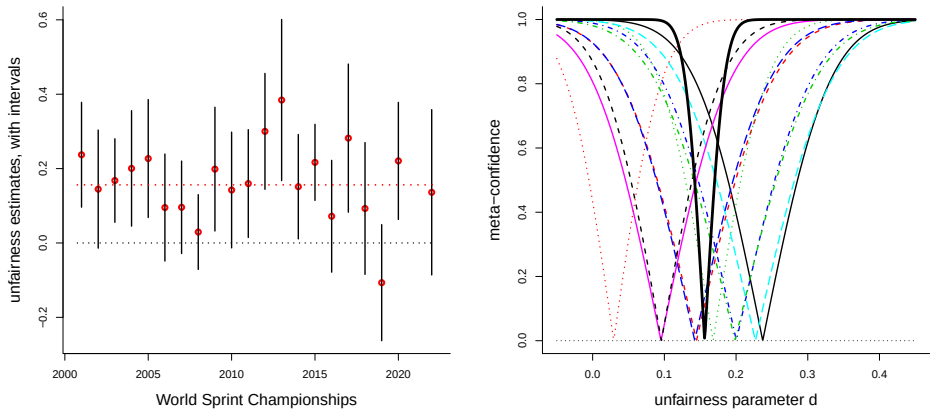


Figure v.12: *Left panel: estimated unfairness parameters* $\widehat{d}_j$*, along with 95 percent confidence intervals, for the World Championships 2001 to 2022 (with no such event for 2021). The y scale is in seconds, and the overall estimate (xx checkit xx)* $\widehat{d} = 0.156$ *is a big one on the Olympic scale. Right panel: confidence curves for the individual unfairness parameters* d_j*, along with a meta-analysis confidence curve* $\mathrm{cc}(d_0)$ *for the overall mean of all the* d_j*.*

(xx still ranting on, will be shortened and cleaned. xx) We may do better, however, by building a coherent statistical model for the (y_1, y_2), incorporating all relevant information, from 200- and 600-passing times to the variability structure (which should encompass both day-to-day variation and skater-to-skater variation), and, crucially, the inner-outer-information. Our considered model takes this form, for 1st race and 2nd race results $(y_{i,1}, y_{i,2})$ for skater no. i:

$$\begin{aligned} y_{i,1} &= a_1 + bu_{i,1} + cv_{i,1} + \tfrac{1}{2}dz_{i,1} + \delta_i + \varepsilon_{i,1}, \\ y_{i,2} &= a_2 + bu_{i,2} + cv_{i,2} + \tfrac{1}{2}dz_{i,2} + \delta_i + \varepsilon_{i,2}, \end{aligned} \tag{v.3}$$

with $(u_{i,1}, u_{i,2})$ passing times after 200 m, $(v_{i,1}, v_{i,2})$ passing times after 600 m; δ_i a parameter following skater i, modelled as from a $\mathrm{N}(0, \kappa^2)$ distribution, and $(\varepsilon_{i,1}, \varepsilon_{i,2})$ are independent error terms, from a $\mathrm{N}(0, \sigma^2)$ distribution. Crucially, the inner-outer information is taken care of via $(z_{i,1}, z_{i,2})$, with $z = -1$ if start in inner and $z = 1$ if start in outer. If $d = 0.08$, for example, it means a time adjustment of -0.04 seconds for the inner guy (good) but $+0.04$ for the outer guy (bad).

The race-to-race variation is in the $\varepsilon_{i,j}$, identified with the σ, whereas the skater-to-skater variation is in the δ_i, identified with the κ. A splendid skater has a negative δ_i (and a not-so-splendid skater has a positive δ_i), though we do not attempt to estimate it directly, for each skater, only through the variation among skaters. It is furthermore practical to slightly modify the variables here, by subtracting overall means for $u_{i,1}$ and $u_{i,2}$ in their definitions, and similarly for the $v_{i,1}$ and $v_{i,2}$; this eases interpretation and also helps stabilise numerics. In particular, a_1 and a_2 may now be seen as the overall expected levels on the 1st and 2nd day of the competitions. Parameters b, c, d are regression structure parameters, signalling the effect of u and v and z on the overall results, whereas κ and σ relate to variability and inter-skater correlation.

(b) (xx building the mixed effects model; interpreration of parameters. xx) The model has both fixed effects, related to the a_1, a_2, b, c, d parameters, and random effects, via the skater-parameter δ_i. We use a_1 for race 1 and a_2 for race 2, since racing conditions might differ, in terms of temperature, humidity, etc. (and wind, if outdoor). The model thus have seven parameters, and may also be represented as

$$y_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} \sim \mathrm{N}_2(x_i\beta, \begin{pmatrix} \sigma^2+\kappa^2 & \kappa^2 \\ \kappa^2 & \sigma^2+\kappa^2 \end{pmatrix})$$

for skaters $i = 1, \ldots, n$, with $\beta = (a_1, a_2, b, c, d)^{\mathrm{t}}$ and with the appropriate 2×5 covariate matrix

$$x_i = \begin{pmatrix} 1 & 0 & u_{i,1} & v_{i,1} & \frac{1}{2}z_{i,1} \\ 0 & 1 & u_{i,2} & v_{i,2} & \frac{1}{2}z_{i,2} \end{pmatrix}$$

(c) Given the combined considerable efforts of the skaters sprinting away in a World Sprint event, we can now fit and assess the model, and, in particular, learn whether the world was unfair then, by seeing if d was close to zero or not. (xx below: direct difference estimator, which is ok, but not the sharpest precision. then full seven-parameter model, used even though our primary interest lies with only one of these, the d. its precision. profiling for $\rho = \kappa^2/(\sigma^2+\kappa^2)$, the interskater correlation, of separate interest. then cc(), for one championships event at the time, followed by meta-analysis, combining $\mathrm{cc}(\dot{d}_j)$ across several championships. include our wilcoxon ranksum model. xx) With $\rho = \kappa^2/(\sigma^2+\kappa^2)$ the intraskater correlation, write

$$\Sigma = \begin{pmatrix} \sigma^2+\kappa^2, & \kappa^2 \\ \kappa^2, & \sigma^2+\kappa^2 \end{pmatrix} = \frac{\sigma^2}{1-\rho}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

and show via the binormal model that the log-likelihood function may be written

$$\ell(\beta, \sigma, \rho) = -2n\log\sigma - \tfrac{1}{2}n\log\frac{1+\rho}{1-\rho} - \tfrac{1}{2}\frac{1}{\sigma^2}\frac{Q(\beta)}{1+\rho},$$

with

$$Q(\beta) = \sum_{i=1}^n (y_i - x_i^{\mathrm{t}}\beta)^{\mathrm{t}} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} (y_i - x_i^{\mathrm{t}}\beta) = Q_1(\beta) + Q_2(\beta) - 2\rho Q_3(\beta),$$

in which $Q_1(\beta) = \sum_{i=1}^n (y_{i,1} - x_{i,1}^{\mathrm{t}}\beta)^2$, $Q_2(\beta) = \sum_{i=1}^n (y_{i,2} - x_{i,2}^{\mathrm{t}}\beta)^2$, and $Q_3(\beta) = \sum_{i=1}^n (y_{i,1} - x_{i,1}^{\mathrm{t}}\beta)(y_{i,2} - x_{i,2}^{\mathrm{t}}\beta)$. Show also that there is an explicit minimiser of $Q(\beta)$, for a given ρ, namely

$$\widehat{\beta}(\rho) = \{M_{1,1} + M_{2,2} - \rho(M_{1,2} + M_{2,1})\}^{-1}\{S_{1,1} + S_{2,2} - \rho(S_{1,2} + S_{2,1}\},$$

where $M_{u,v} = (1/n)\sum_{i=1}^n x_{i,u}x_{i,v}^{\mathrm{t}}$ are 5×5 matrices and $S_{u,v} = (1/n)\sum_{i=1}^n x_{i,u}y_{i,v}$ are 5-vectors, for $u, v = 1, 2$.

(d) Regarding σ, show also that the log-likelihood for given ρ is maximised by

$$\widehat{\sigma}^2(\rho) = \frac{1}{1+\rho}\frac{Q(\widehat{\beta}(\rho))}{2n}.$$

Explain how all of this reduces the numerical optimisation to a one-parameter problem, that of maximising the profiled log-likelihood

$$\ell_{\mathrm{prof}}(\rho) = \ell(\widehat{\beta}(\rho), \widehat{\sigma}(\rho), \rho) = -n\log\{Q(\widehat{\beta}(\rho))/2n\} + \tfrac{1}{2}\log(1-\rho^2) - n.$$

The main focus of the story is to reach inference for the d parameter, for each event with ensuing meta-analysis for a list of events, but take time to explain how the profiled log-likelihood gives rise to estimates and confidence curves for ρ. (xx also mention: can maximise the seven-parameter log-likelihood numerically, but numerics work better via this profiling over ρ and then reading off the rest via formulae; also, we get cc(ρ) via wilks etc. xx)

(e) We spend a few extra efforts on estimating σ precisely, since this matters when assessing the precision of the main regression coefficients a_1, a_2, b, c, d. Assume for a minute that ρ is known. Show that $\widehat{\beta}(\rho)$ is normal, unbiased, and with covariance matrix $(1/n)\sigma^2(1+\rho)M_\rho^{-1}$, where $M_\rho = M_{1,1} + M_{2,2} - \rho(M_{1,2} + M_{2,1})$. Show also that in fact

$$2n\widehat{\sigma}^2(\rho)/(1+\rho) = Q(\widehat{\beta}(\rho))/(1+\rho) \sim 2\sigma^2\chi^2_{2n-p},$$

with p the number of parameters in the β, with $Q(\widehat{\beta}(\rho))$ independent of $\widehat{\beta}(\rho)$. Argue that this invites the unbiased estimator

$$\widehat{\sigma}^2_{\rm un} = \frac{1}{1+\widehat{\rho}}\frac{Q(\widehat{\beta}(\widehat{\rho}))}{2n-p} = \frac{2n}{2n-p}\widehat{\sigma}^2.$$

It is slightly larger than $\widehat{\sigma}^2$, to take estimation variability of the p regression coefficients into account. We similarly use $\widehat{\kappa}^2_{\rm un} = \widehat{\sigma}^2_{\rm un}\widehat{\rho}/(1-\widehat{\rho})$ as a sample-sized adjustment for the ML estimator. These arguments are not disturbed by the insertion of $\widehat{\rho}$ for ρ, since $\widehat{\rho}$ is approximately independent of $\widehat{\beta} = \widehat{\beta}(\widehat{\rho})$, as we show next.

(f) We already have a clear strategy for estimating d, and are now closing on the required extra pieces of knowledge and algorithms required to assess its precision (and yes, we need to work with the full seven-parameter model in order to reach precise inference for this tiny focus parameter). Show that the Fisher information matrix for the seven-parameter model becomes

$$J = \begin{pmatrix} M_\rho/\{\sigma^2(1+\rho)\} & 0 & 0 \\ 0 & 4/\sigma^2 & 2/\{\sigma(1-\rho^2)\} \\ 0 & 2/\{\sigma(1-\rho^2)\} & 2/(1-\rho^2)^2 \end{pmatrix}.$$

Explain that all of this, via general likelihood theory of Ch. 5, leads to approximate normality $\mathrm{N}_p(\beta, (\sigma^2/n)(1+\rho)M_\rho^{-1})$ for $\widehat{\beta}$, and, in particular, that

$$\widehat{d} \approx_d \mathrm{N}(d, (1/n)\sigma^2(1+\rho)k(\rho)),$$

say, with $k(\rho)$ the lower-right element of M_ρ. Use all of this to go through some or all of the World Sprint Championships 2001 to 2022, producing estimates and confidence intervals, leading to Figure v.12. (xx round off sentence. xx)

(g) (xx then need to tend to outlier identification too, to make our Olympic level results statistically robust. motivate these outlier tests. xx)

$$t_{i,1} = \{y_{i,1} - (\widehat{a}_1 + \widehat{b}u_{i,1} + \widehat{c}v_{i,1} + \tfrac{1}{2}\widehat{d}z_{i,1})\}/(\widehat{\sigma}^2_{\rm un} + \widehat{\kappa}^2_{\rm un})^{1/2},$$
$$t_{i,2} = \{y_{i,2} - (\widehat{a}_2 + \widehat{b}u_{i,2} + \widehat{c}v_{i,2} + \tfrac{1}{2}\widehat{d}z_{i,2})\}/(\widehat{\sigma}^2_{\rm un} + \widehat{\kappa}^2_{\rm un})^{1/2}.$$

These should be like realisations from an approximate standard normal. We judge a skater to have a result above normal bounds, and exclude him or her from the final analysis, if either of these two are above 2.50. We do however allow unusually *good* results to remain, so an exceptional result associated with an unusually *low* $t_{i,1}$ or $t_{i,2}$ is kept in the analysis.

(h) (xx on to final primary interest analysis. first simple, assuming all d_j are the same d, across events, with each $\widehat{d}_j$ having a $\mathrm{N}(d, \sigma_j^2)$. then straight meta-analysis. xx) Do all of this, event for event, including care with outliers. Set up a table of $\widehat{d}$, with estimated standard deviation, for each event. do also $\mathrm{cc}(\rho)$, since this is of contextual relevance, the relative stability from day one to day two for the skaters.

(i) (xx then more complex meta-analysis, taking on board that the underlying d_j may not be identical, but exhibit a certain variation across events. so this is $d_j \sim \mathrm{N}(d_0, \tau^2)$, leading to $\widehat{d}_j \sim \mathrm{N}(d_0, \tau^2 + \sigma_j^2)$. first do a CD for τ. then arrive at Figure v.12 (right panel), with many individual $\mathrm{cc}_j(d_j)$, along with the final $\mathrm{cc}^*(d_0)$. xx)

Story v.7 *Olympic Unfairness II: From semifinals to finals*
Data: `crosscountry-data`
Code: `(xx to come, com92* of nilswork22)`
(xx to come. point to Hjort (2017b). balance intro here with the brief description we go for in dataoverview. data: of the type A A B B A A, for about twenty different events, then go for rank sums. nils ranting on before cleaning and editing. sum of many small wilcoxons. check that data are described in dataoverview. check balance between dataoverview and here. xx) The Formula One event for cross-country skiers is the sprint, where the very best athletes need to go through four strenuous three-minute Olympic-intensity competitions in a row: prologue, quarterfinals, semifinals, finals. So after three already quite gruelling tasks, each demanding top manouevring skills and split-second tactical decisions, at barely imaginable speeds of 30 km/hour and more, over about 1.5 km distances in highly varying terrain, the six finalists are ready. If the Olympic World is fair, it will not matter whether they come from Semifinal A or Semifinal B. There is evidence that the world is not fully fair, however, as we are to demonstrate and assess here.

The two best from each semifinal are qualifified for the final, along with the two so-called lucky losers, those with best time among the remaining four + four. Our `semifinals-finals` dataset, compiled over 63 top events (Olympics 2022, 2018, 2014, 2010, World Champiponships 2021, 2019, 2017, 2015, 2013, 2011, and various World Cup events), has the final ranks for the Semifinal A skiers, hence involving three different cases: case (2,4), with 2 A and 4 B; case (3,3), with 3 A and 3 B; and case (4,2), with 4 A and 2 B.

```
From 2022 Winter Olympics:                From 2018 Winter Olympics:
1 J.H. Klaebo    A  1 J. Sundling    A     1 J.H. Klaebo    A  1 S. Nilsson    A
2 F. Pellegrino  A  2 M. Dahlqvist   A     2 P. Pellegrino  A  2 M.C. Falla    A
3 A. Terentyev   B  3 J. Diggins     B     3 A. Bolshunov   A  3 Y. Belorukova B
4 J. Maeki       B  4 R. Brennan     A     4 P. Golberg     A  4 N. Nepryayeva B
5 A. Maltsev     B  5 N. Faehndrich  A     5 O. Svensson    B  5 H. Falk       A
6 O. Svensson    B  6 E. Ribom       B     6 R. Hakola      B  6 J. Diggins    B
```

(a) Introduce the rank sums $Z_A = \sum_i X_i$ and $Z_B = \sum_j Y_j$, where X_i and Y_j are the rank positions for A skiers and B skiers, respectively. For the 2022 Zhangjiakou ski-stadion races, we have (Z_A, Z_B) equal to $(3, 18)$ for the men and $(12, 9)$ for the women; we always have $Z_A + Z_B = 21$ and may hence restrict attention to the Z_A. Explain that under the fairness hypothesis, Z_A is the sum of 2 numbers randomly drawn from $\{1, \ldots, 6\}$, in case (2,4), and similarly the sum of 3 random numbers for case (3,3), and the sum of

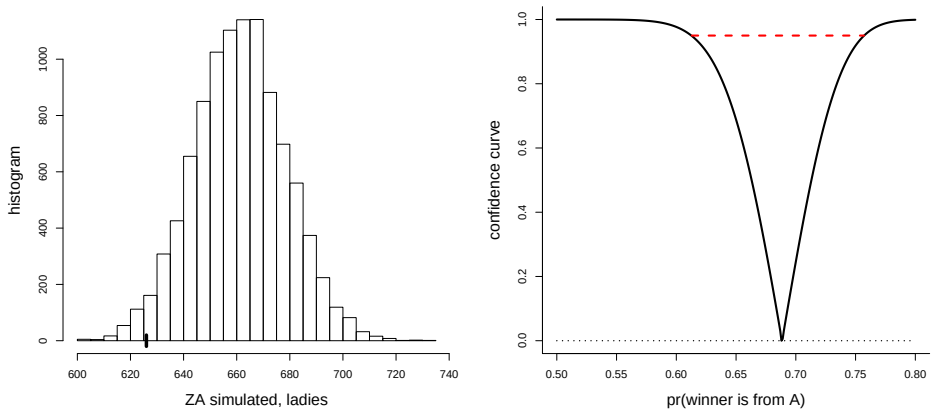


Figure v.13: *Left panel: histogram of* 10^4 *realisations of the combined ranksum* Z_A*, for the ladies skiers, taken over 23 cases with 2 A and 4 B, 33 cases with 3 A and 3 B, and 23 cases with 4 A and 2 B, simulated under the null hypothesis of fairness. The observed* $Z_{\rm obs} = 626$ *is indicated, with p-value 0.022. A corresponding figure for men shows much more dramatic evidence against the fairness assumption, with* $Z_{\rm obs}$ *to the left of almost all of* 10^4 *realisations. Right panel: confidence curve* $\mathrm{cc}(p_A)$ *for the probability that the winner is from Semifinal A, for the men. The 95 percent interval is from 0.613 to 0.757. The corresponding 95 percent interval for the ladies is from 0.504 to 0.658.*

4 random numbers, for case (4,2). Work out that Z_A has mean $7.0, 10.4, 14.0$, for these cases, along with variances $4.67, 5.25, 4.67$ (xx from Wilcoxon test exercise, if we include such a thing, perhaps in Ch4; $(1/12)(n+1)m(n-m)$. xx)

(b) To test the Olympic fairness, consider the overall ranksum test statistic

$$Z = \sum_{j=1}^{63} Z_{A,j} = \sum_{\text{two A}} Z_{A,j} + \sum_{\text{three A}} Z_{A,j} + \sum_{\text{four A}} Z_{A,j},$$

with sums taken over the (2,4), (3,3), (4,2) cases. For the 63 events for the men, use the `semifinals-finals` dataset to learn that there are 17, 9, 37 events of the three types, with total ranksum $Z_{\rm men,obs} = 88 + 68 + 495 = 651$. Show similarly that there are 23, 17, 23 events of the three types, for the ladies, with total ranksum $Z_{\rm women,obs} = 151 + 173 + 302 = 626$.

(c) The question is then how relatively unlikely it is, under fairness, to have Z as low as 651 or lower, for 17 + 9 + 37 events, and as low as 626 or lower, for 23 + 17 + 23 events. Explain why the null distribution for Z must be approximately normal. (xx pointer here to what is soon a wilcoxon exercise in Ch4 with formulae for variances etc. xx) Give such normal approximations, and test the fairness hypothesis, for men and for ladies separately. Carry out this testing also via direct simulation, bypassing the need for means and variances and normal approximations. Construct a version of Figure v.13 (left panel), ending with a p-value of 0.022 for the ladies. Carry out the same work for the men, where the evidence against fairness is much more dramatic, with a p-value close to zero; almost zero of 10^4 null simulated Z are as small as 651.

(d) The efforts above demonstrate that male skiers from Semifinal A have a very clear advantage over those from Semifinal B; the evidence is also there, but less pronounced, for the ladies. This statistical null hypothesis testing does not tell us *the degree* of unfairness, however. To assess the relevant $p_A = \Pr(\text{winner is from A})$ we need a more nuanced framework. Such a model is the following, concerned only with ranks 1 to 6, not with time differences:

$$\begin{aligned} f_{2,4}(i,j,\theta) &= e^{\theta(i+j)}/K_{2,4}(\theta), \\ f_{3,3}(i,j,k,\theta) &= e^{\theta(i+j+k)}/K_{3,3}(\theta), \\ f_{4,2}(i,j,k,\theta) &= e^{\theta(i+j+k+l)}/K_{4,2}(\theta), \end{aligned}$$

for ranks (i,j) for the (2,4) cases, for ranks (i,j,k) for the (3,3) cases, and ranks (i,j,k,l) for the (4,2) cases. Here $K_{2,2}(\theta)$ is a sum over all 15 terms for picking two results for two A skiers, and similarly for $K_{3,3}(\theta)$ and $K_{4,2}(\theta)$, with respectively 20 and 15 terms. Explain that if $\theta = 0$, the world is fair, whereas a negative θ means an advantage for the A skiers. Show that the log-likelihood function can be written

$$\begin{aligned} \ell(\theta) = &\sum_{\text{two } A} \{\theta(X_1+X_2) - \log K_{2,4}(\theta)\} + \sum_{\text{three } A} \{\theta(X_1+X_2+X_3) - \log K_{3,3}(\theta)\} \\ &+ \sum_{\text{four } A} \{\theta(X_1+X_2+X_3+X_4) - \log K_{4,2}(\theta)\}, \end{aligned}$$

with $X_1 + X_2$ notation for the two ranks, for all events of type (2,4), etc. For the men, the sums are over 17, 9, 37 events of the three types, and the $K_{2,4}(\theta)$, $K_{3,3}(\theta)$, $K_{4,2}(\theta)$ are themselves sum over 15, 20, 15 terms. Carry out ML estimation and inference. In particular, construct a confidence curve $\text{cc}(\theta)$, for the men and for the ladies. Explain why these confidence curves are actually optimal, under the assumed model.

(e) We now attempt to estimate and assess uncertainty not merely for the model parameter θ, but for relevant implied quantiies, like the probability that the winner is from A. Express this is $p_A = \Pr_\theta(\text{winner is from } A)$ as

$$\begin{aligned} p_A = &\, q_2 \Pr_\theta(\text{one of } X_1X_2 \text{ is } 1) + q_3 \Pr_\theta(\text{one of } X_1X_2, X_3 \text{ is } 1) \\ &+ q_4 \Pr_\theta(\text{one of } X_1X_2, X_3, X_4 \text{ is } 1), \end{aligned}$$

where e.g. (X_1, X_2, X_3) in the middle term are the three ranks drawn from the $f_{3,3}(i,j,k,\theta)$ model. Show that under neutrality the probability weights (q_2, q_3, q_4), associated with landing skiers in the (2,4), (3,3), (4,2) categories are $(0.25, 0.50, 0.25)$ (xx check this xx), but they may also be adjusted to reflect other purposes. Programme the $p_A(\theta)$ function and plot it.

(f) Use all of this to construct confidence curves for the p_A probability, for mean and for ladies. Construct in particular a version of Figure v.13 (right panel), showing how far the Olympic reality is from its fairness ideal of 0.50, for the men.

(g) (xx round off. cc for the probability that both gold and silver are from A. other weights than 1 for winner 0 for the rest. mention Z sufficient and complete under the model used. is there a use for such a model for ranks, for Wilcoxon, where the sum of ranks is sufficient, and suddenly informative outside the null hypothesis? explore this, briefly. point to Hjort (2017a), and to explanations, related to recuperation time, even more necessary for the men due to more strenuous courses. and also tactics. find out about schemes for landing in A and B. xx)

(h) (xx this to be moved to exercise in Ch2, a little Wilcoxon thing, to illustrate CLT, to get a nonpara test, etc. xx) We follow m individuals, competing with $n-m$ others. When ranking all n, from 1 to n, our group has ranks $X_1, \ldots, X_m$. Under the null assumption that our guys are just as good as the others, the ranks are a random subset of $\{1, \ldots, n\}$. Show that this in particular means $\Pr(X_1 = k) = 1/n$ for $k = 1, \ldots, n$ and $\Pr(X_1 = k, X_2 = \ell) = 1/\{n(n-1)\}$ for $k \neq \ell$. Show that

$$\begin{aligned} \mathrm{E}\, X_i &= \xi = \tfrac{1}{2}(n+1), \\ \mathrm{Var}\, X_i &= \sigma^2 = (1/12)(n^2-1), \\ \mathrm{cov}(X_i, X_j) &= (1/12)\{(n+1)/(n-1)\}(3n^2 - n - 2) - \xi^2, \end{aligned}$$

for $j \neq i$. Then, needing some algebraic pattience for verification, show that for the full rank-sum $Z = X_1 + \cdots + X_m$ we have

$$\mathrm{E}\, Z = \tfrac{1}{2}m(n+1), \quad \mathrm{Var}\, Z = m\sigma^2 + m(m-1)\,\mathrm{cov}(X_1, X_2) = (1/12)(n+1)m(n-m).$$

for semifinals-story, for $n = 6$, for $m = 2, 3, 4$, for normal approximations to Z_A. testing that our guys are from the same distribution as the others. check Lehmann (1975), Lehmann (1950). we invent a parametric model which has Z as sufficient statistic:

$$f(x_1, \ldots, x_m, \theta) = \exp\{\theta(x_1 + \cdots + x_m)\}/K_m(\theta),$$

for ranks $x_1, \ldots, x_m$ inside $\{1, \ldots, n\}$, and with $K_m(\theta)$ the very big sum of all $\exp\{\theta(x_1 + \cdots + x_m)\}$, with $\binom{n}{m}$ terms. can do this via simulation, even when that constant is too impractical to compute. make a separate story on this, with a good dataset. rather than merely testing $F = G$, we give optimal inference for θ, confidence curve for probabilities of interest, etc. xx)

Story v.8 *Who wins? Computing probabilities as a function match time*
Data: `denmark-norway-nov2022-data, handball-data`
Code: `(xx to come, com117* of nilswork22 xx)`
(xx this ought to be good stuff, composed the day after Nor-Den 27-25 November 2022. data1: time points for goals; data2: 117 match results, for correlated Poissons. need to calibrate with what we write elsewhere on Poisson processes. xx) Watching a handball match, the two teams have at time t scored $A(t)$ and $B(t)$ goals. In our continuous excitement we speculate perhaps perplexidly about the final outcome, i.e. $A(60) = A(t) + A'$ and $B(60) = B(t) + B'$. Below we find the dynamically evolving probabilities for team A winning and for team B winning, as a function of time t; see Figure v.14 for how these dramatically panned out for the women's European Championship 2022, with Denmark taking an early lead but Norway prevailing in the end.

(a) Assume the teams are about equally strong, and that goals are scored according to independent Poisson processes with rate $\lambda = 27.00$; this is close to the average number of goals scored by teams in women's Olympic, World, European tournaments. What is the pre-match probability of a draw? What is the most likely result at halftime?

(b) Show that the relevant probabilities, at time t during the match, where $A(t)$ and $B(t)$ have just been observed, are

$$\begin{aligned} p_A(t) &= \Pr(A(t) + A' > B(t) + B') = \Pr(A' - B' > B(t) - A(t)), \\ p_B(t) &= \Pr(A(t) + A' < B(t) + B') = \Pr(A' - B' < B(t) - A(t)), \\ p_D(t) &= \Pr(A(t) + A' = B(t) + B') = \Pr(A' - B' = B(t) - A(t)), \end{aligned}$$

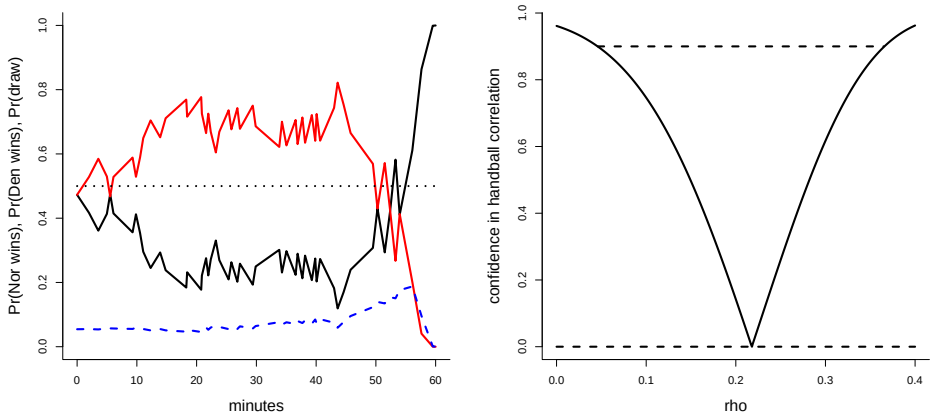


Figure v.14: *Left panel: probabilities that Norway will win, that Denmark will win, or that it will be a draw, as a function of match time, in the women's European Finals November 2022. Right panel: confidence curve* $\mathrm{cc}(\rho)$ *for the dependent Poisson process correlation parameter* ρ*, with point estimate* 0.218 *and 90 percent interval* $[0.046, 0.365]$.

in which A' and B' are independent Poissons with means $\lambda(60-t)$. Find formulae for these probabilities, in terms of sums. Then compute and plot these, from the beginning to the end of the match, for the case of Norway–Denmark in the European 2022 finals; produce indeed a version of Figure v.14, left panel.

(c) When $A(t)$ is a Poisson process with constant rate λ, show that $A(t)$ given $A(60)=m$ is a binomial $(m, t/60)$. Show more generally that with match time $[0,60]$ split into disjoint intervals $C_1,\ldots,C_k$, with lengths $\ell_1,\ldots,\ell_k$, then the goal counts $(A(C_1),\ldots,A(C_k))$ for these intervals have a multinomial distribution $m,(\ell_1/60,\ldots,\ell_k/60)$. For a finished handball match, having observed the $A(t)$ process, explain how a Pearson chi-squared test (see Story vii.1) can be put up to test the constant rate Poisson modelling hypothesis. Carry out such a test, for Norway and for Denmark, in the European 2022 finale, counting the number of goals scored in the six time windows $[0,10],\ldots,[50,60]$.

(d) Another view of the scoring-of-goals processes is as follows. If team A has scored $A(60)=m$ goals, show that the time points $T_1<\cdots<T_m$ at which goals have been scored follows the joint density $m!/60^m$ on the set $t_1<\cdots<t_m$. Explain that this also means that $(T_1,\ldots,T_m)$ behaves as an ordered sample from the uniform distribution on $[0,60]$. Use this again to argue that with $F_m(t)$ the empirical c.d.f. for the data, the process $Z_m(t)=m^{1/2}\{F_m(t)-t/60\}$ is close in distribution to that of $Z(t)=W^0(t/60)$, with W^0 a Brownian bridge; see Ex. 9.23. To check the constant rate Poisson process assumption, therefore, compute and display these Z_m processes, for Norway and Denmark in their 27–25 European finals match. Construct a version of Figure v.15.

(e) There is perhaps a feeling among spectators and handball followers that the top teams to a high degree follow each other during matches; the final scores $A(60), B(60)$ are often close. This motivates Poisson models with positive dependence. For a parameter $\rho\in[0,1]$, consider $A=C+A_0$ and $B=C+B_0$, where A_0,B_0,C are independent

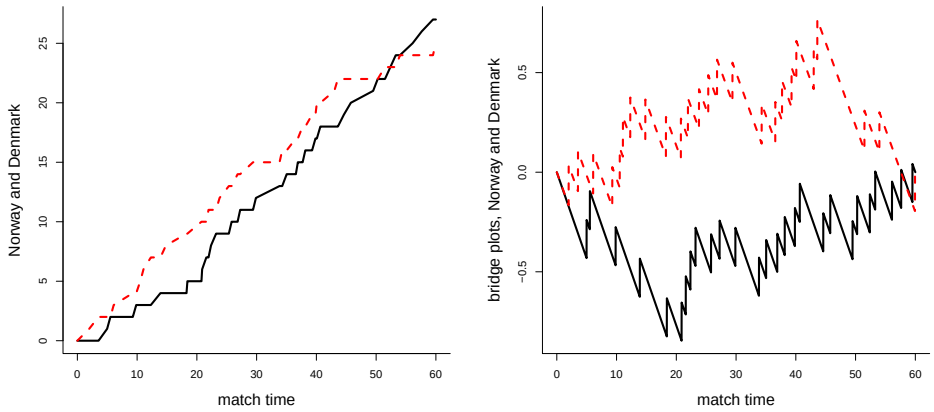


Figure v.15: *Two summary views of the Norway–Denmark European finals 2022. Left panel: the number of goals scored, as a function of match time. Right panel: bridge plots, to assess the Poisson constant rate hypohtesis. These will under that modelling assumption be inside* ± 1.358 *in 95 percent of all cases.*

Poissons, with parameters $\rho\lambda, \rho(1-\lambda), a(1-\lambda)$, and where only (A, B) are observed. Show that A and B are Pois(λ), but dependent, with correlation ρ. Show that

$$\Pr(A=a, B=b) = \sum_{c \le \min(a,b)} g(c, \rho\lambda)\, g(a-c, (1-\rho)\lambda)\, g(b-c, (1-\rho)\lambda),$$

in terms of the point mass function $g(x, \theta)$ for the Poisson with parameter θ. Now access the table of results (y_1, y_2) from 117 top-level women's handball matches (xx described in dataoverview xx). Plot the differences $y_1 - y_2$ divided by standard deviation to argue that matches 33 (Norway vs. Slovenia, 41-18) and 54 (Greece vs. China, 13-33) are clear outliers, and work then with the resulting cleaned table of 115 match results. Programme and graph the profiled log-likelihood function for the ρ parameter, say $\ell_{\text{prof}}(\rho) = \max_{\text{all } \lambda}\{\ell(\rho, \lambda)\}$. Estimate the correlation parameter ρ, and construct a confidence curve $\text{cc}(\rho)$, as with Figure v.14, right panel. This should give the ML estimate $\widehat{\rho} = 0.218$, with 90 percent interval $[0.046, 0.365]$. Thus handball matches at the top international level are positively correlated Poisson processes.

(f) (xx round off. more sophisticated plots for $p_A(t), p_B(t), p_D(t)$, using this dependence model. probably not very different. xx)

Story v.9 *The turn-around operation: from 0-2 to 3-2*
Data: `0-1 and 0.2 at 48th and 52nd, see text`
Code: `(xx to come, com76* of nilswork23)`
In a properly exciting round-of-16 match during the World Cup 2018, Belgium managed to turn 0-2 against Japan into a 3-2 win. This was heralded in media as almost a miracle, the breaking of a 48 year curse, etc.; this had not happened in the World Cup since England saw her 2-0 lead disappear into a 2-3 loss against West Germany in Mexico 1970. Here we shall neutrally assess just how spectacular such an event might be. Consider a match between two essentially equally strong teams A and B. The score

at time t is $(X(t), Y(t))$, with independent Poisson processes with the same rate λ. In a detailed story in Claeskens and Hjort (2008b, Ch. 6), analysing 254 matches to see how FIFA ranking scores may influence these Poisson rates, 627 goals were scored, which means an average of 2.468 goals per match, which we here translate to a common rate of $\lambda = 1.234/90$ per minute, up to match length $T = 90$ minutes (sometimes extended with a few extra minutes for so-called injury time). This little story aims at assessing how small the probability is, for experiencing a match with first has 0-2 and then is turned around to a 3-2 or even better.

(a) What is the probability that the score is still 0-0 after five minutes, and after time t? Plot this probability, for $t \in [0, T]$.

(b) What is the probability that B some time during the match will be leading 2-0 over A? Show that this is

$$p_0 = \int_0^T g_2(s, \lambda) \exp(-\lambda s)\, \mathrm{d}s,$$

with $g_2(s, \lambda) = \lambda^2 s \exp(-\lambda s)$ the Gam$(2, \lambda)$ density, a sum of two Expo(λ). Carry out the integration to find $p_0 = (1/4)\{1 - (1 + 2\lambda T)\exp(-2\lambda T)\}$. This is 0.177. Argue that the frequency of matches where a 2-0 lead will occur, some time during the event, is $2p_0 = 0.353$. – Note that if the teams had been allowed to play on, with T increasing beyond the 90 minutes, the p_0 tends to $1/4$, the chance that when observing two independent undisturbed Poisson processes $X(t)$ and $Y(t)$ over time, with the same intensity, the two first events will occur in the Y process.

(c) Show that in games where team A experiences a 0-2 situation against team B, the random timepoint S where this occurs has probability density,

$$h_2(s, \lambda) = g_2(s, \lambda)\exp(-\lambda s)/p_0 = \frac{\lambda^2 s \exp(-2\lambda s)}{(1/4)\{1 - (1 + 2\lambda T)\exp(-2\lambda T)\}}$$

for $s \in [0, T]$. Show that it peaks at $s_0 = 1/(2\lambda)$, here in about 36 and a half minute. Construct a figure showing this.

(d) Team B now leads 2-0, at time point S, and team A better hurry up. Show that the probability that team A will actually accomplish the 3-2 feat, given that there is a 0-2 time point in the first place, may be expressed as

$$\begin{aligned} p^* &= \int_0^T P(\text{hurry up from } s \text{ to } T \,|\, S = s) h_2(s, \lambda)\, \mathrm{d}s \\ &= \int_0^T G_3(T - s, \lambda) \exp\{-\lambda(T - s)\} h_2(s, \lambda)\, \mathrm{d}s, \end{aligned}$$

with $G_3(T - s, \lambda)$ the cumulative gamma $(3, \lambda)$ distribution function for the sum of three exponential waiting times, evaluated at match time minus s, and $\exp\{-\lambda(T - s)\}$ the probability that team B doesn't score during this remaining time.

(e) Show first that the probability that a Poisson with mean $\lambda(T - s)$ is less than or equal to 2 is $Q(2, \lambda(T - s)) = \exp\{-\lambda(T - s)\}\{1 + \lambda(T - s) + \frac{1}{2}\lambda^2(T - s)^2\}$. Use this to show that the probabiity of experiencing a 0-2 followed by a 3-2 operation is

$$p^* = \int_0^T \{1 - Q(2, \lambda(T - s))\} \exp\{-\lambda(T - s)\} h_2(s, \lambda)\, \mathrm{d}s.$$

(xx check. we find $p^* = 0.014$. xx)

(f) When Belgium see Genki Haraguchi and Takashi Inui score, in the 48th and 52nd minute, they ought to be forgiven for being merely moderately interested in the overall $p^* = 0.014$, but more concerned with the imminent chance that they can still manage, given that they face 0-2 after precisely $s = 52$ minutes. Show that this probability is

$$p^*(s) = G_3(T - s, \lambda) \exp\{-\lambda(T - s)\}.$$

Plot that probability curve, as a function of 0-2 occurrence time s. At time $s = 0$, this is the chance of winning 3-0 or more, namely 3.7 percent, and after 52 minutes, it is about 1.0 percent.

(g) (xx could round off with one or two supplementing questions, using the fifa scores database nils built up for gerda-nils Ch6, to make it statistical too. there $\lambda_{i,j} = h(x_i/x_j, \theta)$, in terms of pre-tournament fifa scores. point finally to magne aldrin and anders løland, their prediction machines at NR, during tournaments. xx)

Story v.10 *The hot hand in basketball.* [xx the code for this story is in `hothandstory.R` xx] Anyone who has ever set foot on a basketball court or watched a few basketball games, have noticed that players sometimes go on streaks, hitting several shots in a row: She or he has the hot hand! There is even a saying in basketball, 'feed the hot hand', meaning that one ought to pass the ball to the player who has the hot hand, and let her or him make an attempt at the basket. During a time out in a game in the 2014–2015 NBA regular season, LeBron James changed the design of plays in order to get the ball to Kevin Love. When a reporter asked James about this change after the game, he James replied: "He had the hot hand, and I wanted to keep going to him." (The Love's-got-the-hot-hand incident is reported in the article Miller and Sanjurjo (2021), which much of the present story builds upon.) The consensus outside of the gym, however, among the allegedly cool headed data crunchers of this world, has for a long time been that the belief in a hot hand is clear evidence of a "massive and widespread cognitive illusion", as stated in Daniel Kahnemann's bestselling book *Thinking Fast and Slow* (Kahneman, 2011). [xx Key articles 'debunking' the hot hand myth are Gilovich et al. (1985), Tversky and Gilovich (1989), and Koehler and Conley (2003). but see also the it' okay in Larkey et al. (1989) and Wardrop (1995). May also point to and use stuff from Story i.3 on Markovian children xx]. (Miller and Sanjurjo, 2018) [xx might also use puzzles from Wissner-Gross (2020) xx]

(a) You make or miss a basket. Let $X_1, \ldots, X_n$ denote n shot attempts by a single player, with $X_i = 1$ indicating a made basket, and $X_i = 0$ a miss. The first difficulty measuring a hot hand is how to define it. The common hot hand intuition of any basketball player and fan is that the probability of making the next shot is higher than what it normally is if all of ones previous $k \geq 1$ shots went in. A sequence of k made shots is called a *hot streak*. This means that if we let $H_i^k = I\{\sum_{j=1}^k X_{i-j} = k\}$ be a hot streak indicator and p_0 be the probability of a make under normal circumstances, the hot hand hypothesis says that $p_{\text{hot}} \coloneqq \Pr(X_j = 1 \mid H_j^k = 1) > p_0$. Some players and fans also think that the hot hand hypothesis implies a cold hand hypothesis according to which the probability of making a basket is lower after k consecutive misses than what it would otherwise be. Let $C_i^k = I\{\sum_{j=1}^k X_{i-j} = 0\}$ with $C_1^k = \cdots = C_k^k = 0$ be cold hand indicators, then the cold hand hypothesis says that $p_{\text{cold}} \coloneqq \Pr(X_j = 1 \mid C_j^k = 1) < p_0$. Though it is less clearly stated, the hot and cold hand hypotheses also say that $\Pr(X_j = 1 \mid X_1 = x_1, \ldots, X_{j-1} =$

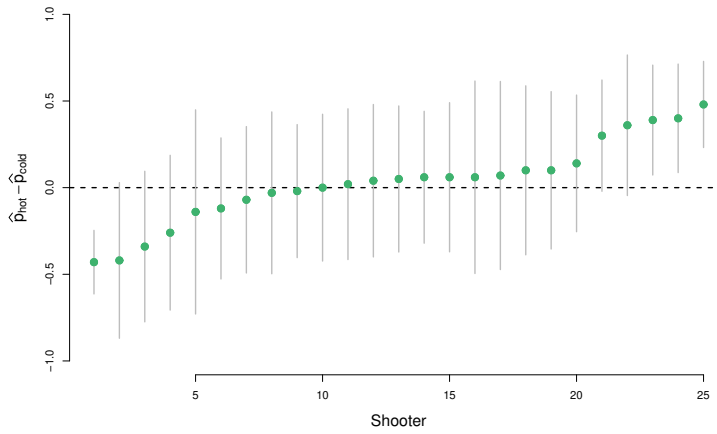


Figure v.16: The differences $\widehat{p}_{\text{hot}} - \widehat{p}_{\text{cold}}$ for each player, along with 95 percent confidence intervals. The estimators are $\widehat{p}_{\text{hot}}$ and $\widehat{p}_{\text{cold}}$ are the estimators from Story v.10(b).

$x_{j-1}) = \Pr(X_j = 1 \mid H_j^k = h_j, C_j^k = c_j)$ for $h_j, c_j \in \{0, 1\}$, and that a player starts a game neither cold nor hot, so $H_1^k = \cdots = H_k^k = 0$ and $C_1^k = \cdots = C_k^k = 0$. Finally, because it is convenient and we have no weighty reason not to, the streak lengths k will be the same for both type of streaks. Based on the above one can device separate hot hand and cold hand hypotheses, and also test the null hypothesis $p_{\text{hot}} = p_{\text{cold}}$ vs. the hot-and-cold alternative $p_{\text{cold}} < p_{\text{hot}}$. This latter hypothesis is perhaps the canonical one to test.

To get going, program a function that takes shots $X_1, \ldots, X_n$ as its input, and returns the hot hand and cold hand indicators $H_1^k, \ldots, H_n^k$ and $C_1^k, \ldots, C_n^k$, respectively; find expression for the maximum likelihood estimators for p_{cold}, p_0, and p_{hot}; and simulate shooting data for some parameter values $p_{\text{hot}} \geq p_0 > p_{\text{cold}}$, n and k of your choosing, and, peeking ahead, look at the statistic $\widehat{p}_{\text{hot}} - \widehat{p}_{\text{cold}}$. What do you notice?

(b) In the classical hot hand study Gilovich et al. (1985), the authors, among other things, asked Cornell University basketball players to participate in a controlled shooting experiment, [xx described on pp. 304–305 in their article xx]. Download the data set `GVT1985_CornellData.csv`, reproduce the plot in Figure (b), and test the null hypothesis $p_{\text{hot}} > p_{\text{cold}}$ for each shooter to see why the hot hand was described as a myth. [xx rewrite xx]

(c) (xx continue (b), something on multiple testing, Bonferroni corrections, etc.. xx)

(d) As a start of trying to understand what goes on in (a), suppose that X_1, X_2, X_3, X_4 are i.i.d. Bernoulli trials with success probability p. We can think of this as the null hypothesis $p_{\text{cold}} = p_0 = p_{\text{hot}} =: p$. Suppose that $k = 1$ and that the estimator $\widehat{p}_{\text{hot}}$ from (a) is used to estimate p_{hot}. Look at Table v.2 and explain what's going on. Can you find an expression for the bias of $\widehat{p}_{\text{hot}}$ as an estimator of p?

(e) We will now see that $\widehat{p}_{\text{hot}}$ undershoots p_{hot} for any of $n \geq 2$ and $k \geq 1$. Let $X_1, \ldots, X_n$ Bernoulli random variables with success probability p. Let $I_k(x) = \{i \in$

seq	shot1	shot2	shot3	shot4	$\sum_{i=1}^4 H_i^4$	$\widehat{p}_{\text{hot}}$
1	0	0	0	0	0	-
2	1	0	0	0	1	0.000
3	0	1	0	0	1	0.000
4	1	1	0	0	2	0.500
5	0	0	1	0	1	0.000
6	1	0	1	0	2	0.000
7	0	1	1	0	2	0.500
8	1	1	1	0	3	0.666
9	0	0	0	1	0	-
10	1	0	0	1	1	0.000
11	0	1	0	1	1	0.000
12	1	1	0	1	2	0.500
13	0	0	1	1	1	1.000
14	1	0	1	1	2	0.500
15	0	1	1	1	2	1.000
16	1	1	1	1	3	1.000
$\mathrm{E}\,\widehat{p}_{\text{hot}}$						0.405

Table v.2: All sequences of length four along with $\widehat{p}_{\text{hot}}$ and $\mathrm{E}\,\widehat{p}_{\text{hot}}$ computed under the assumption that X_1, X_2, X_3, X_4 are i.i.d. Bernoulli(1/2).

$\{1,\dots,n\}$: $H_i^k(x_1,\dots,x_n) = 1\} \subset \{k+1,\dots,n\}$ be the hot streak indices, and let $F = \{(x_1,\dots,x_n)$: $I_k(x)$ is not empty$\}$ be all sequences that contains at least one shot attempted when on a hot streak. In order to show that $\mathrm{E}\,(p_{\text{hot}} \mid F) < p$, suppose that you sample a sequence $(x_1,\dots,x_n)$ from F according to the distribution $\Pr(X_1 = x_1,\dots,X_n = x_n \mid F)$, and then, given $X_1 = x_1,\dots,X_n = x_n$, sample an index J uniformly on $I_k(x_1,\dots,x_n)$, that is $\Pr(J = j \mid X_1 = x_1,\dots,X_n = x_n, F) = 1/|I_k(x_1,\dots,x_n)|$ for $j \in I_k(x_1,\dots,x_n)$, where $|I_k(x_1,\dots,x_n)| = \sum_{i=1}^n H_i^k(x_1,\dots,x_n)$ is the number of elements in $I_k(x_1,\dots,x_n)$. Show that $\Pr(H_J^k X_J = 1 \mid F) = \mathrm{E}\,(\widehat{p}_{\text{hot}} \mid F)$.

(f) In view of the above, in order to show that $\widehat{p}_{\text{hot}}$ is biased for p, it suffices to show that $\Pr(H_J^k X_J = 1 \mid F) < p$. To do so, first write $\Pr(H_J^k X_J = 1 \mid F) = \sum_{j=k+1}^n \Pr(H_j^k X_j = 1 \mid J = j, F)\Pr(J = j \mid F)$, and convince yourself that $\Pr(H_j^k X_j = 1 \mid J = j, F) = \Pr(X_j = 1 \mid J = j, F_j)$ where $F_j = F \cap \{H_j^k = 1\}$. Second, apply Bayes' theorem to the numerator and denominator of the ratio $\Pr(X_j = 1 \mid J = j, F_j)/\Pr(X_j = 0 \mid J = j, F_j)$, and you will see that $\Pr(H_J^k X_J = 1 \mid F) < p$ is implied by the inequality $\Pr(J = j \mid X_j = 1, F_j) < \Pr(J = j \mid X_j = 0, F_j)$ for at least one $j \in \{k+1,\dots,n\}$.

(g) Modify your proof from (f) to show that $\mathrm{E}\,\widehat{p}_{\text{cold}} > p$, where $\widehat{p}_{\text{cold}}$ is the cold hand probability estimator from (a).

(h) In (f) and (g), can you do without the i.i.d. assumption? In particular, can you show that $\mathrm{E}\,\widehat{p}_{\text{hot}} < p_{\text{hot}}$ and $\mathrm{E}\,\widehat{p}_{\text{cold}} > p_{\text{cold}}$ when the $X_1,\dots,X_n$ stem from the process described in (a) and $p_{\text{cold}} < p_0 < p_{\text{hot}}$?

(i) Find the expectation of $\widehat{p}_{\text{hot}}$ when $k = 1$. [xx here we need some more hints, I think xx]. [xx notes xx]

(j) (xx more notes xx) Try to show that the estimator

$$\widetilde{p}_{\rm hot} = \frac{\sum_{i=k+2}^{n}(1-H_{i-1}^k)H_i^k X_i}{\sum_{i=k+2}^{n}(1-H_{i-1}^k)H_i^k},$$

is unbiased for p in the i.i.d. case. [xx well, it is *almost* unbiased I now see xx].

(k) (xx notes xx) . With independent Beta-priors with parameters $(\alpha,\beta) = (a\theta, a(1-\theta))$, with (a_0,θ_0), $(a_{\rm cold},\theta_{\rm cold})$, $(a_{\rm hot},\theta_{\rm hot})$, the posterior expectations (for ith player attempting m_i shots) are

$$\mathrm{E}\,(p_{\rm hot} \mid x_1,\ldots,x_{m_i}) = \frac{a_{\rm hot}}{a_{\rm hot}+\sum_{j=1}^{m_i} H_j^k}\theta_{\rm hot} + \big(1-\frac{a_{\rm hot}}{a_{\rm hot}+\sum_{j=1}^{m_i} H_j^k}\big)\frac{\sum_{j=1}^{m_i} H_j^k X_j}{\sum_{j=1}^{m_i} H_j^k}.$$

Try empirical Bayes and compare to $\widehat{p}_{\rm hot}$.

(l) (xx Try these findings on the three point contest data `threepointcontests.txt` xx)

Story v.11 *BMI for Olympic speedskaters.* Along with other speedskating history enthusiasts, NLH has gathered height and weight data for Olympic speedskating participants, from 1952 to 2018 for $n_M = 1274$ men, and from 1960 to 2018 for $n_W = 907$ women. These give rise to BMI scores.

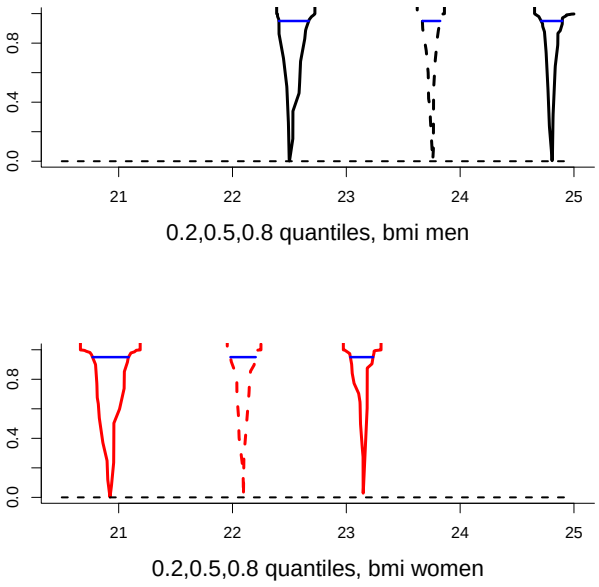


Figure v.17: Confidence curves for the three deciles $F^{-1}(q)$, for levels 0.2, 0.5, 0.8, for the BMI distributions for men (upper panel) and women (lower panel); see Story v.11. 95 percent confidence intervals for the 3 + 3 quantities are indicated with the blue horizontal lines.

(a) (xx first: cc(μ_q) for quantiles 0.2, 0.5, 0.8, for men and women separately. see Figure v.17. xx)

(b) (xx then on to the II-CC-FF story, partly taken from Cunen and Hjort (2022). first $\text{cc}_j(\mu_j)$ for the Olympics no. j. then log-likelihood conversion and a parabola and the max-point. xx)

(c) xx

(d) xx

Notes and pointers

(xx notes and follow-up things for the stories in this chapter. xx)

(xx for Bolt: mention Einmahl and Smeets (2011), Schweder and Hjort (2016, somewhere). xx)

II.vi

Simulated stories

(xx WELL: lots of things to round off and to do, as of 16-January-2025. a partial nils todo list includes: (i) do the beer foam decay processes well and round off. (ii) check if we can invent a simple understandable model for polarisation, with people being pushed from ok pluralism to the boundaries, e.g. inside a $[-1,1]$ scale. try $\exp(\beta h(x))/Z(\beta)$ things. xx)

Story vi.1 *Checking out the CLT.* The Central Limit Theorem, dealt with at length in Ch. 2, says that with $X_1, X_2, \ldots$ i.i.d., with mean μ and standard deviation σ, the distribution of

$$Z_n = (X_1 + \cdots + X_n - n\mu)/(\sqrt{n}\sigma) = \sqrt{n}(\bar{X}_n - \mu)/\sigma \tag{vi.1}$$

becomes close to the standard normal as n increases. We should wonder how close Z_n is to the normal, for different distributions and sample sizes, and such questions may be answered both by mathematics and by simulations.

(a) To illustrate closeness-or-not of Z_n to the standard normal, make the Beta distribution the start distribution, with parameters $(a,b) = (1,5)$; see Ex. 1.19. Display the density of this distribution and compute its mean, variance, and skewness $\gamma_3 = \mathrm{E}\,(X-\xi)^3/\sigma^3$. Show also that $\mathrm{skew}(Z_n) = \gamma_3/\sqrt{n}$.

(b) Your task is now to simulate the high number $\mathrm{sim} = 10^4$ realisations of the variable Z_n, for small to moderate sample size n, which we for Figure vi.1 have taken to be $10, 11, \ldots, 100$. The idea is to see if natural goodness-of-fit tests are able to see that the big simulated dataset, drawn from the exact distribution of Z_n, are not from the standard normal (though close, for growing n). For each n, and for each such simulated dataset $Z^*_{n,1}, \ldots, Z^*_{n,\mathrm{sim}}$, compute the Kolmogorov–Smirnov and Pearson test statistics

$$D_{\mathrm{sim}} = \sqrt{n}\max_{i\le n} |i/n - \Phi(Z^*_{n,(i)})|, \qquad K_{\mathrm{sim}} = \sum_{j=1}^{m} \frac{(N_j - \mathrm{sim}\, p_{0,j})^2}{\mathrm{sim}\, p_{0,j}}.$$

Here $Z^*_{n,(i)}$ is the i-th order statistic, N_j the number of datapoints landing in cell j, and $p_{0,j}$ the standard normal probability for that cell; see Story vii.1. The cells can be constructed as one pleases, but here we have taken $(\Phi^{-1}((j-1)/m), \Phi^{-1}(j/m))$, so that each of these have probability $p_{0,j} = 1/m$ under standard normality.

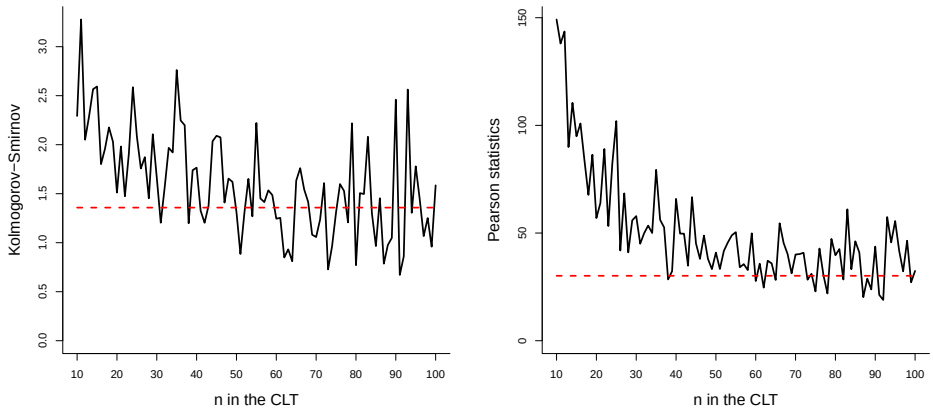


Figure vi.1: *For each* $n = 10, 15, 20, \ldots, 295, 300$*, we have simulated* 10^4 *realisations of* Z_n *of (vi.1), and then computed the Kolmogorov–Smirnov (left panel) and the Pearson chi-squared test statistic* $K_n = \sum_{j=1}^{20}(N_j - 10^4\, p_{0,j})^2/(10^4\, p_{0,j})$ *(right panel). The red horizontal lines are at 1.358 (left panel) and 30.144 (right panel), the the 0.95 points of these statistics under standard normality.*

(c) Construct a version of Fig. vi.1, where we have used $m = 20$ cells for the Pearson test (right panel). It appears that for n as big as around 75, the exact distribution of Z_n is so close to the standard normal that normality tests will typically not pick up the small difference, even with samples of size 10000. You may play with your code using even bigger samples, to check that the goodness-of-fit tests *do* pick up the tiny differences, requiring even bigger n to pass muster.

(d) Try to answer the same type of questions as above, but with yet another goodness-of-fit test for standard normality.

(e) With the code you have written up to carry out the tasks pointed to above, it should be easy to change to other start distributions than the Beta, with other and perhaps larger simulation sizes sim with different sequences of n, etc. Try to find a start distribution for the X_i for which the convergence $Z_n \to_d \mathrm{N}(0,1)$ is rather slow.

Story vi.2 *An infinite weighted sum of Bernoullis.* We start out this story with pieces of clear probability theory, regarding ways of constructing random numbers on the unit interval. Versions of such schemes lead to clear construction recipes, but where it becomes too difficult to determine the precise distributions involved. The generally valid point, illustrated below, is then that simulations are sometimes simple and often useful.

(a) Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli variables with probabilities $\frac12, \frac12$ for 0, 1. Form from these $Y = \sum_{j=1}^{\infty} X_j/2^j$; in other words, the X_j are the decimals in the binary expansion. Show from this definition that Y has mean $\frac12$ and variance $1/12$.

(b) Show that the moment-generating function (m.g.f.) for the partial sum $Y_n = \sum_{j=1}^{n} X_j/2^j$ can be written

$$M_n(t) = \prod_{j=1}^{n} \{\tfrac12 + \tfrac12 \exp(t/2^j)\} = (1/2)^n \prod_{j=1}^{n} \{1 + \exp(t/2^j)\}.$$

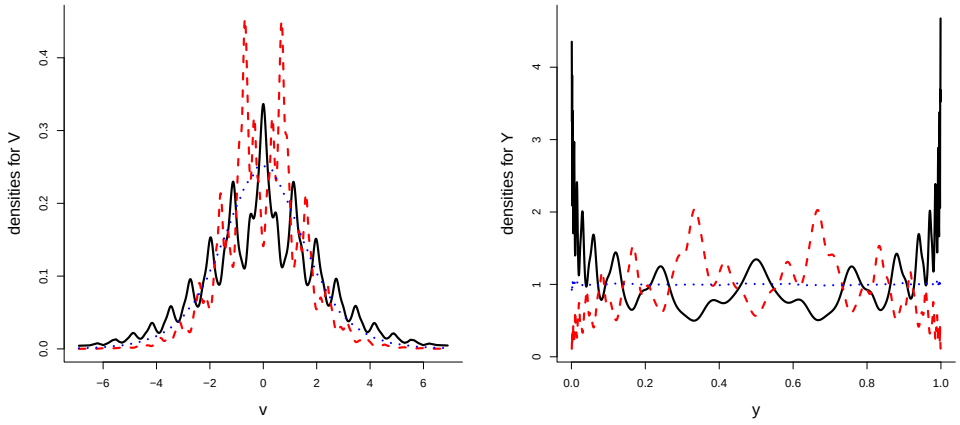


Figure vi.2: *Densities for V (left panel) and for $Y = \exp(V)/\{1+\exp(V)\}$ (right panel), for the model (xx below xx), with $a = 0.40, 0.50, 0.60$ (black, blue, red), using kernel density estimation with half a million simulated outcomes.*

(c) Verify that for any positive x, we have

$$(x^{1/16}-1)(x^{1/16}+1)(x^{1/8}+1)(x^{1/4}+1)(x^{1/2}+1) = x-1.$$

Generalise to $(x^{1/2^n}-1)\prod_{j=1}^n(x^{1/2^j}+1) = x-1$. Going back to Y_n and its m.g.f., show that

$$M_n(t) = \frac{\exp(t)-1}{2^n\{\exp(t/2^n)-1\}}.$$

Prove from this that the infinite sum Y has a uniform distribution on the unit interval. Attempt to demonstrate $Y_n \to_d$ unif also in another way, by viewing Y as a random number inside the unit interval, with the $X_1, X_2, X_3, \ldots$ being its binary digits.

(d) It is clear from the representation above that a uniform Y may be represented as a sum of two independent variables, in several different ways. For any n, show that if $Y^* \sim$ unif$[0,1]$, and $X_1, \ldots, X_n$ are i.i.d. symmetric Bernoulli, then $Y = \sum_{j\le n} X_j/2^j + (\frac12)^n Y^*$ is also uniform on $[0,1]$. Also, write $Y = \sum_{j=1}^\infty X_j/2^j$ as a sum $Y_{\rm odd} + Y_{\rm even}$, summing over odd and even numbers. Simulate a high number of $Y_{\rm odd}$ and $Y_{\rm even}$ to see their densities.

(e) Some formulae become simpler or more illuminating when turning to the uniform on $[-1,1]$. Let X_j be i.i.d. with values ±1 with probabilities $\frac12, \frac12$; such are sometimes called Rademacher variables. Show that $Y = \sum_{j=1}^\infty X_j/2^j$ is uniform on $[-1,1]$. Working with characteristic functions, show that

$$\frac{\sin t}{t} = \prod_{j=1}^\infty \cos\Big(\frac{t}{2^j}\Big).$$

(f) Another take on the problem of demonstrating that Y is uniform is as follows. Use $Y = X_1/2 + \sum_{j=2}^\infty X_j/2^j$ to show that we have the representation

$$Y = \begin{cases} 0 + \frac12 Y_1 & \text{with probability } \frac12, \\ \frac12 + \frac12 Y_2 & \text{with probability } \frac12, \end{cases}$$

where Y_1 and Y_2 have the same distribution as Y. Deduce from this that for the m.g.f. M, we must have $M(t) = \{\frac12 + \frac12 \exp(\frac12 t)\}M(\frac12 t)$. This agrees with $M(t) = \{\exp(t) - 1\}/t$ (xx but can we also prove that this must be the only solution xx).

(g) An extension of the uniform distribution is now to allow the binary digits to be dependent. For some probability parameter a, assume that $X_1, X_2, \ldots$ forms a symmetric Markov chain on $\{0, 1\}$, with transition probabilities

$$\begin{pmatrix} p_{0,0} & p_{0,1} \\ p_{1,0}, & p_{1,1} \end{pmatrix} = \begin{pmatrix} 1-a & a \\ a, & 1-a \end{pmatrix}.$$

We take the first binary digit X_1 in $Y = \sum_{j=1}^\infty X_j/2^j$ to have probabilities $\frac12, \frac12$ for $0, 1$; this is also the equilibrium distribution of such a chain, see Ex. 12.4. Show that

$$\mathrm{E}\, Y = \tfrac12 \quad \text{and} \quad \mathrm{Var}\, Y = \frac{1}{12} + \frac16 \frac{1-2a}{1+2a}.$$

(h) It appears difficult to give a clear formula for this generalisation of the uniform distributiom. Before turning to simulations, show the following. Let $M_0(t) = \mathrm{E}\, \exp(tY_0)$ and $M_1(t) = \mathrm{E}\, \exp(tY_1)$ be the m.g.f.s for Y, given that the $X_1, X_2, \ldots$ chain has started in respectively $X_0 = 0$ and $X_1 = 1$. Show that

$$Y = \begin{cases} 0 + \frac12 Y_0 & \text{with probability } \frac12, \\ \frac12 + \frac12 Y_1 & \text{with probability } \frac12, \end{cases}$$

where Y_0 and Y_1 have distributions reflecting $X_0 = 0$ and $X_1 = 1$. Use this to show that the m.g.f. of Y can be written

$$M(t) = \tfrac12 M_0(\tfrac12 t) + \tfrac12 \exp(\tfrac12 t) M_1(\tfrac12 t).$$

(i) To learn about M_0 and M_1, show that

$$\begin{aligned} M_0(t) &= (1-a)M_0(\tfrac12 t) + a \exp(\tfrac12 t) M_1(\tfrac12 t), \\ M_1(t) &= aM_0(\tfrac12 t) + (1-a) \exp(\tfrac12 t) M_1(\tfrac12 t). \end{aligned}$$

Express $M_0(\frac12 t)$ and $M_1(\frac12 t)$ in terms of $M_0(t)$ and $M_1(t)$.

(j) For Y having a density $f(y)$ on the unit interval, consider the logistic representation $Y = \exp(V)/\{1 + \exp(V)\}$. Show that V has density

$$g(v) = f\Big(\frac{\exp(v)}{1+\exp(v)\}}\Big) \frac{\exp(v)}{\{1+\exp(v)\}^2},$$

and that if we start with $g(v)$ for V, then $f(y) = g(\log\{y(/1-y)\})/\{y(1-y)\}$ is the density for Y. Work in particular this through for the case of Y having the uniform distribution.

(k) For the non-uniform case of $a \neq \frac12$ it appears too difficult to squeeze out clear formulae for the distribution of Y. It is then fruitful to produce versions of Fig. vi.2, by simulating say half a million realisations of Y, for given values of a; that figure shows the result, for $a = 0.40, 0.50, 0.60$. For this purpose we use kernel density estimation techniques from Ch. 13. To avoid the boundary problem, the difficulty of estimating the density $f_a(y)$ for Y coming close to 0 and 1, it pays off to estimate the density $g_a(v)$ for V separately first, using the $Y = \exp(V)/\{1 + \exp(V)\}$ representation of the previous point. Construct a version of Fig. vi.2.

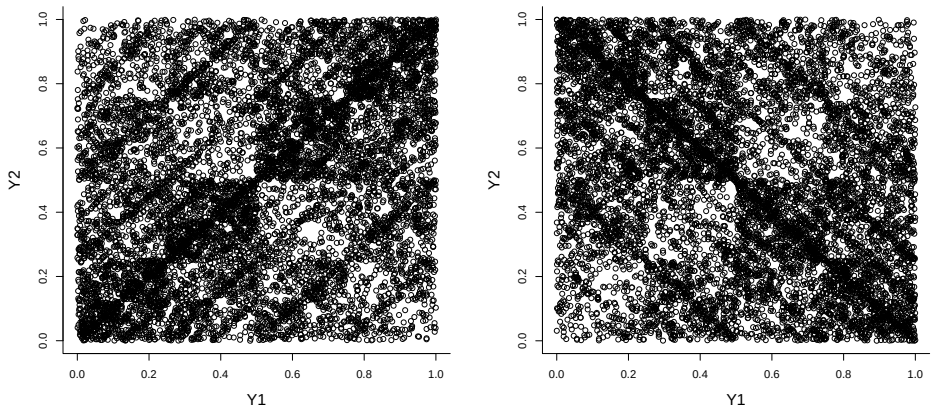


Figure vi.3: *Simulated* (Y_1, Y_2) *from the (vi.2) model, with* 10^4 *pairs, with a positive parameter* b *(left panel) and a negative parameter (right panel). The marginals have uniform distributions.*

(l) We round off this story by using the infinite Bernoulli sum representation to create a distribution for two dependent uniforms. For some $b \in [-\frac{1}{4}, \frac{1}{4}]$, let (X, X') have the distribution with probabilities $\frac{1}{4}+b, \frac{1}{4}-b, \frac{1}{4}-b, \frac{1}{4}+b$ for outcomes $(0,0), (0,1), (1,0), (1,1)$. Show that X and X' both have $\frac{1}{2}, \frac{1}{2}$ probabilities for $0, 1$, and that $\mathrm{corr}(X, X') = 4b$. Then form

$$Y = \sum_{j=1}^{\infty} X_j/2^j \quad \text{and} \quad Y' = \sum_{j=1}^{\infty} X'_j/2^j, \tag{vi.2}$$

where the (X_j, X'_j) are independently drawn from the distribution described. Draw say 10^4 realisations from this joint distribution with uniform marginals, and plot them in a diagram, for a few different values of b. Also show that $\mathrm{corr}(Y, Y') = 4b$. From the examples of Figure vi.3 (left and right panels), can you estimate b?

Story vi.3 *Finding magic squares by MCMC.* "Thirty-four", she said. "Every direction adds up to thirty-four." "Exactly", Langdon said. "But did you know that this magic square is famous because Dürer accomplished the seemingly impossible?" He quickly showed Katherine that in addition to making the rows, columns, and diagonals add up to thirty-four, Dürer had also found a way to make the four quadrants, the four center squares, and even the four corner squares add up to that number. "Most amazing, though, was Dürer's ability to position the numbers 15 and 14 together in the bottom row as an indication of the year in which he accomplished this incredible feat!"

This is from Dan Brown's symbolic 2009 thriller *The Lost Symbol*, and the Dürer's *Melencolia 1* from 1514 is seen in Figure vi.4, left panel. How can we construct, or sample our ways to finding, magic squares, of the Dürer type, with sums of rows, columns, diagonals equal to 34? We may attempt to be clever in other ways than Dürer (1457–1528), exploiting specially designed probability distributions and then ingenious sampling from these. The first task is to set up probability distributions, on the set of all 4×4 squares with the numbers 1-to-16, which favour realisations close to being magic. The second task is to simulate squares, from such a distrribution, hoping that sooner or later

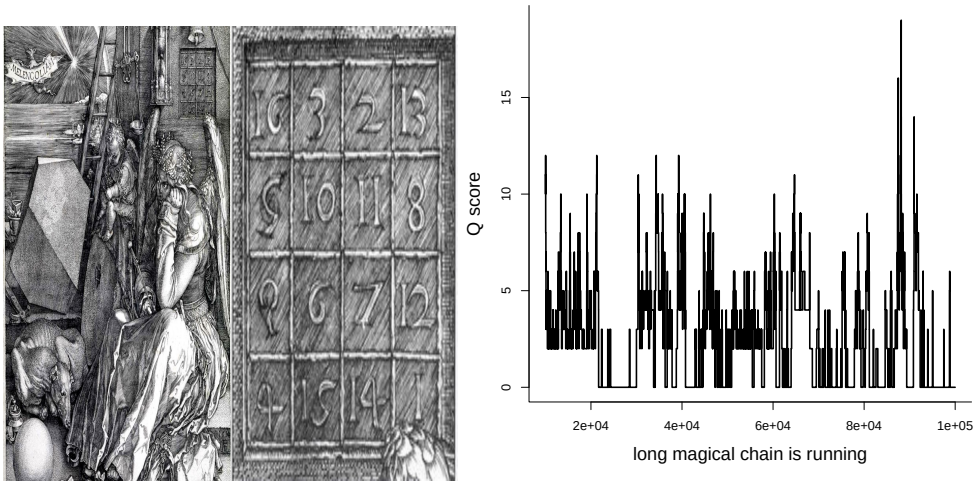


Figure vi.4: *Left panel: part of Dürer's Melancolie I, from 1514. Right panel: the scores $Q(x)$ from the Markov chain of squares, having used $\lambda = 1.234$. It hits zero after 21684 steps (amounting to just a few seconds), having then found the first of many magic squares.*

we will see fully magic squares. (xx to be edited and cleaned, with references to relevant mcmc exercises, in Ch6 and in ch12. point to Hjort (2019a,b). xx)

(a) Show first that if a 4×4 square is to be magic, with the same sum for all rows and columns, then that sum needs to be 34. Next, argue that there are $16! \doteq 2.0923 \cdot 10^{13}$ different ways of placing the 1-to-16 numbers in a 4×4 square. For squares x in this enormous sample space, consider $f(x, \lambda) = k(\lambda)^{-1} \exp\{-\lambda Q(x)\}$, where

$$Q(x) = \sum_{i=1}^{4} |R_i(x) - 34| + \sum_{j=1}^{4} |C_j(x) - 34| + \sum_{d=1}^{2} |D_d(x) - 34|,$$

writing $R_i(x)$, $C_j(x)$, $D_d(x)$ for sums over rows i, columns j, diagonals d. Show that this becomes a genuine probability distribution, with the right choice of $k(\lambda)$. For a positive λ, what are the outcomes with highest probability?

(b) (xx here we point suitably to relevant mcmc exercises in Ch6 and Ch12. and take a bit of care to match notation for the MH things. xx) The idea is to choose a λ and then to sample squares from the $f(x, \lambda)$ distribution. Direct sampling algorithms are unfeasible since the sample space is so enormous. Instead, set up a MCMC, a Markov chain $x_0, x_1, x_2, \ldots$ of squares, in the following fashion. For x_0, start anywhere. To go from x_{t-1} to x_t, choose row-and-column positions (i, j) and (i', j') at random. The proposal $x_{\rm prop}$ for x_t is to switch values, at these two positions; $x_{\rm prop}[i', j'] = x_{t-1}[i, j]$ and $x_{\rm prop}[i, j] = x_{t-1}[i', j']$. Then accept this proposed change with probability

$$\text{pr} = \min(1, \exp[-\lambda\{Q(x_{\rm prop}) - Q(x_{t-1})\}]).$$

That is, we let $x_t = x_{\rm prop}$, if accepted, and $x_t = x_{t-1}$ otherwise. Show that this works, in the sense that the equilibrium distribution for the chain is exactly that of $f(x, \lambda)$.

(c) Implement such a Markov chain of squares, run it for perhaps 10^5 steps. Construct a version of Figure vi.4, right panel, and record a few magic squares found in this fashion. In the actual run associated with this figure, having used $\lambda = 1.234$, the $Q(x)$ first hits zero after 21684 steps, and then yielded the 4×4 square to the left here. Adjusting your code, find similarly a magical 6×6 square, with sums of rows, columns, both diagonals all equal to 111. One of many such magical squares, hit by our MCMC code, is the one to the right here.

9	7	14	4
15	1	6	12
2	16	11	5
8	10	3	13

1	28	24	10	12	36
26	14	15	31	7	18
13	30	19	8	21	20
2	32	23	27	16	11
35	3	25	6	33	9
34	4	5	29	22	17

(d) Play a bit with your code, perhaps attempting different fine-tuning parameters λ. This is a balancing game; describe what happens if λ is set too low, or too large. Modify your code to respect Dürer's choice of having 15 14 in the middle of the last row, and to having sums over the four 2×2 blocks also equal to 34. Produce in this fashion a few other magic Dürer squares.

he played the guitar, the harp, the viola da gamba, and for good measure invented his own glass armonica

(e) Benjamin Franklin (1706–1790), statesman, inventor, scientist, inventor, philosopher, economist, printer, and musician, had the talent to be creatively inventive when he was bored. He must have been a clever doodler and droodler and riddler. Once upon a time he constructed a rather beautiful 8×8 square, with lots of sums equal to 260. As he rather modestly writes in his autobiography (published 1793), "I was at length tired with sitting there to hear debates, in which, as clerk, I could take no part, and which were often so unentertaining that I was induc'd to amuse myself with making magic squares or circles." – We might not be quite as creative or musical as Franklin (1793), but we may construct special probability models and sampling from them, in ways pointed to above. Consider the sample space of all 8×8 squares, filled with numbers 1-to-64; show that the number of elements in this sample space is bigger than the number of atoms in the known universe. By adjusting your MCMC code, as appropriate, running in this big space, solve three problems. (i) Find a square which has sums of each row, each column, and both diagonals, equal to 260. (ii) In addition to these contraints, put in that also the sum of the central 4×4 block, i.e. $x[3:6, 3:6]$, needs to have sum 260. (iii) For an even more magical 8×8 square, set up a probability distribution for outcomes x of the type $f(x) \propto \exp\{-\lambda Q(x)\}$, with $Q(x) = Q_1(x) + Q_2(x) + Q_3(x) + Q_4(x)$ over the four main 4×4 blocks, and with these being nonnegative and zero only for perfect blocks, with sums of rows and columns equal to 130. Our MCMC algorithm found this, after a few milllion iterations:

12	11	51	56	44	31	50	5
34	54	25	17	8	53	24	45
64	10	26	30	21	39	9	61
20	55	28	27	57	7	47	19
36	41	37	16	38	46	3	43
58	14	18	40	42	22	60	6
35	13	23	59	48	29	4	49
1	62	52	15	2	33	63	32

(f) (xx point to such models being used also in other and more statistical applications. find ML estimate for λ, exponential family theory, by solving $\xi(\lambda) = \bar{Q}$, where $\xi(\lambda) =$

$\mathrm{E}_\lambda\, Q(X)$. simulate 10^6 realisations from the Markov chain above, using say $\lambda = 1.234$; sample every 1000th from this chain, taking these 1000 as independent; find the ML. xx)

(g) (xx make a simple addendum, after having set up things in the kortstokkproblemet, in Ch5 and Ch7. a statistician runs an algorithm, taking about 1 minute each time, to find 4×4 squares. how many such magic squares are there in total? need $\mathrm{cc}(N)$ from having found r different such in a total of n runs. true answer is perhaps 880, but no precise answer is known for sizes 6×6 and above. xx)

Story vi.4 *Reconstructing allegedly exponential decay beer foam processes.* In his alleged spare time, German physicist professor Arndt Leike studied the alleged exponential decay over time of beer foam, for as many as $7 + 4 + 4$ beer mugs, for beer makes A, Erdinger Weißbier, B, Augustinerbräu München, and C, Budweiser Budvar. This involved recording the decreasing volume of froth over time (proportional to height, for the cylindrical mugs used), and fitting the observed averages to exponential decay curves. Here we actually raise doubts over his claims, and argue that the decay processes have not stayed quite exponential over time. For beer froth scientists it would have been best if Leike had given the individual decay observations over time; he gave only the means and standard deviations over time, however, for the three makes A, B, C (and has informed us, in private communication, that the individual $7 + 4 + 4$ decay-over-time numbers have been lost). Presenting our arguments therefore involves the non-trivial task of reconstructing these individual decay processes.

his paper Leike (2001) summarising these efforts won him the *Ig Nobel Prize of Physics* for 2002

In more detail, for beer mug i, let $U_i(t)$ be the volume of froth after time t, for $i = 1, \ldots, m$ (with m equal to 7, 4, 4, for the three makes). Leike (2001) gives the mean and standard deviation for these, say $\bar{U}(t)$ and $s(t)$, at the start $t_0 = 0$ seconds, and then after $t_1 = 15, \ldots, t_{14} = 360$ seconds. Our statistical interest lies with the decay, not the actual volume, so it would have been best to have had the individual $V_i(t) = U_i(t)/U_i(0)$, or at least means and standard deviations for these. With Leike not having kept his original data, we translate his summary data to the table given here, with $v(t) = \bar{U}(t)/\bar{U}(0)$, starting at 1, and the similarly scaled $\widehat{\sigma}(t) = s(t)/\bar{U}(0)$ for $t > 0$.

	time	VtA	sigmatA	VtB	sigmatB	VtC	sigmatC
0	0	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
1	15	0.9471	0.0176	0.8429	0.0176	0.8643	0.0235
2	30	0.8765	0.0235	0.7500	0.0176	0.7786	0.0235
3	45	0.8235	0.0235	0.6643	0.0294	0.7143	0.0235
4	60	0.7765	0.0235	0.6071	0.0353	0.6643	0.0235
5	75	0.7353	0.0353	0.5500	0.0353	0.6143	0.0235
6	90	0.7000	0.0235	0.5071	0.0412	0.5714	0.0176
7	105	0.6588	0.0235	0.4643	0.0471	0.5357	0.0176
8	120	0.6294	0.0235	0.4286	0.0471	0.5000	0.0176
9	150	0.5706	0.0235	0.3786	0.0647	0.4429	0.0176
10	180	0.5235	0.0176	0.3143	0.0706	0.3929	0.0235
11	210	0.4882	0.0235	0.2500	0.0529	0.3214	0.0235
12	240	0.4412	0.0235	0.2071	0.0647	0.2500	0.0294
13	300	0.3706	0.0294	0.0929	0.0412	0.1429	0.0294
14	360	0.3059	0.0294	0.0500	0.0294	0.0643	0.0235

(a) Since the $\widehat{\sigma}(t)$ do not differ very much, we may at least initially view the $v(t)$ data as stemming from the model with independent $v(t_j) \sim \mathrm{N}(g(t_j), \sigma^2/m)$, with perhaps

different parametric forms for the mean function $g(t)$. Show that fitting the data to some $g(t, \theta)$, via maximum likelihood, is then equivalent to ordinary least squares, minimising $\sum_{j=1}^{k} \{v(t_j) - g(t_j, \theta)\}^2$. Do this, for the three beer makes, for the exponential $\exp(-\lambda t)$, and construct versions of Figure vi.5 (xx left panel xx). This is essentially what Leike (2001) did. Carry out also related weighted least squares fitting, minimising $\sum_{j=1}^{k} m\{v(t_j) - g(t_j, \theta)\}^2 / \widehat{\sigma}(t_j)^2$, and comment.

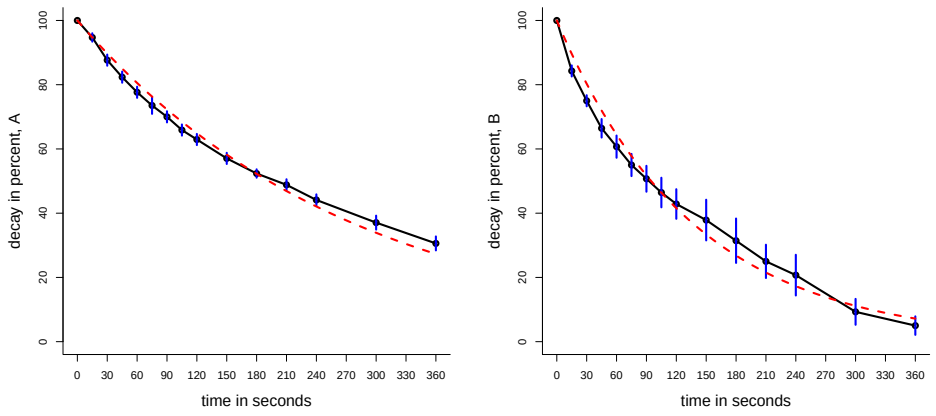


Figure vi.5: *Decay process $v(t)$ over time, for beer makes A (left panel) and B (right panel), with 95 percent intervals, along with the fitted exponential decay curves* $\exp(-\widehat{\lambda}_A t)$ *and* $\exp(-\widehat{\lambda}_B t)$.

(b) The simple exponential $\exp(-\lambda t)$ model assumes that the λ is constant, for different beer mugs of the same make. Show that if λ is allowed to vary, across mugs, according to a $\mathrm{Gam}(a, b)$ distribution, then the decay function becomes rather $\exp\{-a \log(1 + t/b)\}$. Fit these curves to the three sequences of $v(t_j)$ too. Compare the two-parameter (straight exponential) and three-parameter (gamma influenced curve) normal regression models using the AIC of Ch. 11, and comment on your findings.

(c) Before we come to more complicated models, and to the beer foam decay process reconstruction, answer the following. For beer make A, for example, use the normal non-linear regression model with the two-parameter gamma-influenced decay curve to estimate t^*, the time at which precisely half of the initial beer froth has evaporated, along with a 95 percent confidence interval.

(d) Model selection arguments, via e.g. AIC, indicate that there might be better decay models for the beer foam than the simple exponential. In order to come closer to clear alternatives, let us represent the decay processes as $V_i(t) = \exp\{-Z_i(t)\}$, where $Z_i(t)$ needs to be nondecreasing over time. Argue that 'exponential decay', for such random processes, could be reasonably defined by demanding that increments $Z_i(t + h) - Z_i(t)$ ought to have the same distribution, for each time window $[t, t + h]$ of the same length h. In that case the random fraction $V_i(t + h)/V_i(t)$ kept has the same distribution, whether at the start, or middle, or near the end of the decay. Again, since Leike has not given us more than the $v(t)$ and $\widehat{\sigma}(t)$, consider $z(t) = -\log v(t)$, and from these

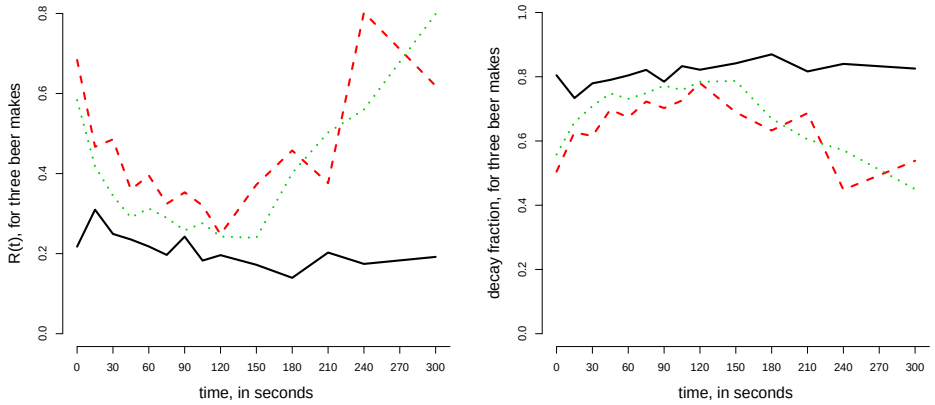


Figure vi.6: *Left panel:* $R(t)$ *over time, the estimated acceleration of* $z(t) = -\log v(t)$, *for beer makes A (full curve), B (slanted), C (dotted). Right panel:* $\exp\{-R(t)\}$ *over time, the estimated fraction kept, per minute.*

compute $R(t_j) = c\{z(t_{j+1}) - z(t_j)\}/(t_{j+1} - t_j)$, for $j = 0, \ldots, 14$. We choose the factor $c = 60$ here, for ease of interpretation; argue that $\exp\{-R(t)\}$ is the fraction of foam kept per minute. If $R(t)$ is about 0.75, for example, then $0.75^3 = 0.422$ is kept after 3 minutes and $0.75^6 = 0.178$ is kept after 6 minutes. Explain that under the exponential decay hypothesis, the $R(t_j)$ ought to stay approximately constant over time. Construct versions of Figure vi.6, with $R(t_j)$ and $\exp\{-R(t_j)\}$, for the three beer makes, and comment.

(e) What we find via the $R(t_j)$ and $\exp\{-R(t_j)\}$ plots indicates that the $Z_i(t+h) - Z_i(t)$ increments are not constant over time, which again means that the decay processes are of a more complicated nature than with exponential decay (more so for beer makes B, C than for A). This motivates *reconstruction of the decay processes* from Leike's summary data. As an introduction question, before we attempt reconstructing $Z_i(t)$ and then $\exp\{-Z_i(t)\}$, assume $Y_1, \ldots, Y_7$ stem from the normal (ξ, σ^2) model, and that you learn the mean 5.55 and standard deviation 2.22 for these. Reconstruct the seven data points, in the sense of drawing them from the normal model, conditional on knowing these two summary numbers. Argue also that in a statistical sense, no vital information has been lost. – Explain next that this normal reconstruction procedure cannot be used for the Leike data, however; attempting to draw normal realisations $V_1(t_j), \ldots, V_m(t_j)$ from the known $v(t_j)$ and $\widehat{\sigma}(t_j)$ will lead to 'wrong values'. They might not be monotone in t_j, and they may fall outside the $(0, 1)$ range.

(f) Explain that modelling $Z_i(t) = -\log V_i(t)$ as independent increments processes, with $Z_i(u) - Z_i(s) \sim \text{Gam}(a(u-s), b)$ over time intervals $[s, u]$, is a coherent construction, in the sense that different binnings of the time axis give the same probabilistic results for $Z_i(t)$. This motivates using $V_i = \exp(-Z_i)$ with gamma increments for the Z_i, and leads to the following reconstruction problem. Consider ratio type data $v_1, \ldots, v_m$ in $(0, 1)$, where a document has reported mean v_0 and standard deviation σ_0 but not the dataset. We wish to model the observations as $V_i = \exp(-D_i)$, with $D_i, \ldots, D_m$ coming

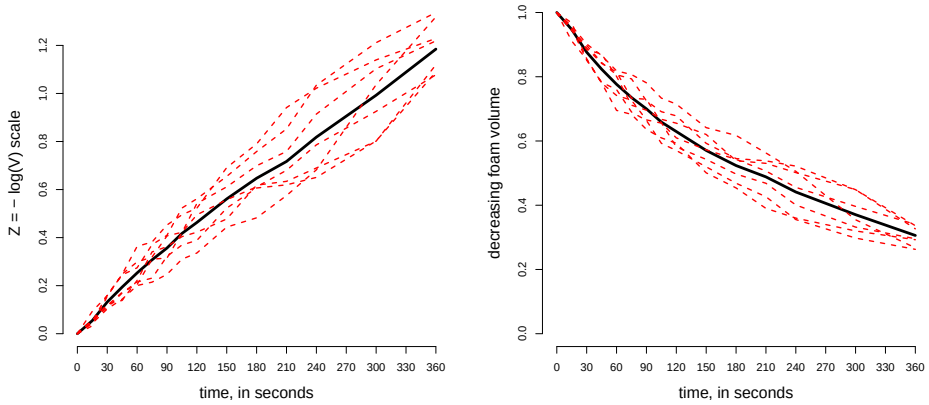


Figure vi.7: *Reconstructed beer foam volume decay processes* $V(t) = \exp\{-Z(t)\}$*, from seven glasses of Erdinger Weißbier, based on the incomplete data information in Leike (2001). Left panel: on the* $Z(t)$ *scale; right panel: on the decaying volume scale. The fat full curves are the averages, from Leike's data, whereas the seven individual paths have been reconstructed through conditional simulation. The simulated paths are such that the* $V_i(t_j)$ *for* $i = 1, \ldots, m$ *have the given mean and standard deviation, at each time point* t_j*.*

from a Gam(a, b) distribution. How can we reconstruct versions of the original dataset, matching these characteristics? Show first that

$$\begin{aligned} f_1(a,b) &= \mathrm{E}\exp(-D_i) = \{b/(b+1)\}^a, \\ f_2(a,b) &= \mathrm{Var}\exp(-D_i) = \{b/(b+2)\}^a - \{b/(b+1)\}^{2a}. \end{aligned}$$

Solving $f_1(a,b) = v_0$ and $f_2(a,b) = \sigma_0^2$ sometimes involves rather large values of a, b, and it is numerically safest to find the solutions (a_0, b_0) in two steps. Show first that solving $f_1(a,b) = v_0$ gives $a(b) = \log(1/v_0)/\log(1+1/b)$, which then leads to the variance formula $g(b) = \{b/(b+2)\}^{a(b)} - \{b/(b+1)\}^{2a(b)}$. Show that g is decreasing, from top value $g(0) = v(1-v)$ (xx check this xx) down to zero. Let this be Step 1 of a data reconstructing strategy, finding (a_0, b_0) matching mean and standard deviation, the solution to $\mathrm{E}\exp(-D_i) = v_0$ and $\mathrm{Var}\exp(-D_i) = \sigma_0^2$ through the numerical scheme pointed to above, by solving $g(b) = \sigma_0^2$ and then computing $a_0 = a(b_0)$. Step 2 is to simulate $D_{1,0}, \ldots, D_{m,0}$ from the Gam(a_0, b_0), and then stretch or fine-tune these, via the two-parameter transformation $D_1 = cD_{i,0}^d, \ldots, D_m = cD_{m,0}^d$, to achieve mean and empirical standard deviation equal to the given v_0 and σ_0. With acceptable start variables, (c_0, d_0) is not far from $(1,1)$. Carry this out for beer A, after $t = 15$ seconds, i.e. for $(v_0, \sigma_0) = (0.9471, 0.0176)$ with $m = 7$.

(g) For the beer foam processes, write again $V_i(t) = \exp\{-Z_i(t)\}$, with the $Z_i(t)$ increasing in t. The reconstruction recipe above gives us $V_i(t_1) = \exp(-D_i)$ at time t_1, based on the reported mean and standard deviation $v(t_1)$ and $\sigma(t_1)$. Next consider the follow-up reconstruction question, where we need $V_1(t_{j+1}), \ldots, V_m(t_{j+1})$, after having successfully reconstructed the $V_1, \ldots, V_m$ processes up to time t_j, as $V_i(t_j) = \exp(-D_i)$, say. We are then after $V_i(t_{j+1}) = \exp(-D_i - E_i) = V_i(t_j)\exp(-E_i)$, say, to match mean $v(t_{j+1})$

and standard deviation $\sigma(t_{j+1})$. Explain how the reasoning above leads to the following required extension recipe. First find a new Gamma distribution parameter pair (a_0, b_0) to match

$$\frac{1}{m}\sum_{i=1}^{m}\exp(-D_i)f_1(a,b) = v(t_{j+1}), \quad \text{or } f_1(a,b) = v(t_{j+1})/v(t_j),$$
$$\frac{1}{m}\sum_{i=1}^{m}\exp(-2D_i)f_2(a,b) = \sigma(t_{j+1})^2, \quad \text{or } f_2(a,b) = \sigma(t_{j+1})^2/\Big\{\frac{1}{m}\sum_{i=1}^{m}\exp(-2D_i)\Big\}.$$

Finding this (a_0, b_0) hence involves the same numerical procedure as for the first time point. Then go on to Step 2: simulate $E_1, \ldots, E_m$ from $\mathrm{Gam}(a_0, b_0)$, and then stretch these to $c_0 E_1^{d_0}, \ldots, c_0 E_m^{d_0}$ to in the end match $v(t_{j+1})$ and $\sigma(t_{j+1})$ for $V_1(t_{j+1}), \ldots, V_m(t_{j+1})$.

(h) Reconstruct the $m = 7$ beer foam processes from the summary statistics $v(t_j)$ and $\sigma(t_j)$ given in the table, for $t_1 < \cdots < t_k$, for beer make A, leading to a version of Figure vi.7. Carry out this also for beer makes B, C.

(i) For your reconstructed decay processes $V_i(t) = \exp\{-Z_i(t)\}$, for $i = 1, \ldots, m$, as in Figure vi.7, fit these via maximum likelihood to the model where the $Z_i(t)$ have independent gamma distributed increments, with $Z_i(t_{j+1}) - Z_i(t_j) \sim \mathrm{Gam}(a_j(t_{j+1} - t_j), b_j)$. Under the exponential decay hypothesis, the (a_j, b_j) parameters should not be unreasonably different. Plot the estimated gamma means a_j/b_j, or perhaps medians $G^{-1}(0.50, a_j, b_j)$, over time. Explain, after repeating all of this several times, for beer makes A, B, C, that the exponential decay hypothesis cannot hold.

Story vi.5 *From falling ill to having recovered.* Mr Jones has fallen ill. Thanks to a clear diagnosis and correct medicine, along with his own good constitution, he'll soon enough have fully recovered, however. In fact he reports that every day he is about ten percent better than the day before. When asked to make this optimistic statement a bit more precise, he explains that on his scale of wellbeing, normal health corresponds to level 2.0, and this particular illness to level 1.0. His health index is $H_0 = 1.0$ on the day when he fell ill, but then $H_m = (1 + \delta_1) \cdots (1 + \delta_m)$ after m days, where he experiences the δ_j as i.i.d. with mean 0.10 and standard deviation 0.03. So when will he have fully recoverd?

(a) Clearly precise answers, about the day M where H_m first exceeds 2.0, depends on the particularities of the distribution of the δ_j. One modelling possibility, which captures the essence of the setup here, is to take $1 + \delta_j = \exp(y_j)$, with the y_j i.i.d. from an appropriate $\mathrm{Gam}(a, b)$. Show that

$$\mathrm{E}\,\delta_j = \Big(\frac{b}{b-1}\Big)^a - 1, \quad \mathrm{Var}\,\delta_j = \Big(\frac{b}{b-2}\Big)^a - \Big(\frac{b}{b-1}\Big)^{2a}.$$

(b) Find the Gamma distribution parameters to match $\mathrm{E}\,\delta_j = 0.10$ and $\mathrm{Var}\,\delta_j = 0.03^2$. (xx check this with care: Answer: $(12.151, 127.991)$. xx) Produce a version of Figure vi.8, displaying densities for his health index after 2, 4, 6 days.

(c) Find and present the full distribution for M, the number of days it takes Mr Jones to be fully recovered. What is the most probable number of days needed for his recovery, based on the 'about ten percent better each new day' starting assumption?

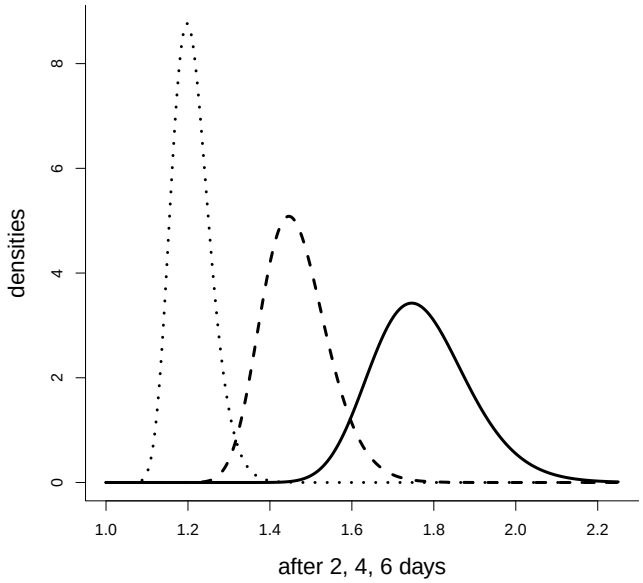


Figure vi.8: The densities for Mr Jones's health index, after 2, 4, 6 days; he has fully recovered when this index reaches 2.0.

(d) Investigate a few other possibilities, regarding the distribution of the δ_j, but still keeping the $\mathrm{E}\,\delta_j = 0.10$ and $\mathrm{Var}\,\delta_j = 0.03^2$ starting information. Are the gamma-based results reached fairly robust?

(e) (xx then generalised to a full log-gamma process. do the ProcMod selling point. pointer to Ch. 15. and be specific about simulation to illustrate and to check plausibility of assumptions. xx)

Story vi.6 *Causal inference and potential outcomes.*
Code & Data: `someRfile.R`, `somePyfile.py`
Many if not most empirical questions posed and claims made in economics, political science, sociology, medicine, and epidemiology are causal. A recent example is, 'Boys should start school a year later than girls', as argued by the economist Richard Reeves (Reeves, 2022b,a). Reeves' argument is causal. Notice that Reeves does not merely make the descriptive claim that 'boys starting school a year later do better than the boys starting school with kids their age', which would just be a statement about a empirical association. Rather, Reeves' normative and thefore causal claim is that a boy starting school a year later (called redshirting) would do better in life than if *that same boy* were to start school with children his age. For a reasearcher intending to test Reeves' hypothesis, the challenge is that we don't observe the *counterfactual*, that is, we never observe a boy both being redshirted and not being redshirted.

One way of resolving the fundamental problem of causal inference is to conduct a randomised experiment. The random assignment of treatment would assure that differences in the outcome can be attributed to the effects of the treatment. Randomised experiments are, however, perhaps more often than not, not feasible. Who will ran-

domly redshirt their child?, for example. This is where the theory of *causal inference* for observational data comes in (or the 'credibility revolution', as it is called (Angrist and Pischke, 2010; Ashworth et al., 2021)). At the heart of the theory of causal inference lies an intense scrutiny of various ways of exploiting sources of random or as-good-as-random variation in the treatment. The point is to leverage so-called natural experiments, that is, empirical settings where the observational data – looked at and used the right way – may mimic or approximate a bona fide randomised experiment. The theory of causal inference consists of a language for talking about causality, and of statistical tools tailored to help tease out causal information from observational data. In this story we introduce this language, which is the theory of potential outcomes (Rosenbaum and Rubin, 1983; Holland, 1986; Imbens and Rubin, 2015); and, by way of simulations, we aim to get a feel for when and why some of the more popular research designs work.

(a) All notions of causality involves the notion that by changing the treatment, we can change the likely value of the outcome. The potential outcomes theory formalises this idea in that there, for one and the same unit, exists parallel worlds, where each world corresponds to a value of the treatment. When treatment, say $X_i = \{0, 1\}$, is binary, one associates two random variables $Y_i(0)$ and $Y_i(1)$ to each unit. Here, $Y_i(0)$ is the outcome in the world where the ith unit is not treated, while $Y_i(1)$ is the outcome in the world where the ith unit is treated. The individual level causal effect of treatment is then some contrast of the two potential outcomes, of which the most obvious and common one is

$$Y_i(1) - Y_i(0),$$

Since we, the data collectors, only live in one of the two parallel worlds, we only get to see $Y_i = Y_i(X_i) = X_i Y_i(1) + (1 - X_i)Y_i(0)$. Inference on the difference above is therefore impossible, a fact known as the *fundamental problem of causal inference* (Holland, 1986, p. 947). With a sample of units, however, we do observe the potential outcome under treatment for some units, and the potential outcome under control for others. This may make it possible to say something about various averages of the difference above. For example, the average treatment effect $\text{ATT} = \text{E}\left(Y_i(1) - Y_i(0)\right)$; the average treatment effect on the treated $\text{ATT} = \text{E}\left(Y_i(1) - Y_i(0) \,|\, X_i = 1\right)$; or the other conditional average treatment effects, $\text{CATE}(z) = \text{E}\left(Y_i(1) - Y_i(0) \,|\, Z_i = z\right)$, for some covariate (vector) Z_i. Describe situations where the ATE, the ATT, or the CATE might be the estimand of interest.

(b) The ubiquitous challenge with causal inference is *confounders*. A confounder is a variable that affects *both* the treatment and the outcome, thereby inducing a correlation between the treatment and the outcome that may not be due changes in the treatment *causing* changes in the outcome. This is where the correlation-is-not-causation adage, uttered in every stat101 course, comes from. A small example of confounding is given in the R-script below.

```
n <- 10^3
z <- rnorm(n) # the confounder
x <- 1*(z >= 0) # the treatment
a <- 1.23; b <- 2.34
y <- a + b*z + rnorm(n,0,1) # the outcome
summary(lm(y ~ x)) # coeff on x is significant
summary(lm(y ~ x + z)) # coeff on x is zero
```

In this script, changing the treatment `x` by one unit does not lead to any change in the outcome `y`. Nevertheless, as the first regression of this R-script illustrates, `x` and `y` are highly correlated. As the second regression of this R-script illustrates, however, the problem of confounding dissappers if all confounders are observed and can be controlled for. With observational data, however, it is rather bold to confidently assume that all confounders are observed, and thus can be controlled for.

Now that we have used the terms observational and experimental, we should be more precise about what we mean with these terms. It all comes down to the *assignment mechanism*. The assignment mechanism is, roughly, the mechanism by which treatment and control (assuming a binary treatment) is allocated, or choosen, or determined, in a population. For example, in the R-script above, $\Pr(\mathtt{x}_i = 1 \,|\, \mathtt{z}_i) = \Pr(\mathtt{z}_i \geq 0) = 1 - \Phi(\mathtt{z}_i)$ for $i = 1, \ldots, n$, completely describes the assignment mechanism. The key difference between the two types of studies, is that in a *randomised experiement* the assignment mechanism is controlled by the researcher, thus fully known to her; while in an *observational study* the assignment is unknown.

Suppose that we have independent data $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ for some outcome Y_i, treatment indicators X_i, and vectors of covariates Z_i. The potential outcomes theory tells us to regard these observed outcomes as $Y_i = XY_i(1) + (1 - X_i)Y_i(0)$ for $i = 1, \ldots, n$. This means that, in the background, there are independent replicates $\{Y_i(0), Y_i(1), X_i, Z_i\}_{i \leq n}$ of a vector $(Y(0), Y(1), X, Z)$. Can we, with these data, make inference on any causal estimand? Suppose that these data stem from an experimental setup where it is decided that n_t units will undergo treatment, and $n_c = n - n_t$ will be assigned to the control group. We draw balls from a hat containing n_t red balls and n_c blue balls. If the first ball is red, the unit with label 1 is assigned to treatment, of the second ball is blue, the unit labelled 2 is assigned to the control group, and so on. Describe the assignment mechanism of the experiment, explain that

$$\{Y(0), Y(1)\} \perp\!\!\!\perp X \quad \text{and} \quad X \perp\!\!\!\perp Z, \tag{vi.3}$$

and show that $\widehat{\text{ATT}}_n = n_t^{-1} \sum_{i=1}^n X_i Y_i - n_c^{-1} \sum_{i=1}^n (1 - X_i) Y_i$ is unbiased for the ATE. Show also that

$$\text{Var}(\widehat{\text{ATE}}_n) = \frac{1}{n} \Big\{ \frac{\text{Var}\, Y(1)}{p} + \frac{\text{Var}\, Y(0)}{1 - p} \Big\},$$

where $p = n_t/n$ is the shared of treated units. The take away here, which you probably knew already, is that with a randomised experiment of this kind, we don't need to control for anything. Practically, the Z_is do not enter into our estimation of the ATE.

(c) With observational data the assumption in (vi.3) is quite unrealistic. Think of a sociological study where the treatment is $X = I\{\text{higher education}\}$ and the outcome is $Y = \text{salary}$ There are certainly innumerable background variables confounding this relationship. In other words, the treated units may differ systematically from the units that are not treated. Okay, so suppose that Z is a vector containing all these background variables. Then, comparing like with like, all else equal, *ceteris paribus*, etc., means that for a given value of $Z = z$, the world is basically performing a randomised experiment of X on Y. More formally, if Z really contains all confounders of the X–Y relationship, it is reasonable to assume that

$$\{Y(0), Y(1)\} \perp\!\!\!\perp X \,|\, Z. \tag{vi.4}$$

If we couple this with the assumption $\Pr(X = 1 \mid Z = z) > 0$ for all z, then treatment assignment is said to be *strongly ignorable*. Let $\{Y_i(0), Y_i(1), X_i, Z_i\}_{i \le n}$ be independent replicates of $(Y(0), Y(1), X, Z)$, and assume that (vi.4) holds. The following R-script provides an example.

```
n <- 10^3
z <- rnorm(n,1,1); u <- 0.5*(z - 1) + sqrt(1 - 0.5^2)*rnorm(n)
x <- 1*(u >= 0) # the treatment indicator
a <- 1.23; b <- 2.34; gamma0 <- 3.45 ; gamma1 <- 4.56
# The potential outcomes
y0 <- a + gamma0*z + rnorm(n,0,1)
y1 <- a + b + gamma1*z + rnorm(n,0,1)
# We observe x,z and
y <- x*y1 + (1 - x)*y0
```

What is the ATE in this example? What is the ATT? More generally, suppose the following model for the potential outcomes

$$Y(0) = Z^{\mathrm{t}}\gamma_0 + \sigma_0\varepsilon(0), \quad \text{and} \quad Y(1) = \beta + Z^{\mathrm{t}}\gamma_1 + \sigma_1\varepsilon(1), \tag{vi.5}$$

where $\varepsilon(0)$ and $\varepsilon(1)$ are independent of X (so that (vi.4) holds). Find an unbiased estimator for the ATE, and compute its variance.

(d) Trouble arises because we seldom observe all confounders. If Z are the confounders we observe and U are the unobserved confounder, then a common state of affairs is that

$$\{Y(0), Y(1)\} \perp\!\!\!\perp X \mid Z, U. \tag{vi.6}$$

How may we then draw causal conclusions about the causal effect of X? This is where various natural experiments might, if we are lucky, come to the rescue. We first look at the *instrumental variable* design in a simple setting. Suppose that the following model for the potential outcomes

$$Y(0) = \alpha + \gamma_0 U + \varepsilon(0), \quad \text{and} \quad Y(1) = \alpha + \beta + \gamma_1 U + \varepsilon(1),$$

where $U \in \mathbb{R}$, and $\varepsilon(0), \varepsilon(1) \perp\!\!\!\perp U$, so (vi.6) is satisfied. Suppose that U is a confounder related to the treatment X by $X \sim \text{Bernoulli}(p(a + bU))$, where $p(\eta) = 1/\{1 + \exp(\eta)\}$. Find expression for the ATE, the ATT, and the CATE(z).

In the R-script below we simulate data $\{(Y_i(X_i), X_i)\}_{1 \le i \le n}$ from this model, giving U a standard normal distribution. [xx fix this xx]

(e) In (a) we made a subtle assumption: The potential outcomes of the ith unit does not depend on the treatment status of any other unit. In the causality literature, this is known as the stable unit treatment value assumption, or SUTVA. A potential outcomes model that does not make this assumption is one in which, in a population of n units, all 2^n different assignments of treatment amount to potential worlds. For example, if $n = 2$ we have the potential outcomes $Y_i(0,0), Y_i(0,1), Y_i(1,0)$, and $Y_i(1,1)$, corresponding to both units being treated, unit 1 not being treated and unit 2 being treated, and so on, for $i = 1, 2$. Situations where the potential outcome of one unit might depend on the treatment status of other units are known as spillover effects. Give an example of an empirical situation where such spillover effects might be more likely than not to occur (i.e., a situation where SUTVA does not hold). We'll make the stable unit treatment value assumption throughout this story.

Story vi.7 *Finding your counterfactual cousin.* (xx polish and round off. choose perhaps a different figure, where the angles are more clearly not orthogonal. use only one n, but invent a different right panel figure, with different push factors $\widehat{c}$. xx) You're part of life's multitude of logistic regressions. You hope to achieve A, for which there are relevant data for a flock of other individuals, associated with the probability

$$\Pr(A \,|\, x) = H(x^{\mathrm{t}}\beta), \quad \text{with } H(u) = \frac{\exp(u)}{1+\exp(u)},$$

in terms of the individual's covariate vector $x = (x_1, \ldots, x_p)^{\mathrm{t}}$. Your current estimate, since you are equipped with your own covariate vector x_0, is $\widehat{p}(x_0) = H(x_0^{\mathrm{t}}\widehat{\beta})$, with $\widehat{\beta}$ the standard maximum likelihood (ML) estimate obtained from logistic regression analysis of the available dataset. In the case of $\widehat{p}(x_0)$ being disappointingly low, the question is *how to change*, getting from your x_0 to a better x_{new}, with say $\Pr(A \,|\, x_{\mathrm{new}}) \geq p_0 = 0.90$.

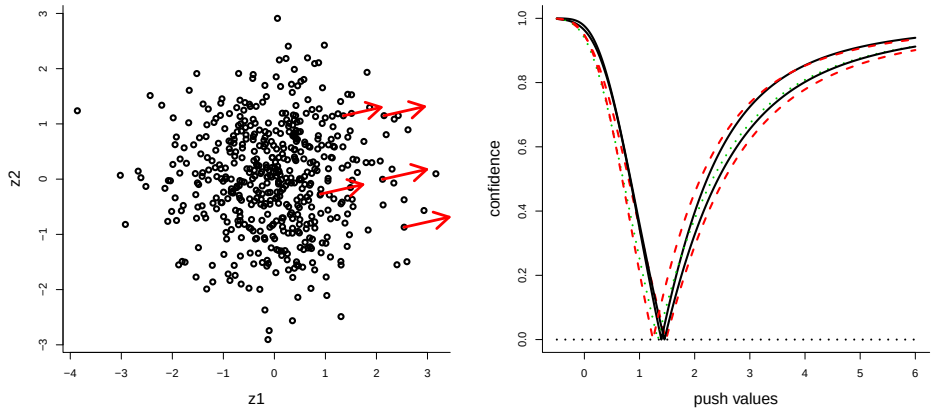


Figure vi.9: *The logistic regression model* $H(\beta_0 + \beta_1 x_1 + \gamma_1 z_1 + \gamma_2 z_2)$ *is used, with parameters* $0.25, 0.44, 0.66, 0.22$. *The covariates are centred, with means zero, and are here taken as independent standard normals. For a sample of size* $n = 500$, *the left panel shows the* (z_1, z_2), *with arrows indicating how five individuals with estimated* $\Pr(A \,|\, x, z)$ *in the range* $[0.80, 0.85]$ *need to move in order to attain the 0.90 probability. The right panel shows confidence curves* $\mathrm{cc}(\mathrm{pu}_j)$ *for the five push factors.*

(a) Show that the formal answer to this question is to achieve $x_{\mathrm{new}}^{\mathrm{t}}\beta \geq d_0$, with $d_0 = H^{-1}(p_0) = \log\{p_0/(1-p_0)\}$, like 2.197 for a hoped for $p_0 = 0.90$.

(b) God, give us grace to accept with serenity the things that cannot be changed, courage to change the things which should be changed, and the wisdom to distinguish the one from the other. We therefore change the notation slightly, sorting covariates into a set of x_j which cannot be changed and another set of z_j which at the outset can be changed. We write the regression model as $p(x,z) = H(x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma)$, with regression coefficients β and γ of dimensions p and q. When (x_0, z_0) is changed to $(x_{\mathrm{new}}, z_{\mathrm{new}}) = (x_0, z_0 + u)$, explain that this leads to new predictor $x_0^{\mathrm{t}}\beta + (z_0+u)^{\mathrm{t}}\gamma$. Then use Cauchy–Schwarz to show that the best you can do is to take $u = \mathrm{pu}\,\gamma$ proportional to γ, yielding

$$x_{\mathrm{new}}^{\mathrm{t}}\beta + z_{\mathrm{new}}^{\mathrm{t}}\gamma = x_0^{\mathrm{t}}\beta + z_0^{\mathrm{t}}\gamma + \mathrm{pu}\,\gamma^{\mathrm{t}}\gamma;$$

you are to push $z_{0,j}$ up if the γ_j coefficient is positive, and push it down if it is negative.

(c) If you need $\Pr(A \,|\, x, z) \geq 0.90$, explain that your strategy should be to push your z_j (well, if possible) until $(z_0 + \mathrm{pu}\,\gamma)^{\mathrm{t}}\widehat{\gamma}$ is high enough, and that this means

$$\widehat{\mathrm{pu}} = (d_0 - x_0^{\mathrm{t}}\widehat{\beta} - z_0^{\mathrm{t}}\gamma)/\|\widehat{\gamma}\|^2.$$

Hence $(x_{\mathrm{new}}, z_{\mathrm{new}}) = (x_0, z_0 + \widehat{\mathrm{pu}}\,\widehat{\gamma})$ is your estimated counterfactual cousin, an inspiring individual hopefully not too far away from yourself in the space of covariates, and who has the envisaged alluring 0.90 chance of achieving A in *her* life. Indeed show that $\widehat{p}(\widehat{x}_{\mathrm{new}}, z_{\mathrm{new}}) = 0.90$.

(d) The practical $\widehat{z}_{\mathrm{new}}$ is really an estimate of the correct underlying

$$z_{\mathrm{new}} = z_0 + \mathrm{pu}\gamma, \quad \text{with } \mathrm{pu} = (d_0 - x_0^{\mathrm{t}}\beta - z_0^{\mathrm{t}}\gamma)/\|\gamma\|^2.$$

To assess how precise $\widehat{\mathrm{pu}}$ is, write $\alpha = (\beta, \gamma)$ for the true parameter vector, and start from

$$\begin{pmatrix} \sqrt{n}(\widehat{\beta} - \beta) \\ \sqrt{n}(\widehat{\gamma} - \gamma) \end{pmatrix} \to_d Z = \begin{pmatrix} Z_b \\ Z_c \end{pmatrix} \sim \mathrm{N}_r(0, J^{-1}),$$

as per standard theory for logistic regressions; see Ex. 5.45. Here J is the probability limit of $\widehat{J}$, the normalised Fisher information matrix. We shall also need $\widehat{Q}$, the $q \times q$ lower-right submatrix of $\widehat{J}^{-1}$. Explain first that

$$\sqrt{n}(x_0^{\mathrm{t}}\widehat{\beta} + z_0^{\mathrm{t}}\widehat{\gamma} - x_0^{\mathrm{t}}\beta - z_0^{\mathrm{t}}\gamma) \to_d x_0^{\mathrm{t}}Z_b + z_0^{\mathrm{t}}Z_c,$$

and show also that

$$\sqrt{n}(\|\widehat{\gamma}\|^2 - \|\gamma\|^2) \to_d \sum_{j=1}^{q} 2\gamma_j Z_{c,j} = 2\gamma^{\mathrm{t}}Z_c.$$

Then apply the delta method to show that there is zero-mean limiting normal distribution

$$\sqrt{n}\Big(\frac{d_0 - x_0^{\mathrm{t}}\widehat{\beta} - z_0^{\mathrm{t}}\widehat{\gamma}}{\|\widehat{\gamma}\|^2} - \frac{d_0 - x_0^{\mathrm{t}}\beta - z_0^{\mathrm{t}}\gamma}{\|\gamma\|^2}\Big) \to_d L.$$

Work further with the limit variable to reach

$$L = \frac{1}{\|\gamma\|^2}(-x_0^{\mathrm{t}}Z_b - z_0^{\mathrm{t}}Z_c) - \frac{d_0 - x_0^{\mathrm{t}}\beta - z_0^{\mathrm{t}}\gamma}{\|\gamma\|^4}2\gamma\mathrm{t}Z_c = -\frac{1}{\|\gamma\|^2}(x_0^{\mathrm{t}}Z_b + z_0^{\mathrm{t}}Z_c + 2\mathrm{pu}\,\gamma^{\mathrm{t}}Z_c).$$

Show that this limit distribution variance is

$$\tau^2 = \frac{1}{\|\gamma\|^4}\begin{pmatrix} x_0 \\ z_0 + 2\mathrm{pu}\,\gamma \end{pmatrix}^{\mathrm{t}} J^{-1} \begin{pmatrix} x_0 \\ z_0 + 2\mathrm{pu}\,\gamma \end{pmatrix},$$

and explain how this quantity can be estimated from the data.

(e) (xx writing out the Wald tests, $W = \widehat{c}/(\widehat{\tau}/\sqrt{n})$. xx)

(f) The distribution of ratios $\widehat{\rho} = \widehat{a}/\widehat{b}$ are often skewed, making delta method type normal approxmations vulnerable. It is often better to construct confidence intervals and confidence distributions for the underlying ratio $\rho = a/b$ by working with $\widehat{a} - \rho\widehat{b}$,

particularly when $(\widehat{a}, \widehat{b})$ is close to having a binormal distribution, as with the Fieller problem exercises in Ch. 7 (xx point to exercises xx). To this end, write first $\widehat{\beta} = \beta + Z_{n,b}/\sqrt{n}$ and $\widehat{\gamma} = \gamma + Z_{n,c}/\sqrt{n}$, where $(Z_{n,b}, Z_{n,c}) \to_d (Z_b, Z_c)$ defined above. For any candidate value pu, work with

$$\begin{aligned} A_n(\mathrm{pu}) &= \sqrt{n}\Big\{d_0 - x_0^{\mathrm{t}}\widehat{\beta} - z_0^{\mathrm{t}}\widehat{\gamma} - \mathrm{pu}\sum_{j=1}^{q}(\widehat{\gamma}_j^2 - \widehat{\kappa}_j^2/n)\Big\} \\ &= -x_0^{\mathrm{t}}Z_{n,b} - z_0^{\mathrm{t}}Z_{n,c} - 2\,\mathrm{pu}\sum_{j=1}^{q}\gamma_j Z_{n,c,j} - \mathrm{pu}\sum_{j=1}^{q}(Z_{n,c,j}^2 - \widehat{\kappa}_j^2)/\sqrt{n}, \end{aligned}$$

writing $\widehat{\kappa}_j^2$ for the diagonal elements of $\widehat{Q}$. Explain that $A_n(\mathrm{pu})$ is approximately normal with mean zero, and variance close to

$$V_n(\mathrm{pu}) = Var\,\{x_0^{\mathrm{t}}Z_b + (z_0 + 2\mathrm{pu}, \gamma)^{\mathrm{t}}Z_c\} = \begin{pmatrix} x_0 \\ z_0 + 2\mathrm{pu}\,\gamma \end{pmatrix}^{\mathrm{t}} J^{-1} \begin{pmatrix} x_0 \\ z_0 + 2\mathrm{pu}\,\gamma \end{pmatrix}.$$

Inserting $\widehat{\gamma}$ and $\widehat{J}$ for γ and J gives the variance estimator $\widehat{V}_n(\mathrm{pu})$. Now explain that all of this leads to the statement that $A_n(\mathrm{pu})^2/\widehat{V}_n(\mathrm{pu})$ is approximately a χ_1^2, and to the confidence curve

$$\mathrm{cc}_n(\mathrm{pu}) = \Gamma_1(A_n(\mathrm{pu})^2/\widehat{V}_n(\mathrm{pu})).$$

(g) Your task is now to use the following description to construct a version of Figure vi.9 (and you may use `set.seed(11)` for making an exact replicate). The data are generated with centred covariates x_1, z_1, z_2, here taken as independent and standard normal, and with logistic regression models $H(\beta_0 + \beta_1 x_1 + \gamma_1 z_1 + \gamma_2 z_2)$, with true values $0.25, 0.44, 0.66, 0.22$, and sample size $n = 500$. The left panel shows (z_1, z_2), with five individuals selected from the segment of those with estimated $p(x_1, z_1, z_2)$ inside $[0.80, 0.85]$. Estimate their pushes $\widehat{\mathrm{pu}}_j$, and construct arrows, as shown, indicating how they need to move from current positions (x_0, z_0) to (x_0, z_{new}), in order to have estimated $p(x, z_{\mathrm{new}}) = 0.90$. For the right panel, construct confidence curves $\mathrm{cc}_n(\mathrm{pu}_j)$ for the push factors.

Story vi.8 *Restoration of blurred images.* Blurred images, where an underlying true image has been observed with statistical noise, can be statistically restored, or estimated. This prototypically needs (i) a model for the true image, perhaps involving prior knowledge, landmarks, degree of contextual smoothness, etc.; and (ii) a model for how the observation process has been working, with the degree of noise, etc. Here we tell a rather simplified story of this type, involving a simple one-dimensional image or chain $x = (x_1, \ldots, x_n)$, where the x_i are 0s and 1s, and an observed image-with-noise $y = (y_1, \ldots, y_n)$. Once the mechanics of this proptotype setup are well understood, various generalisations, primarily from one to two dimensions, can be put up and worked with.

(a) Suppose there is a model $\pi(x)$ for the real image, meant to drive the degree of contextual continuity, how long the typical stretches of 0s and 1s are. Suppose next that the blurring process, leading to the observed $Y = (Y_1, \ldots, Y_n)$, has $Y_1, \ldots, Y_n$ independent given the image, with a density $f(y\,|\,0)$ if $x_i = 0$ and $f(y\,|\,1)$ if $x_i = 1$. Show that

$$\pi(x\,|\,y) \propto \pi(x_1, \ldots, x_n)\prod_{i=1}^{n} f(y_i\,|\,x_i).$$

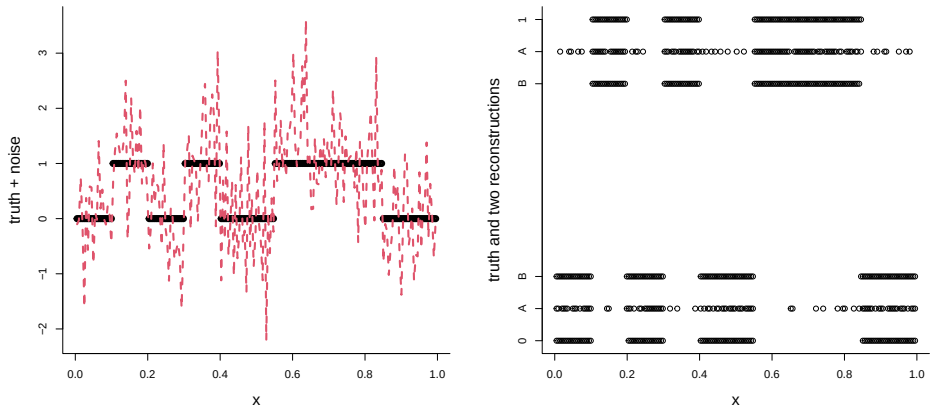


Figure vi.10: *Left panel: the true image, size $n = 200$, plotted normalised to the unit interval (horizontal fat line segments), as 0s and 1s, with the noisy observations y_i on top (slanted lines). Right panel: again the true image, placed at 0 and 1, along with restoration A, the direct one, and restoration B, the contextual one. The error rate is reduced from 0.195 to 0.010.*

This is a probability distribution over the finite space $\{0,1\}^n$ of all the 2^n possible chains.

(b) To make this more concrete, suppose the two densities, for the 0s and the 1s, are normal, with $Y_i \,|\, x_i \sim \mathrm{N}(x_i, \sigma^2)$, with σ the degree of statistical noise. Show that $f(y\,|\,1)/f(y\,|\,0) = \exp\{(y - \frac{1}{2})/\sigma^2\}$. With the simple direct classifier, estimating x_i to be 1 if $f(y_i\,|\,1) \geq f(y_i\,|\,0)$, and vice versa, show that leads to $x_i^* = 1$ if $y_i \geq \frac{1}{2}$ and $x_i^* = 0$ if $y_u < \frac{1}{2}$. For this method, determine the error rate for the two cases, say $\mathrm{err}_0 = \Pr(\text{say } 1 \,|\, x_i = 0)$ and $\mathrm{err}_1 = \Pr(\text{say } 0 \,|\, x_i = 1)$. They are indeed equal here, to $\Phi(-\frac{1}{2}/\sigma)$. Explain that if σ is small, error rates will be low, and there might not be a need for improved restoration methods.

(c) We now introduce a simple contextual model for the x chain; more elaborate models can be built and worked with in similar fashions. With $H(x) = \sum_{i=1}^{n-1} I(x_i = x_{i+1})$, let $\pi(x) = Z(\beta)^{-1} \exp\{\beta H(x)\}$. The normalisation constant $Z(\beta)$ is a sum over 2^n terms, but we may manage without it, as we shall see. For this model, show that

$$\begin{aligned} \Pr(x_i = 1 \,|\, (x_{i-1}, x_{i+1}) = (1,1)) &= \exp(2\beta)/\{1 + \exp(2\beta)\}, \\ \Pr(x_i = 0 \,|\, (x_{i-1}, x_{i+1}) = (1,1)) &= 1/\{1 + \exp(2\beta)\}. \end{aligned}$$

Explain that a positive β means more spatial continuity in the chain, whereas a negative β would lead to a high number of changes from 0 to 1 and from 1 to 0. The value $\beta = 0$ corresponds to a random salt-and-pepper chain, with no continuity.

(d) Show that the probability distribution for the chain, given the observations, is

$$\pi(x \,|\, y) \propto \exp\{\beta H(x)\} \exp\{-\tfrac{1}{2} \sum_{i=1}^{n} (y_i - x_i)^2/\sigma^2\}.$$

A natural idea is now to simulate chains from this distribution. If we succeed, we may read off the most probable x_i for each position i. Using the Metropolis MCMC algorithm (xx

pointer to MCMC section in Ch6 xx), explain that the following simulation scheme will work, in the sense that the chain of chains $x^* = (x_1^*, \ldots, x_n^*)$ will converge in distribution to precisely $\pi(x \,|\, y)$. One starts anywhere in the space of chains, with the following recipe from getting to $x^{\rm old}$ to $x^{\rm new}$: (i) draw a proposal $x^{\rm prop}$ from some proposal distribution, symmetric in $(x^{\rm old}, x^{\rm prop})$; (ii) accept it, with probability $\text{pr}_{\rm accept} = \min(1, r)$, with

$$r = \frac{\pi(x^{\rm prop} \,|\, y)}{\pi(x^{\rm old} \,|\, y)} = \frac{\exp\{\beta H(x^{\rm prop})\} \exp\{-\frac{1}{2}\sum_{i=1}^n (y_i - x_i^{\rm prop})^2/\sigma^2\}}{\exp\{\beta H(x^{\rm old})\} \exp\{-\frac{1}{2}\sum_{i=1}^n (y_i - x_i^{\rm old})^2/\sigma^2\}}.$$

Note that the normalisation constant $Z(\beta)$ has disappeared. Operationally, this scheme leads to $x^{\rm new} = (1 - {\rm ok})\, x^{\rm old} + {\rm ok}\, x^{\rm prop}$, where ok is an indicator for the proposal being accepted, e.g. computed as ok $= I(u \leq \text{pr}_{\rm accept})$, with u drawn from the uniform.

(e) The algorithms simplify and run faster when the proposal $x^{\rm prop}$ does not alter $x^{\rm old}$ much. A simple scheme is to pick an index i randomly, inside $\{2, \ldots, n-1\}$ to avoid boundary issues, and then propose to switch $x_i^{\rm old}$ to $x_i^{\rm prop} = 1 - x_i^{\rm old}$. Show that leads to the following simplification for the ratio r above:

$$\begin{aligned} r &= \exp\{\beta G(x_{i-1}, x_i, x_{i+1})\} \exp[\tfrac{1}{2}\{(y_i - x_i^{\rm old})^2 - (y_i - x_i^{\rm prop})^2\}/\sigma^2] \\ &= \exp\{\beta G(x_{i-1}, x_i, x_{i+1})\} \exp\{(x_i^{\rm prop} - x_i^{\rm old})(y_i - \tfrac{1}{2})/\sigma^2\}, \end{aligned}$$

in which $G(x_{i-1}, x_i, x_{i+1})$ is equal to

$$\begin{aligned} H(x^{\rm prop}) - H(x^{\rm old}) = {} & I(x_{i-1}^{\rm prop} = x_i^{\rm prop}) + I(x_i^{\rm prop} = x_{i+1}^{\rm prop}) \\ & - I(x_{i-1}^{\rm old} = x_i^{\rm old}) - I(x_i^{\rm old} = x_{i+1}^{\rm old}). \end{aligned}$$

The crucial point here is that one avoids having to compute quantities at all positions in the chain, for each new proposal.

(f) Figure vi.10 gives an illustration of these methods. We have started with a given true image $x_{\rm true} = (x_1, \ldots, x_n)$, of length $n = 200$, plotted normalised to the unit interval as 0s and 1s, and then added noise, with $y_i = x_i + \varepsilon_i$ (left panel), with ε_i being i.id. $\text{N}(0, \sigma^2)$, here with $\sigma = 0.75$. Create such a situation in your computer, and then simulate a Markov chain of chains, per the recipe above. For this illustration we have used $\beta = 2.22$; the restoration code may of course be used for different values of this parameter. With loss function $L(x, \widehat{x}) = n^{-1} \sum_{i=1}^n I(\widehat{x}_i \neq x_i)$, show that the optimal solution is taking $\widehat{x}_i$ equal to the most probable value of 0 and 1, in the distribution $\pi(x_i \,|\, y)$. Explain how this may be read off from a long MCMC run; $\widehat{x}_i = 1$ if 1 is seen more than half the time in position i, and correspondingly $\widehat{x}_i = 0$ if 1 is seen less than half the time in that position. Implement this, and create a version of Figure vi.10 (right panel). It gives the crude context-free x^* as restoration A (here with error rate 0.195), and the spatial continuity model based restoration $\widehat{x}$ as restoration B (here with error rate 0.010).

(g) Consider the quantity $\alpha = n^{-1} \sum_{i=1}^n I(x_i = 1)$, the percentage of 1s in the 0-1 image. For the simulated dataset, compute first the crude estimate $\widetilde{\gamma} = n^{-1} \sum_{i=1}^n I(\widehat{x}_i = 1)$, based on the Bayesian restoration. Show that the actual Bayes estimate is rather $\widehat{\gamma} = n^{-1} \sum_{i=1}^n \Pr(x_i = 1 \,|\, \text{data})$. Compute and plot these probabilities, then give a histogram of the distribution of γ given data, and compare the two estimates of γ.

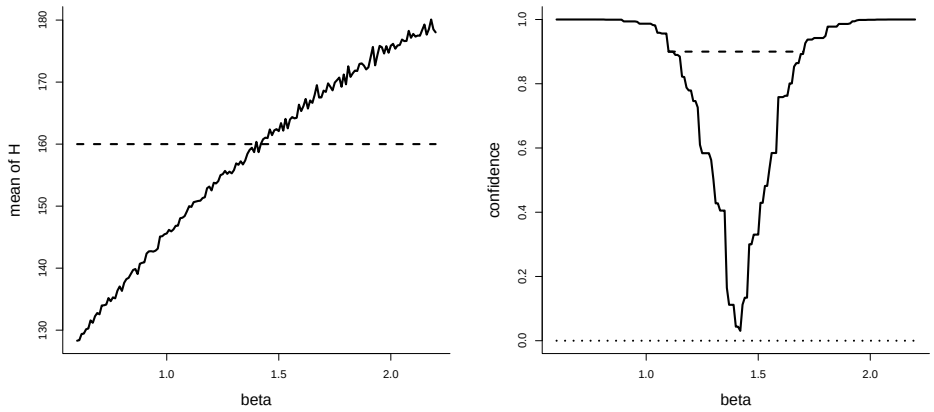


Figure vi.11: *Left panel:* $\xi(\beta)$ *to find ML,* $\widehat{\beta} = 1.42$. *Right panel: optimal confidence ruve* $\mathrm{cc}(\beta)$, *via MCMC simulations and isotonic repair. The 90 percent confidence interval is* $[1.11, 1.70]$.

(h) The $\pi(x) \propto \exp\{\beta H(x)\}$ model has a simple structure, but is not easy to handle since the normalisation constant $Z(\beta)$ is unfeasible. We use the occasion to show how the ML estimate $\widehat{\beta}$ can be computed, in such situations, via simulations. Assume that a certain image of size $n = 200$ has $H(x_{\mathrm{obs}}) = 160$. Show that the ML estimate is the solution to the equation $\xi(\beta) = H(x_{\mathrm{obs}})$, where $\xi(\beta) = \mathrm{E}_\beta\, H(x)$. Now devise a scheme to estimate $\xi(\beta)$, via simulations. This admittedly takes an MCMC chain of chains for each of a grid of β values, but is feasible. Created indeed a version of Figure vi.11, left panel, with $\xi(\beta)$ computed via simulations; the ML value $\widehat{\beta} = 1.42$ is read off from this. Show also, via the theory of Ch. 7 that there is an optimal confidence distribution here, $C(\beta) = \Pr_\beta(H(x) > H_{\mathrm{obs}}) + \frac{1}{2}\Pr_\beta(H(x) = H_{\mathrm{obs}})$. We compute this via the same simulations, and then add on isotonic repair, as per Ex. 7.6. Produce a version of Figure vi.11, right panel; there we have used $1000 + 5 \cdot 10^4$ simulations for each of the MCMCs, then discarding the first 1000. The 90 percent confidence interval is $[1.11, 1.70]$, for the envisaged $H_{\mathrm{obs}} = 160$ for a chain of length $n = 200$.

(i) (xx round off. the Bayes word does not have to be used, but of course the model $\pi(x)$ has the feature of being a prior for the unknown parameters $(x_1, \ldots, x_n)$. about two-dimensional versions being very similar, though requiring a more careful bureaucratic book-keeping code. briefly about β, an Ising type parameter; ideally one would train on real images to elicit that value. xx)

Notes and pointers

(xx notes and follow-up things for the stories in this chapter. xx)

(xx Ising model. Besag, Ripley. extensions and modifications when the image blurring does not operate with the y_i independent, but context there too; Hjort and Mohn. xx)

II.vii

Miscellaneous stories

Story vii.1 *Karl Pearson 1900: goodness-of-fit and the chi-squared.* In 1900, Karl Pearson (1857–1936) published *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* in *Philosophical Magazine, Series 5*. This is indeed a remarkably elaborate and informative title for a journal article, and rightly so. (i) He invents a very useful general test, to check whether a probability vector is equal to a set of specified values; (ii) he shows that the test statistic can be approximated with a new distribution, which is the first-ever published chi-squared distribution, which conveniently does not depend on the specified probability vector, but only the number of boxes under consideration; and (iii) he develops logically sound arguments for when should keep one's theory, and when one should reject it. In yet other words, he invents the notion of statistical testing, via a test statistic, which he shows has a limit distribution, and he almost touches on p-values. In one of perhaps several nutshells, Pearson (1900) builds a full apparatus *to test a given theory*.

Here we go through the ideas and details for the Pearson statistic. Let $N = (N_1, \ldots, N_k)$ be a multinomial vector, with n independent draws for k given boxes, and probability vector $p = (p_1, \ldots, p_k)$; see Ex. 1.5. A simple example to point to is to roll your die n times, count the numbers $(N_1, \ldots, N_6)$ of the different outcomes 1, 2, 3, 4, 5, 6; if your die is fair, this is a multinomial vector with $p = (1/6, \ldots, 1/6)$. The Pearson statistic is

$$K_n = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} = \sum_{j=1}^k \frac{(\text{obs}_j - \text{exp}_j)^2}{\text{exp}_j} = \sum_{j=1}^k \frac{n(\widehat{p}_j - p_j)^2}{p_j},$$

with the familiar estimators $\widehat{p}_j = N_j/n$ and squared ratios $r_j = (\text{obs}_j - \text{exp}_\text{j})/\text{exp}_j^{1/2}$, involving 'observed' and 'expected' numbers. If the die you roll many times is not a completely fair one, the Pearson statistic may be used to detect this, i.e. that some of the r_j deviate too much from zero, and that the real p is not equal to $(1/6, \ldots, 1/6)$. As part of what goes on in Story vii.2 we use the Pearson statistic, and relatives, to check whether the first million digits of π have a uniform distribution on $0, 1, \ldots, 9$.

(a) We have seen in Ex. 2.48 that there is full vector convergence in distribution $X_n = \sqrt{n}(\widehat{p} - p) \to_d X \sim \text{N}_k(0, \Sigma)$, with $k \times k$ variance matrix $\Sigma = D - pp^\text{t}$, writing D for the diagonal matrix with $p_1, \ldots, p_k$ in its diagonal. Show that indeed $\sum_{j=1}^k X_j = 0$. For any

linear combination $\phi = a^{\mathrm{t}}p = \sum_{j=1}^k a_j p_j$, with estimator $\widehat{\phi} = a^{\mathrm{t}}\widehat{p} = \sum_{j=1}^k a_j \widehat{p}_j$, show that $\sqrt{n}(\widehat{\phi} - \phi) \to_d \mathrm{N}(0, \tau^2)$, where $\tau^2 = a^{\mathrm{t}}\Sigma a = \sum_{j=1}^k a_j^2 p_j - (\sum_{j=1}^k a_j p_j)^2$.

(b) Show that $K_n = \sum_{j=1}^k X_{n,j}^2/p_j \to_d K = \sum_{j=1}^k X_j^2/p_j$, with the $X \sim \mathrm{N}_k(0, \Sigma)$ above. This is 'the main job' (now accomplished); the rest of the story is to demonstrate that this K has a χ^2_{k-1} distribution. Show, directly, that $\mathrm{E}\, K = k - 1$.

(c) For the $(k-1) \times (k-1)$ submatrix Σ_0 of Σ, corresponding to the first $k-1$ elements $X_0 = (X_1, \ldots, X_{k-1})^{\mathrm{t}}$ of X, show that $\Sigma_0 = D_0 - p_0 p_0^{\mathrm{t}}$, where D_0 is diagonal with $p_0 = (p_1, \ldots, p_{k-1})^{\mathrm{t}}$ on its diagonal. Use algebraic results from Ex. 3.44 to demonstrate that K may be expressed as $X_0^{-1}\Sigma_0^{-1}X_0$, and explain that this proves that Pearson reached more than a hundred years ago, but with other words and symbols, that $K \sim \chi^2_{k-1}$.

(d) An alternative to the classic K_n is

$$K_n' = \sum_{j=1}^k \frac{(N_j - np_j)^2}{N_j} = \sum_{j=1}^k \frac{(\mathrm{obs}_j - \mathrm{exp}_j)^2}{\mathrm{obs}_j} = \sum_{j=1}^k \frac{n(\widehat{p}_j - p_j)^2}{\widehat{p}_j},$$

i.e. using observed and not expected in the denominator. Show that K_n' and K_n must have identical limit distributions; hence $K_n' \to_d \chi^2_{k-1}$ too.

(e) Another option for testing a given value p for a multinomial is to use maximum likelihood and Wilks testing, from Ch. 5. With $\ell_n(p)$ the log-likelihood function, with $\ell_{n,\max}$ its maximum and $\ell_{n,0}$ the value at the null hypothesis, show that $D_n = 2(\ell_{n,\max} - \ell_{n,0}) = 2n\sum_{j=1}^k \widehat{p}_j \log(\widehat{p}_j/p_j)$. Explain that Wilks theory from Ch. 5 implies $D_n \to_d \chi^2_{k-1}$ at the fixed p. To show that there is a close connection between D_n and K_n, start from the Kullback–Leibler distance

$$\sum_{j=1}^k p_j \log \frac{p_j}{\widehat{p}_j} = \sum_{j=1}^k -p_j \log\Big(1 + \frac{\widehat{p}_j - p_j}{p_j}\Big) \doteq \sum_{j=1}^k \Big\{-p_j \frac{\widehat{p}_j - p_j}{p_j} + \tfrac{1}{2} p_j \Big(\frac{\widehat{p}_j - p_j}{p_j}\Big)^2\Big\},$$

and show that $D_n - K_n \to_{\mathrm{pr}} 0$.

(f) When using $K_n = \sum_{j=1}^k (N_j - np_{j,0})^2/(np_{j,0})$ to test the hypothesis $p = p_0$, study also the local power, at nearby alernative position $p_j = p_{0,j} + \delta_j/\sqrt{n}$. Show that $X_n \to_d X \sim \mathrm{N}_k(\delta, \Sigma)$, and that $K_n \to_d \chi^2_{k-1}(\lambda^2)$, a noncentral chi-squared with $\lambda^2 = \sum_{j=1}^k \delta_j^2/p_j$. Check Story vii.2 to see this applied. (xx just a bit more, making CD, similar for Story iii.8. testing $\lambda = 0$, $C(\lambda) = \Pr_{p_0 + \delta/\sqrt{n}}(K_n \geq K_{n,\mathrm{obs}}) \doteq 1 - \Gamma_{k-1}(K_{n,\mathrm{obs}}, \lambda^2)$. pointmass at zero. may read off 95 percent interval for λ. xx)

(g) (xx one more theme rounding this off. closeness to ML, estimator determined by $\sum_{j=1}^k (N_j/p_j(\theta))u_j(\theta) = 0$. use this in Geissler data stories. xx)

(h) (xx check and point to Story ii.7. xx) Consider two multinomial vectors $M = (M_1, \ldots, M_k)$ and $N = (N_1, \ldots, N_k)$, with comparable probability vectors $p = (p_1, \ldots, p_k)$ and $q = (q_1, \ldots, q_k)$. How can we naturally test the hypothesis H_0 that $p = q$? Writing n_1 and n_2 for the two sample sizes, with sum n, write $\widehat{p}_j = M_j/n_1$ and $\widehat{q}_j = N_j/n_2$. Explain first that

$$\sqrt{n}(\widehat{p} - p) \to_d (1/c_1^{1/2})X \sim \mathrm{N}_k(0, \Sigma_1), \quad \sqrt{n}(\widehat{q} - q) \to_d (1/c_2^{1/2})Y \sim \mathrm{N}_k(0, \Sigma_2),$$

where $\Sigma_1 = D_1 - pp^{\rm t}$, $\Sigma_2 = D_2 - qq^{\rm t}$, 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. Under H_0, then, show that $\sqrt{n}(\widehat{p} - \widehat{q}) \to_d (1/c_1 + 1/c_2)^{1/2} Z$, with $Z \sim {\rm N}_k(0, D - pp^{\rm t})$. Deduce that

$$K_{n_1,n_2} = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{k} \frac{(\widehat{p}_j - \widehat{q}_j)^2}{\widehat{r}_j} \to_d \chi^2_{k-1},$$

as long as $\widehat{r}_j \to_{\rm pr} p_j$ for each j; it is natural to take $\widehat{r}_j = c_1\widehat{p}_j + c_2\widehat{q}_j = (n_1 X_j + n_2 Y_j)/(n_1 + n_2)$.

(i) (xx perhaps room and need for this little Anhang too. xx) the KP is neutral, all cells equally important. could have weights $w_1, \ldots, w_k$ reflecting importance over the different cells. show

$$K_{n,w} = \sum_{j=1}^{k} w_j \frac{(N_j - np_j)^2}{np_j} \to_d K_w = \sum_{j=1}^{k} w_j X_j^2/p_j.$$

the mean of the limit variable is $\sum_{j=1}^k w_j(1-p_j)$. distribution can be simulated. power gets better in certain directions. for the thing where we test uniformity of $u_1, \ldots, u_n$, but go to $v_i = u_{(i+1)}/u_{(i)}$ and then to $w_i = {\rm Beta}(v_i, i, 1)$, can use weights 1 for the two end cells. this test can pick up behaviour of ratios v_i being often closer to 0 or 1 than under uniform conditions. limit is $X_1^2/p_1 + X_k^2/p_k$, mean $2 - (p_1 + p_2)$ under the null.

Story vii.2 *Decimals of* π. The decimals of π have fascinated mathematicians and amateurs endlessly (and some can recite several thousands of decimals faultlessly). One intriguing perspective is that the digits really appear to behave fully randomly, as i.i.d. variables with probabilities $0.1, \ldots, 0.1$ for the ten digits $0, 1, \ldots, 9$.

(a) Here are the decimal counts among the first million such digits of π, not including the 3. start, alongside what we term Pearson residuals, from Story vii.1, which you are asked to compute below. These are $r_j = (N_j - np_{j,0})/(np_{j,0})^{1/2} = (N_j - 10^5)/10^{5/2}$, here with $n = 10^6$ and $p_{j,0} = 0.10$. Plot them, and carry out goodness-of-fit tests, in particular using the Pearson test $K_n = \sum_{j=0}^9 r_j^2$. Find corresponding tables on the net, with bigger sample sizes than a million, and carry out similar analyses.

```
0  99959 -0.1297    5 100359  1.1353
1  99758 -0.7653    6  99548 -1.4293
2 100026  0.0822    7  99800 -0.6325
3 100229  0.7242    8  99985 -0.0474
4 100230  0.7273    9 100106  0.3352
```

(b) Imagine that the creation of π really involved i.i.d. digits with probabilities $p_j = 0.10 + \delta_j$, for some conceivably very small but non-zero δ_j (summing to zero). Show that the Pearson test statistic K_n is then close to a $\chi^2_9(n\lambda)$, with $\lambda = \sum_{j=0}^9 \delta_j^2/p_{j,0} = 10\sum_{j=0}^9 \delta_j^2$. Suppose now that a person sneaks into your π laboratory in the middle of the night, switching for every 2000th decimal $0, 1, 2, 3, 4$ with $5, 6, 7, 8, 9$. Show that the Pearson test is actually able to detect this departure from plain $0.10, \ldots, 0.10$, at the 0.01 testing level.

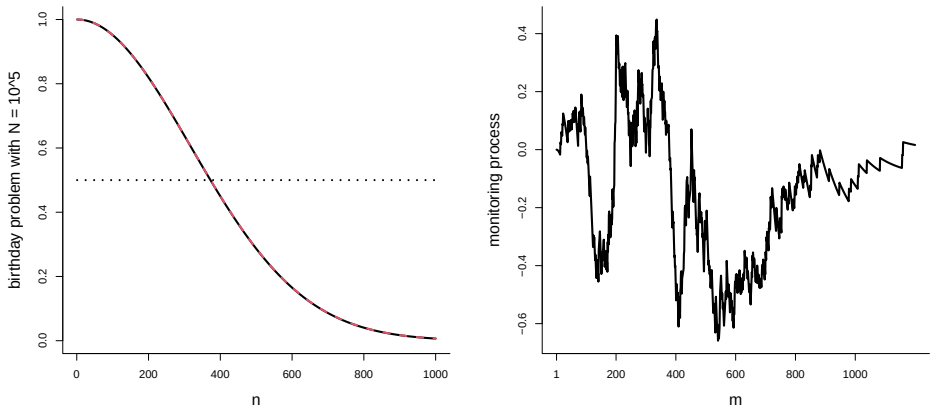


Figure vii.1: *Birthday problem, with $N = 10^5$ days in the year, left panel: probabilities p_n, along with approximation q_n, for n people to have all different birthdays, with median as low as 373. Decimals of π, right panel: computing the $k = 490$ birthday-match numbers $M_1, \ldots, M_k$ from the first 200,000 5-decimal numbers $u = d_1/10 + \cdots + d_5/10^5$, the model monitoring process $Z_k(m) = k^{1/2}\{G_k(m) - G(m)\}$.*

(c) Suppose you meet n people, with birthdays assumed to be independent and uniform on $1, \ldots, N$, with N the number of days in a year. Explain that the probability of these all having different birthdays is

$$p_n = (1 - 1/N)(1 - 2/N) \cdots (1 - (n-1)/N).$$

For n small to moderate, show that this can be well approximated with $q_n = \exp\{-\frac{1}{2}(n-1)n/N\}$. Explain that this leads to the approximation $n = \{(2\log 2)\, N\}^{1/2} = 1.177\sqrt{N}$ for the required number of people to be in the room such that the chance is 0.50 for having at least one common birthday. For $N = 365$, compute and plot these p_n, q_n, where you should find that already for $n = 23$ people, the chance of someone in the group having the same birthday is 0.50. Then take $N = 10^5$, and find the n such that $p_n \doteq 0.50$. Create a version of Figure vii.1, left panel.

a Plutonian number of days in a year

(d) Consider M, the first time we meet a person with birthday common to some of those already in the room. Show that

$$\Pr(M = m) = \Big\{ \prod_{j=1}^{m-2} (1 - j/N) \Big\} (m-1)/N \quad \text{for } m = 1, 2, \ldots.$$

Show via the p_n and q_n that the median is close to the solution to $n = 1.177\, N^{1/2}$. Carry out these calculations, for p_n, q_n and for the distribution of M, for the much bigger number of birthdays $N = 10^5$. For $N = 365$ and $N = 10^5$, compute the median and the mean of M, and comment. Perhaps counterintuitively, for n as relatively small as 373, there will with probability 0.50 be common birthdays in the group, even when the year has hundred thousand days. Then apply these birthday problem ideas to the first million digits of our friend π. Separate repeatedly after 5 decimals, creating the sequence $u_1, u_2, \ldots$ with 0.14159, 0.26535, etc., in the space with all $N = 10^5$ such 5-decimal

these numbers being small is sometimes referred to as 'the birthday paradox'

numbers $u = d_1/10 + \cdots + d_5/10^5$ on the unit interval. Organise the million digits into 5-digits numbers $u_1, \ldots, u_n$, with $n = 10^6/5 = 200 \cdot 10^3$. Construct a code from the million digits yielding $M_1, \ldots, M_k$, the first time one meets a number having been seen before, with appropriate restart each time. It turns out that $k = 490$ is the associated sample size, with the M_j ranging from 14 to 1160. Put up tests for the $M_1, \ldots, M_k$ to match the theoretical distributions. With $G(m) = \Pr(M \le m)$ the c.d.f., plot in particular $Z_k(m) = k^{1/2}\{G_k(m) - G(m)\}$, with the empirical c.d.f. G_k for the 490 M_i numbers, as in Figure vii.1, right panel. Explain via Ch. 9 theory that $Z_k(m)$ should be close to the Brownian bridge process $W^0(G(m))$, under the hypothesis that the decimals of π really are i.i.d. uniform on $0, 1, \ldots, 9$.

(e) If the grand hypothesis of i.i.d. uniform digits holds, show using Ex. 2.77 that the waiting times V_{10} required to have seen all ten decimals (where one starts counting again after having found one such complete cycle), must follow the distribution

$$g_{10}(v) = \Pr(V_{10} = v) = \sum_{j=1}^{9} (-1)^{j-1} \binom{9}{j-1} (1 - j/10)^{v-1} \quad \text{for } v \ge 10.$$

We can get hold of these decimals, starting with
3.1415926535 8979323846 2643383279 5**0**28841971 6939937510 5820974944 5923078164
and then counting away to get to the V_{10}; the first few are 32, 18, 19, 27, 15 ... We have actually managed to get hold of the first 10^9 decimals of π (!), and have written up a simple code to extract from these the first half million of the cycle lengths V_{10}. Compute the exact mean ξ and standard deviation σ for the V_{10} distribution, using the formula above or other results from Ex. 2.77; you should find $\xi = 29.2897$ and $\sigma = 11.2367$. Compute also skew and kurt, the skewness and the kurtosis. From the data file `V10counts`, with all these V_{10} variables $x_1, x_2, \ldots$, compute successive means $\bar{x}_n$ and standard deviations $\widehat{\sigma}_n$. Produce a version of Figure vii.2 (left panel for means, right panel for standard deviayions). The left panel has $\xi \pm 1.96\,\sigma/\sqrt{n}$ as a band (here displayed for $n = 1001$ to our upper limit). Explain how this illustrates both the Law of Large Numbers and the Central Limit Theorem. The right panel similarly has $\sigma \pm 1.96\,\kappa/\sqrt{n}$ lines, with $\kappa = (\frac{1}{2} + \frac{1}{4}\text{kurt})\sigma$; check with Ex. 2.51, and compute the kurt number in question. Supplement this with a plot of $\sqrt{n}(\widehat{\sigma}/\sigma - 1)/(\frac{1}{2} + \frac{1}{4}\gamma_4)^{1/2}$, also of the running $t_n = \sqrt{n}(\bar{x}_n - \xi)/\sigma$ as a function of n, and comment.

(f) The π decimals give us the rare chance of testing whether a given hypothesised (and perhaps esoteric) distribution is fully correct, with an enormous sample size. Use the $m = 5 \cdot 10^5$ cycle lengths V_{10} to count $N(v)$, the number of times $V_{10} = v$ is observed, and compare $N(v)/m$ with the hypothesised $g(v)$, for $v \ge 10$. Produce a version of Figure vii.3 (right panel), which has the Pearson residuals $m^{1/2}\{N(v)/m - g(v)\}/g(v)^{1/2}$. Argue that these Pearson residuals should be approximately independent and standard normal. Carry out a Pearson chi-squared test, and comment. Inspecting these numbers and tables (xx check with com69d xx), we have $N(93) = 51$ sightings of $V_{10} = 93$, in this first half a million times, creating a Pearson residual of 3.628. Is this too big?

(g) (xx edit and polish this. xx) For $v \ge 50$, say, show that the $N(v)$ counts must be close to independent Poissons, with parameters $\lambda_v = mg(v)$. Work with $Z = \sum_v \{N(v) - mg(v)\}^2/\{mg(v)\}$, summing over an index set in this terrain of $v \ge 50$. Show that $\mathrm{E}\, Z = 2k + \sum_v 1/\{mg(v)\}$. Use this to test whether the $N(v)$ counts behave suspiciously.

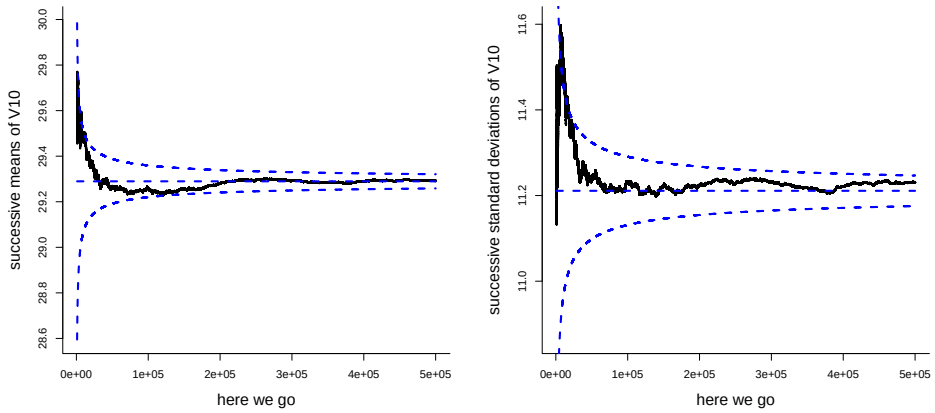


Figure vii.2: *Decimals of π: computing the first half million successive means (left panel) and standard deviation (right panel), based on some 15 million decimals, for the V_{10} cycle lengths. They are in convergence towards $\xi = 29.28979$ and $\sigma = 11.2367$, indicated with horizontal lines. The bands represent $\pm 1.96\,\tau/\sqrt{n}$ lines, with $\tau = \sigma$ for the means and $\tau = (\frac{1}{2} + \frac{1}{4}\mathrm{kurt})^{1/2}\sigma$ for the standard deviations.*

Story vii.3 *Random integers via prime numbers.* The first few prime numbers are of course $2, 3, 5, 7, 11, \ldots$; denote these $p_1, p_2, p_3, \ldots$. Every integer N can be uniquely represented as $N = \prod_{j=1}^{\infty} p_j^{X_j}$, with finitely many non-zero X_j. One may hence create models for random integers by placing probability distributions on the X_j.

(a) Let $X_1, X_2, \ldots$ be independent random variables, with values in $\{0, 1, 2, \ldots\}$. Show using the Borel–Cantelli Lemma that the infinite product N is a well-defined random integer variable if and only if $\sum_{j=1}^{\infty} \Pr(X_j \geq 1) = \sum_{j=1}^{\infty} \{1 - \Pr(X_j = 0)\}$ is finite. Show that the division here is sharp: if the sum diverges, then not only is $N = \infty$ with positive probability, but with probability one.

(b) Let first $X_j \sim \mathrm{Pois}(d_j)$; show that N is finite if and only if $\sum_{j=1}^{\infty} d_j$ is finite. Then let the X_j be independent Bernoulli variables with $\Pr(X_j = 1) = r_j$, and with $r_j = 1/j^{1.5}$, say. Show that this leads to a well-defined probability distribution on the set of modest integers, those where all exponents $x_j \in \{0, 1\}$. Simulate say 1000 random modest integers from this distribution. Find the mean of both N and $\log N$.

(c) Now suppose we give X_j independent geometric distributions, of the form $\Pr(X_j = x) = (1 - c_j)c_j^x$ for $x = 0, 1, \ldots$. Explain that (xx check with care xx) $\mathrm{E}\, X_j = c_j/(1 - c_j)$, and that N is well-defined if and only if $\sum_{j=1}^{\infty} c_j$ is finite.

(d) Study in particular the case of $c_j = 1/p_j^{\alpha}$. Explain that this leads to a well-defined random integer N, provided $\alpha > 1$, with $\mathrm{E}\, X_j = 1/(p_j^{\alpha} - 1)$. Show indeed that

$$\Pr(N = n) = \frac{1}{\zeta_0(\alpha)} \frac{1}{n^{\alpha}} \quad \text{for } n = 1, 2, 3, \ldots, \text{ where } \zeta_0(\alpha) = \prod_{\text{primes}} \frac{p^{\alpha}}{p^{\alpha} - 1}.$$

Explain that this $\zeta_0(\alpha)$ must be identical to the famous Riemann zeta function, $\zeta(\alpha) = \sum_{n \geq 1} 1/n^{\alpha}$. We have hence reached a simple probabilistic proof of the zeta function

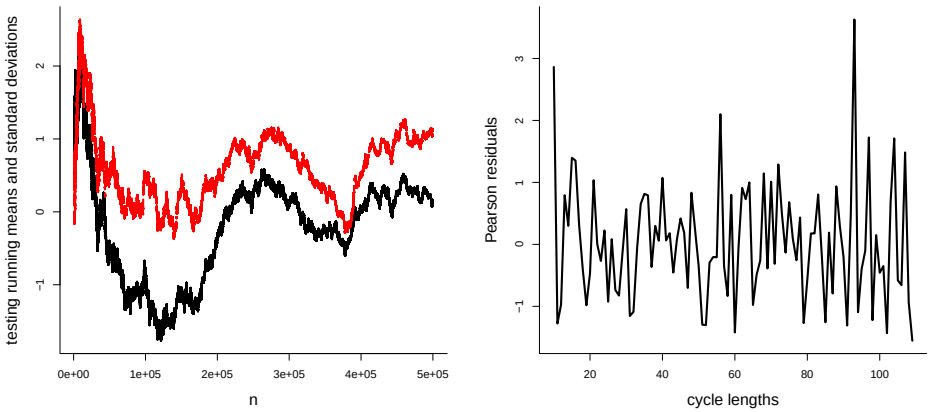


Figure vii.3: *Left panel: plots of running consecutive tests, for* $\xi = 29.2897$ *(black) and for* $\sigma = 11.2110$ *(red), with values of* n *from 1001 to the max number half a million. Right panel: Pearson residuals* $m^{1/2}\{N(v)/m - g(v)\}/g(v)^{1/2}$*, comparing observed frequencies* $N(v)/m$ *with the hypothesised* $g(v)$ *distribution, for the half a million cycle lengths.*

representation in terms of products over all primes (proven first by Euler in 1737, though via very different means). In particular,

$$\frac{\pi^2}{6} = \sum_{j=1}^{\infty} \frac{p_j^2}{p_j^2 - 1} = \frac{4}{3}\frac{9}{8}\frac{25}{24}\frac{49}{48}\cdots, \qquad \frac{\pi^4}{90} = \sum_{j=1}^{\infty} \frac{p_j^4}{p_j^4 - 1} = \frac{16}{15}\frac{81}{80}\frac{625}{624}\frac{2401}{2400}\cdots.$$

You know that π is irrational. Prove, as Euclid did about three hundred years b.C., that there is an infinitude of primes. Prove also from this that $\zeta(\alpha) \to \infty$ as $\alpha \to 1$.

(e) Say that N is square-free if all $X_j \leq 1$. Show that the probability of this happening is $1/\zeta(2\alpha)$.

(f) Let N be such a random natural number, drawn from what we may term the zeta distribution, with probabilities proportional to $1/n^\alpha$, for some $\alpha > 1$. Prove the following. (i) N is odd with probability $1 - (\frac{1}{2})^\alpha$, e.g. $3/4$ for $\alpha = 2$. (ii) N is a square with probability $\zeta(2\alpha)/\zeta(\alpha)$, e.g. $\pi^2/15$ for $\alpha = 2$. (iii) The number 100 will be a factor in N with probability $(1/100)^\alpha$. Generalise.

(g) Consider the Möbius function from number theory, defined by setting $\mu(1) = 1$ and $\mu(p_{j_1} \cdots p_{j,r}) = (-1)^r$ if the number is a product over distinct primes; for all other n, $\mu(n) = 0$. Show that $\mu(n)$ is nonzero precisely for what we termed modest integers above. Now prove the intriguing number theory formula

$$\sum_{n=1}^{\infty} \frac{1}{n^\alpha} \sum_{n=1}^{\infty} \frac{\mu(n)}{n^\alpha} \equiv 1,$$

by working with the mean of $\mu(N)$ in different ways.

(h) Let $N_1, \ldots, N_m$ be independently drawn from the zeta distribution with $\alpha = 2$. Find the probability limit of $(N_1 \cdots N_m)^{1/m}$.

(i) We've drawn $m = 25$ numbers from the zeta distribution, for a particular value of α:
`1 1 1 1 1 1 1 1 1 2 2 2 2 5 5 10 11 12 12 36 47 63 234 464`
Estimate the α we have used, with a 95 percent confidence interval.

Story vii.4 *Time-to-failure for machine components.* (xx polish. check with other place we use nlm. xx) Among the aims of the present story is to showcase how the general maximum likelihood theory of Ch. 5 can be applied also in new situations, perhaps with models outside the usual repertoire. Even with a freshly invented model one may fit parameters, read off approximate standard deviations, construct confidence intervals, test hypotheses, as long as the log-likelihood can be programmed. The machinery also applies to any interest functions of the model parameters, via the delta method. It is to be noted that general-purpose numerical optimisation methods and algorithms, along with routines for computing gradients and Hessians, i.e. first and second order derivatives, are wondrously helpful here, essentially with only modest extra efforts needed beyond having programmed the log-likelihood.

Of course methods also apply for standard models, where there might be packages or established routines accomplishing the fitting and testing, but the spirit for the modern statistician should be that of building and trying out also new models for new purposes. This might be even more important for regression type models, where similar programming and implementation schemes work; see e.g. Story iv.7.

We build our present illustrations around the following dataset, with $n = 201$ time-to-failure measurements for certain machine parts, taken from the SAS User's Guide Ch. 44. We shall fit the data to the gamma and Weibull models, and also to a three-parameter extension of these, which we call the gamma-Weibull model.

```
  [1]  620  470  260   89  388  242  103  100   39  460  284 1285  218  393  106
 [16]  158  152  477  403  103   69  158  818  947  399 1274   32   12  134  660
 [31]  548  381  203  871  193  531  317   85 1410  250   41 1101   32  421   32
 [46]  343  376 1512 1792   47   95   76  515   72 1585  253    6  860   89 1055
 [61]  537  101  385  176   11  565  164   16 1267  352  160  195 1279  356  751
 [76]  500  803  560  151   24  689 1119 1733 2194  763  555   14   45  776    1
 [91] 1747  945   12 1453   14  150   20   41   35   69  195   89 1090 1868  294
[106]   96  618   44  142  892 1307  310  230   30  403  860   23  406 1054 1935
[121]  561  348  130   13  230  250  317  304   79 1793  536   12    9  256  201
[136]  733  510  660  122   27  273 1231  182  289  667  761 1096   43   44   87
[151]  405  998 1409   61  278  407  113   25  940   28  848   41  646  575  219
[166]  303  304   38  195 1061  174  377  388   10  246  323  198  234   39  308
[181]   55  729  813 1216 1618  539    6 1566  459  946  764  794   35  181  147
[196]  116  141   19  380  609  546
```

(a) Consider the three-parameter density function

$$f(y, a, b, c) = k(a, b, c) y^{a-1} \exp(-by^c) \quad \text{for } y > 0.$$

Prove that the normalisation constant must be $k(a, b, c) = cb^{a/c}/\Gamma(a/c)$. Show that $c = 1$ gives the $\text{Gam}(a, b)$ distribution, and that the special case $a = c$ corresponds to c.d.f. $F(y) = 1 - \exp(-by^c)$, which is a Weibull (though with a different parametrisation than with Ex. 1.58). Due to these special cases we may call this three-parameter family the gamma-Weibull distribution. Prove also the mean formula

$$\text{E}\, Y = \frac{\Gamma(a/c + 1/c)}{\Gamma(a/c)} \frac{1}{b^{1/c}},$$

and verify that mean formulae for the gamma and the Weibull indeed are special cases. (xx nils notes: a/b for $c = 1$ and $\Gamma(1 + 1/c)/b^{1/c}$ for $a = c$, the weibull case. xx)

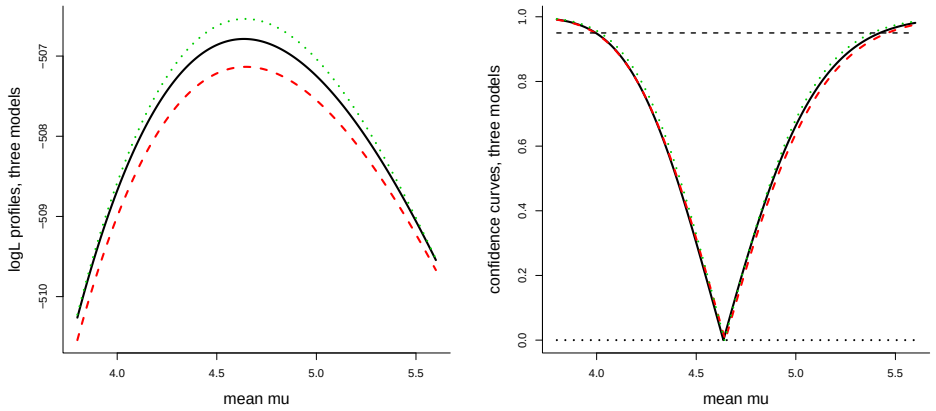


Figure vii.4: *For the time-to-failure data, left panel: log-likelihood profile functions for the mean μ, via the* Gam(a, b) *model, the Weibull* (b, c) *model, and the three-parametric* (a, b, c) *model. Right panel: these are transformed to confidence curves via the Wilks theorem, see Ex. 7.11, where e.g. 95 percent intervals can be read off. These are in essential agreement here.*

(b) First read the data into your computer suitably. It helps accurate numerics to scale them with a factor of e.g. 1/100; in the scripts below this is what we call `yy`. Show that the following little script succeeds in computing the log-likelihood for the Gam(a, b) model:

```
logL1 = function(para)
{
a = para[1]
b = para[2]
aux = (a-1)*log(yy)-b*yy + a*log(b) - lgamma(a)
sum(aux)
}
```

To maximise the log-likelihood, along with the Fisher information matrix $\widehat{J} = -\partial^2 \ell_n(\widehat{\theta})/\partial\theta\,\partial\theta^{\mathrm{t}}$, it is practical to use the general-purpose non-linear minimisation algorithm `nlm` in R. Define the function `minuslogL1` as `-logL1`, and then use

```
starthere1 = c(1,1)
fit1 = nlm(minuslogL1,starthere1,hessian=T)
ML1 = fit1$estimate
Jhat1 = fit1$hessian
se1 = sqrt(diag(solve(Jhat1)))
show1 = cbind(ML1,se1)
```

Carry out this scheme, and explain what the different steps involve and achieve; sometimes a bit of fiddling might be required with the start point `starthere1` to secure convergence of the numerical iterative minimisation procedure. Read off both ML estimates

$(\widehat{a},\widehat{b})$ and approximate 95 percent confidence intervals for them. Test the hypothesis that the data are actually from the simpler exponential model.

(c) For any focus parameter $\mu = \mu(\theta)$ of the model parameters, the delta method says that the approximate variance of $\widehat{\mu}_{\rm ml} = \mu(\widehat{\theta}_{\rm ml})$ is $\widehat{\kappa}^2 = \widehat{d}^{\rm t}\widehat{J}^{-1}\widehat{d}$, with $\widehat{d} = \partial\mu(\widehat{\theta})/\partial\theta$. To illustrate the general machinery, consider the mean $\mu = {\rm E}\,Y$, which for the gamma model is the simple a/b. Here partial derivatives etc. are easily found, but to explain the general practical principle start by defining the function `mu1 = function(para)` as `a/b`, and then carry out

```
mu1hat = mu1(ML1)
der1 = grad(mu1,ML1)
kappa1 = sqrt( t(der1) %*% solve(Jhat1) %*% der1 )
```

where `grad` is available via the library `numDeriv`. Construct a 95 percent interval for the mean using this. Modify your code to similarly find estimate and interval for the median.

(d) Having accomplished the above for the gamma model, modify your code to handle also the Weibull model, with $F(y) = 1-\exp(-by^c)$. Find ML estimates, their estimated standard deviations, test exponentiality; then find estimates and intervals for the mean and the median. Part of the intended experience here is that passing from one model to another often does not take many extra efforts, as results flow from having programmed the log-likelihood.

(e) There are packages and routines available handling the gamma and Weibull models, but perhaps not our three-parameter extension. Programme the appropriate $\ell_n(a,b,c)$, then find ML estimates $(\widehat{a},\widehat{b},\widehat{c})$, along with estimates and confidence intervals for the mean and median. For these tasks it is indeed helpful to have an explicit formula for the normalisation constant $k(a,b,c)$, but it is useful for other models and situations to learn that one may manage without, via numerical integration routines, typically in the format of `integrate(g,0,Inf)$value`. For the learning experience, redo the fitting of the (a,b,c) model without using the $k(a,b,c)$ formula.

(f) The methods and programmes above lead to confidence intervals of the first-order large-sample approximation type, say $\widehat{\mu}\pm 1.96\,\widehat{\kappa}$. Supplement your efforts by programming also the log-profile-likelihood functions, $\ell_{n,{\rm prof}}(\mu) = \max\{\ell_n(\theta)\colon \mu(\theta)=\mu\}$, for the three models. Here you are helped by having explicit formulae for the μ. Construct a version of Figure vii.4, left panel. The profiles are in good agreement hee, and the three-parameter model does not lead to any significant increase over the two two-parameter models. Then construct a version of the *confidence curves* in the right panel, as follows. With $D(\mu) = 2\{\ell_{n,\max} - \ell_{n,{\rm prof}}(\mu)\}$ the deviance, explain that $D(\mu) \approx_d \chi^2_1$, at the true position in the parameter space, via the Wilks theorem. Deduce as with Ex. 7.11 that ${\rm cc}(\mu) = \Gamma_1(D(\mu))$ has the uniform distribution, and that the random set $\{\mu\colon {\rm cc}(\mu)\le 0.95\}$ has probability 0.95 of covering the true value.

(g) On this particular occasion the different models produce rather similar confidencec intervals for the mean μ, also when it comes to comparing direct $\widehat{\mu}\pm 1.96\,\widehat{\kappa}/\sqrt{n}$, say, with the Wilks theorem based ones, with $\{\mu\colon \Gamma_1(D(\mu))\le 0.95\}$. Go through the required calculations, yielding the following little table, with direct symmetric intervals to the left and Wilks based on the right.

```
                    low   up      low   up
gamma(a,b)          3.931 5.342   3.996 5.421
weibull (b,c)       3.923 5.368   3.992 5.459
three-para (a,b,c) 3.964 5.321   4.019 5.382
```

(h) Use the occasion to check one or two more models. For each, programme the log-likelihood function, find ML estimates and their estimated standard deviations. Compute also the log-likelihood maxima and the associated AIC scores, as per (11.1). One particular model is what we may term the Beta envelope around the exponential, with c.d.f. $F(y,\theta,a,b) = \mathrm{Be}(1-\exp(-\theta y),a,b)$. The exponential model then corresponds to $(a,b)=(1,1)$. We learn that the two-parameter Gamma model is the best.

```
 logLmax   dim   aic
-509.3245    1  -1020.649  expo
-506.7866    2  -1017.573  gamma
-507.1344    2  -1018.269  weibull
-506.6254    3  -1019.251  gamma-weibull
-506.5368    3  -1019.074  beta envelope
```

Story vii.5 *Speed of light in 1882, and BHHJ estimation.* Impressively, Simon Newcomb managed to estimate the speed of light after audacious experiments in 1882. These amounted in part to measuring the time t_i, in 10^{-6} seconds, it took light to travel 7442.42 m. He then transformed these to observations y_i, via the equation $y_i = 10^3(t_i - 24.8)$. His 66 y_i data were as follows, with two conspicuous outliers:

```
28 22 36 26 28 28 26 24 32 30  27 24 33 21 36 32 31 25 24 25
28 36 27 32 34 30 25 26 26 25 -44 23 21 30 33 29 27 29 28 22
26 27 16 31 29 36 32 28 40 19  37 23 32 29 -2 24 25 27 24 16
29 20 28 27 39 23
```

Translating the 1882 experiments and intentions to data analysis, we formulate the essential task being to estimate the correct mean of the underlying distribution, along with its spread. We take this to mean fitting a $\mathrm{N}(\xi,\sigma^2)$ to the data. Using the direct mean $\bar y$ and standard deviation s will produce a very biased view, since the two outliers exert very notable influence. Below we repair these estimates using the BHHJ method of Ex. 5.9. Incidentally, the true speed of light value 299856.2 km/s translates to $\xi_0 = 33.2$ on Newcomb's y scale.

(a) Compute and display a nonparametric kernel-type estimate $\widehat f(y)$ of the underlying density f, along with direct $\mathrm{N}(\widehat\xi_{\rm ml},\widehat\sigma^2_{\rm ml})$, using ML estimates, as in Figure vii.5, left panel. We see that the ML estimates are far off, as is the estimated normal density.

(b) We therefore turn to the likelihood robustification BHHJ method, worked with in Ex. 5.9, 5.19. Show that this for the normal model amounts to minimising

$$H_n(\xi,\sigma) = (2\pi)^{-1/2}(1+a)^{-1/2}\sigma^{-a} - (1+1/a)\frac{1}{n}\sum_{i=1}^n\Big\{\phi\Big(\frac{y_i-\xi}{\sigma}\Big)\frac{1}{\sigma}\Big\}^a.$$

The a is a positive tuning parameter, with a small corresponding to coming close to the ML method, and bigger a yielding more robustness. Carry out BHHJ estimation with the Newcomb data, for a grid of a values over say $[0,0.50]$, with $a=0$ being the ML method. Construct versions of the plots of Figure vii.6, with $\widehat\xi_a$ and $\widehat\sigma_a$ as functions of a. These start at the ML values $26.212, 10.745$, and then become gradually less influenced by the outliers as a increases. The horizonal lines indicate ML values for the cleaned dataset, with the two outliers removed.

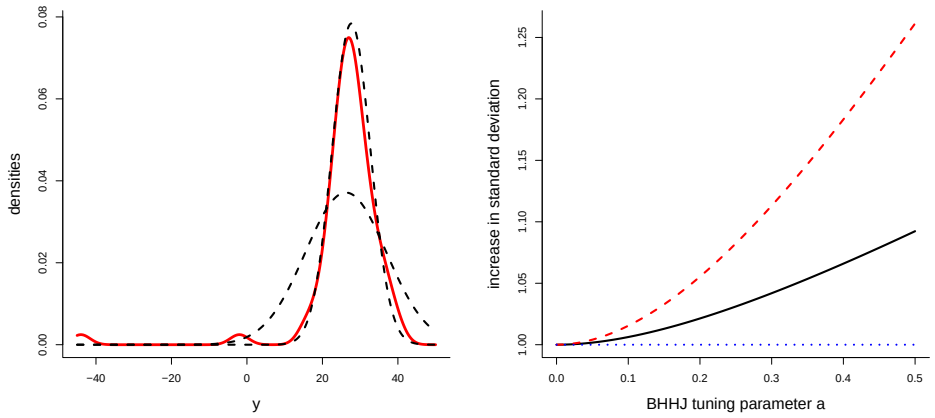


Figure vii.5: *Left panel: nonparametric estimate* $\widehat{f}$ *of the underlying density, with two bumps for the two outliers, along with two fitted normal densities. The lower one, which is considerably off, is from ML estimation, and the upper one, providing a very good fit, is from using BHHJ with* $a = 0.20$. *Right panel: relative efficiency of the BHHJ method, in terms of standard deviation divided by optimal standard deviation, under normal model conditions, as a function of the BHHJ tuning parameter* $a \in [0, 0.50]$*; for* ξ *(smooth curve) and for* σ *(dashed curve).*

(c) Using a positive a rather than $a = 0$ means gaining robustness, as seen here, but sacrificing a certain amount of efficiency under model circumstances. The limiting variance matrix is of the form $J_a^{-1} K_a J_a^{-1}$, see Ex. 5.19, which can be compared to J_0^{-1} for the optimal-under-model method. Use this occasion to investigate $r_1(\xi, a)$, the limit standard deviation for $\widehat{\xi}_a$ divided by that for the ML for ξ, and $r_2(\sigma, a)$, the limit standard deviation for $\widehat{\sigma}_a$ divided by that for the ML for σ. Construct a version of Figure vii.6, right panel. The efficiency loss is small, for a small; for $a = 0.20$, for example, show that one loses 2.2 percent for ξ estimation and 5.5 percent for σ estimation, in terms of widths of confidence intervals. We have used this $a = 0.20$ value to display the associated estimated normal density in Figure vii.5, left panel; the density comes very close to the nonparametric one, for the relevant range of data.

(d) Newcomb did a good job, in 1882: show that the true value $\xi_0 = 33.2$ is well inside relevant confidence intervals for ξ, based on the BHHJ analysis of his 66 datapoints.

(e) It might not be necessary for the data analysis here, but carry out similar analysis, with both ML and BHHJ for a range of a, with the three-parameter t model, i.e. using $y_i = \xi + \sigma\varepsilon_i$ with $\varepsilon_i \sim t_\nu$.

Story vii.6 *How many of those born now will become at least 90?* (xx nilsdemography rant, so far; will be matched with emil life expectancy Story i.17; human mortality databases; need serious sttructuring and polish. xx) will calibrate notation and methods with that emil story. $F(t) = 1 - \prod_{i \le t}(1 - \alpha_i)$. Figure vii.7 displays the quantiles, at level 0.1, 0.3, 0.5, 0.7, 0.9, for Norwegian women and men. then do modelling and prediction. xx)

(a) Consider a time to event random variable T, on the continuous scale, with cumulative

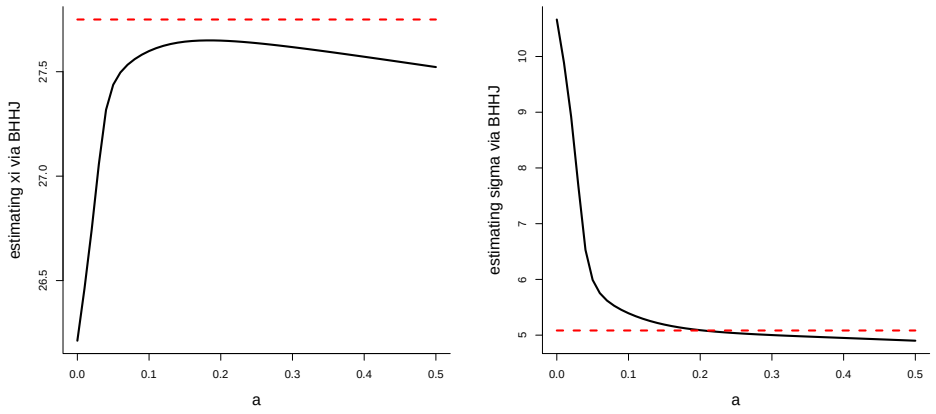


Figure vii.6: *BHHJ method estimates* $\widehat{\xi}_a$ *(left panel) and* $\widehat{\sigma}_a$ *(right panel), for the* $\mathrm{N}(\xi, \sigma^2)$ *model, based on Newcomb's 66 measurements, as a function of tuning parameter* a*. ML estimates* $(26.212, 10.745)$ *are at* $a = 0$*. The horizontal lines at* $\xi = 27.75$ *and* $\sigma = 5.083$ *indicate the 'best' estimates, computed for the cleaned dataset with the outliers removed. The BHHJ estimates chosen for Figure vii.5, left panel, are for* $a = 0.20$*.*

hazard rate $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ and survival function $S(t) = \exp\{-A(t)\}$. If data are observed on a discrete grid of intervals $[j, j+1)$, for $j = 0, 1, \ldots$, show that the chance for an individual still at risk at time j of surviving also $j+1$ is $S(j+1)/S(j)$. Deduce that the time-discrete hazards are

$$\begin{aligned} \alpha_j &= \Pr(T \in [j, j+1) \,|\, T \geq j) \\ &= 1 - S(j+1)/S(j) = 1 - \exp[-\{A(j+1) - A(j)\}]. \end{aligned} \tag{vii.1}$$

In particular, any parametric model for the continuous survival function, or for the cumulative hazard, translates to a parametric model for the α_j. If increments are small, or the time grid fine, we have $\alpha_j \doteq A(j+1) - A(j) \doteq \alpha(j)$, but when data are observed on the discrete grid we need in general to work with the α_j in their $[0, 1]$ space.

(b) Suppose survival data are available of the form $N_j \sim \mathrm{binom}(Y_j, \alpha_j)$, with Y_j at risk at the start of $[j, j+1)$, of whom N_j die inside that time window. We view the data as a chain of binomials. (xx will give the right asymptotics, with martingale argument. xx) The direct binomial estimator, without further modelling assumptions, is $\widehat{\alpha}_j = N_j/Y_j$. With a parametric model, say $\alpha_j(\theta)$, show that the log-likelihood function can be written

$$\begin{aligned} \ell_n &= \sum_{j=0}^{m} \big[N_j \log \alpha_j(\theta) + (Y_j - N_j) \log\{1 - \alpha_j(\theta)\}\big] \\ &= \sum_{j=0}^{m} Y_j \big[\widehat{\alpha}_j \log \alpha_j(\theta) + (1 - \widehat{\alpha}_j) \log\{1 - \alpha_j(\theta)\}\big]. \end{aligned}$$

Write down the basic first-order properties of the maximum likelihood estimator $\widehat{\theta}$. (xx can we have a very simple illustration? can ask for α_j constant. xx)

(c) When working with parametric hazards models it might be important to have better fit for e.g. the higher values than the lower. Suppose this is translated to weights w_j

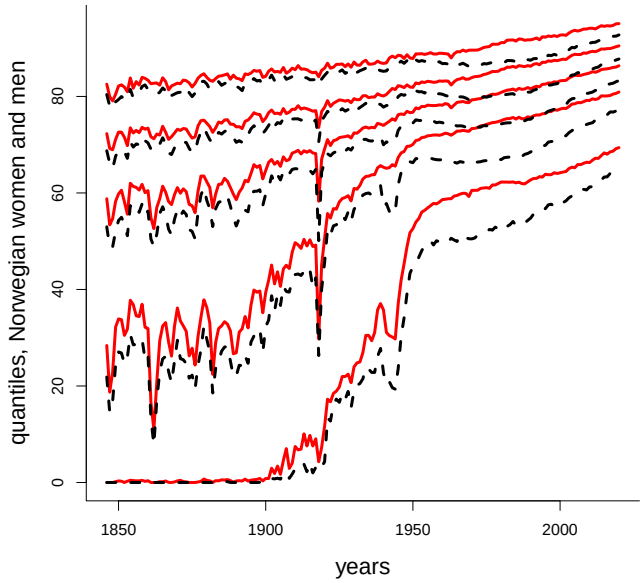


Figure vii.7: *A demographic view of Norwegian history, 1846–2020: 0.1, 0.3, 0.5, 0.7, 0.9 quantiles of lifelength distribution, for women (full curves, above) and men (broken curves, below).*

reflecting such relative importance. Define the weighted log-likelihood function as

$$\ell_n = \sum_{j=0}^{m} w_j Y_j \big[\widehat{\alpha}_j \log \alpha_j(\theta) + (1-\widehat{\alpha}_j) \log\{1-\alpha_j(\theta)\}\big].$$

(xx a bit more. but wish to do norway with gompertz and skewed gompertz before coming to large-sample theory. xx)

(d) Using $\widehat{F}_t = 1 - \prod_{j\le t}(1-\widehat{\alpha}_j)$, with $\widehat{\alpha}_j = \mathrm{d}N(j)/Y(j)$, implement, and read off the quantiles, and construct a version of Figure vii.7. The 0.90 quantile, for example, for a given year 1960, is found by checking where the $\widehat{F}_{1960}(t)$ line crosses 0.90, where we use linear approximation to have a smoother plot than if we merely compute the cumulatives as constant inside each year.

(e) For Norway data we shall use the time-continuous Gompertz model, which is $A(t) = (a/b)\{\exp(bt)-1\}$. Show that this via (vii.1) implies $\alpha_j = 1-\exp\{-d\exp(bj)\}$, with $d = (a/b)\{\exp(b)-1\}$. Nils will also try out a three-parameter skewed generalisation of the Gompertz, using

$$A(t) = (a/b)\{\exp(bt)-1\}^c, \quad \text{again with} \quad \alpha_j = 1-\exp[-\{A(j+1)-A(j)\}]$$

The plan is to fit a couple of these models to Norwegian demographic data, men and women, 1846 to 2020, with weights reflecting interest in the higher ages. We then find $(\widehat{a}_t, \widehat{b}_t)$, year for year, and may then meta-model the underlying (a_t, b_t) over time.

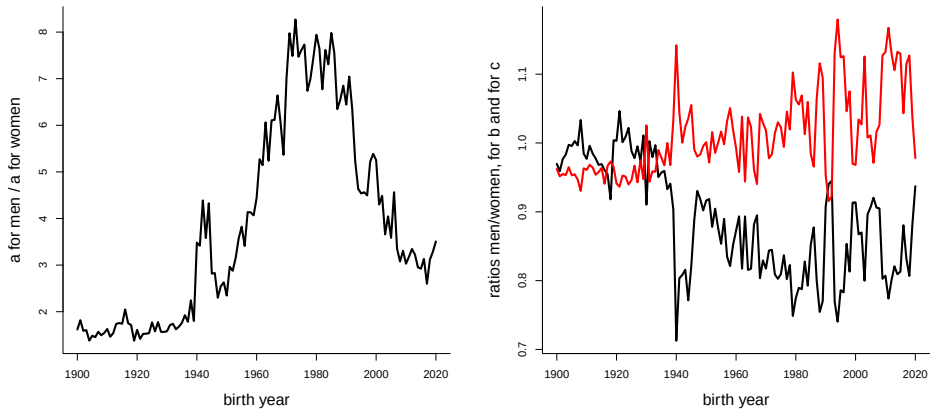


Figure vii.8: *For the skewed Gompertz model, with parameters* (a, b, c)*, the men to women parameter ratios are computed as a function of birth year. Left panel:* $\widehat{a}_m/\widehat{a}_w$*, with a clear peak around 1980; right panel:* $\widehat{b}_m/\widehat{b}_w$*, becoming stable at around 0.85, and* $\widehat{c}_m/\widehat{c}_w$*, becoming stable at around 1.00.*

(f) (xx then nils does gompertz, birth year by birth year, finding smooth changes in the parameters. this is then used to track the 0.90 quantile and also to estimate $F(90)$ over time. xx) can estimate the Gompertz parameters in $A(t) = \beta\{\exp(\gamma t) - 1\}$, $S(t) = \exp\{-A(t)\}$. (i) can work with this, for fixed t_0, e.g. 90 years; how may born in 2025 will be alive in 2135? (ii) time point t^* where $F(t^*) = 0.90$, or $S(t^*) = 0.10$.

(g) (xx to come here, with suitable things to monitor, based on fitting the three-parameter skewed Gompertz model (a, b, c) to the men and women year groups. xx) So $A(t) = a[(1/b)\{\exp(bt) - 1\}]^c$. Show that $F(t) = q$ means $A(t) = z = -\log(1 - q)$, which means quantile $t^* = (1/b)\log(1 + b(z/a)^{1/c})$. Figure vii.8 shows parameter ratios $\widehat{a}_m/\widehat{a}_w$, with a clear peak around 1980, $\widehat{b}_m/\widehat{b}_w$, becoming stable at around 0.85 $\widehat{c}_m/\widehat{c}_w$, becoming stable at around 1.00. this has consequences for longer lives and prediction. further: (i) plot $\log \widehat{a}_m$ and $\log \widehat{a}_w$, approximately linear with downward trend, and predict to 2030, 2040, 2050. (ii) plot $\widehat{b}_m$ and $\widehat{b}_w$, to see that these have become about stable. (iii) plot $\widehat{c}_m$ and $\widehat{c}_w$, approximately linear with a small upward trend. then use all of this to fremskrive (a, b, c) to 2030, 2040, 2050, for men and women. via a little set of monitoring functions $M(a, b, c)$, prediction of these will imply evolution of (a, b, c).

(h) We need to understand the behaviour of the maximum weighted likelihood estimator $\widehat{\theta}$. Show first that with $\alpha_j^*(\theta) = \partial\alpha_j(\theta)/\partial\theta$, we have

$$U_n = \frac{\partial \ell_n}{\partial \theta} = \sum_{j=0}^{m} w_j \frac{N_j - Y_j\alpha_j(\theta)}{\alpha_j(\theta)\{1 - \alpha_j(\theta)}\alpha_j^*(\theta),$$

and that we at the true parameter value have $U_n \approx_d \mathrm{N}_p(0, K_n)$, with

$$K_n = \sum_{j=0}^{m} w_j^2 \frac{Y_j}{\alpha_j(1 - \alpha_j)}\alpha_j^*(\alpha_j^*)^{\mathrm{t}}.$$

Show further via Taylor expansion arguments (xx pointer to Ch4-Ch5 things xx) that $\widehat{\theta} \approx_d \mathrm{N}_p(\theta_0, \Sigma_n)$, with the sandwich matrix $\Sigma_n = J_n^{-1} K_n J_n^{-1}$, where

$$J_n = \sum_{j=0}^m w_j \Big\{ \frac{N_j}{\alpha_j^2} + \frac{Y_j - N_j}{(1-\alpha_j)^2} \Big\} \alpha_j^*(\alpha_j)^{\mathrm{t}} \doteq \sum_{j=0}^m w_j \frac{Y_j}{\alpha_j(1-\alpha_j)} \alpha_j^*(\alpha_j^*)^{\mathrm{t}}.$$

Here we may also use $\widehat{J}_n = -\partial^2 \ell_n(\widehat{\theta})/\partial\theta\,\partial\theta^{\mathrm{t}}$, the Hessian at position $\widehat{\theta}$.

(i)

Story vii.7 *Stout's physician and the last n.* In his book *Almost Sure Convergence*, Stout (1974, p. 9) draws up a little scenario, which we here re-tell with a bit of notation and soon enough additional comments. It involves a physician who treats patients with a drug having the same unknown cure rate p for each patient, and who will be using the same drug as long as no superior alternative is found. From time to time he estimates p, using the binomial proportion $\widehat{p}_n$ after n patients. Now suppose the physician wishes to estimate p within a tolerance $\varepsilon > 0$. He asks whether he will ever reach a point in time such that with high probability, *all subsequent estimates* will fall in the $[p-\varepsilon, p+\varepsilon]$ interval. Is there a finite N such that

$$\Pr(\max_{n \geq N} |\widehat{p}_n - p| \leq \varepsilon) \geq 0.95, \tag{vii.2}$$

say? The point conveyed, later echoed by Serfling (1980, p. 49), who essentially repeats this story, is that this question is *not* answered by the convergence in probability statement $\Pr(|\widehat{p}_n - p| \leq \varepsilon) \to 1$. But the strong law of large numbers [xx pointer to Ch. 2 xx] comes to the partly informative rescue, essentially saying that since $\Pr(\widehat{p}_n \to p) = 1$, there is such a finite N, with probability one. Such theorems, and the basic literature concerning strong convergence, say nothing about its expected size, or indeed distribution (nobody knows the future, so nobody can know its precise size), however. In this story we reach precise limit distributions results for

$$N_\varepsilon = \max\{n \geq 1 \colon |\widehat{p}_n - p| \geq \varepsilon\},$$

'the last n in the SLLN', also for rather more general cases than for the binomial setup.

(a) To understand the sample sizes we need to encounter, start with the classic approximate 0.95 interval $\widehat{p}_n \pm 1.96\,\widehat{\sigma}_n/\sqrt{n}$, consider n_ε, the size needed for such an interval to have size 2ε or smaller. Show that this is the same as requiring $\Pr_p(|\widehat{p}_n - p| \leq \varepsilon) \geq 0.95$, and that indeed $\sqrt{n_\varepsilon} \doteq 1.96\,\widehat{\sigma}_n/\varepsilon$, or $n_\varepsilon \doteq 1.96^2\,\widehat{\sigma}_n^2/\varepsilon^2$.

(b) For the case of Stout's physician, with the variance formula $\widehat{\sigma}_n^2 = \widehat{p}_n(1-\widehat{p}_n)$, how large must n be, in order for the 0.95 interval to be of length less than $2 \cdot 0.05$, and less than $2 \cdot 0.01$?

(c) With $y > 0$, let $m = \langle y/\varepsilon^2 \rangle$ be the smallest integer greater than of equal to y/ε^2. Then $y_0 = m\varepsilon^2$ is close to y; show that $y_0 - \varepsilon^2 < y \leq y_0$. Then show that

$$\Pr(\varepsilon^2 N_\varepsilon \geq y) = \Pr(N_\varepsilon \geq m) = \Pr(\sqrt{m} \max_{n \geq m} |\widehat{p}_n - p| \geq y_0^{1/2}).$$

(d) This leads to investigating the closeness of a sequence of sample means to its target, which we can do via a different perspective on the Donsker theorem of Ch. 9. Let

$U_1, U_2, \ldots$ be i.i.d. with mean zero and standard deviation σ, with sample averages $\bar{U}_m, \bar{U}_{m+1}, \ldots$, for suitably high m. Consider the process $Z_m(t) = m^{1/2}\bar{U}_{[mt]}$ for $t \geq 1$, to be worked with in the space $D[1,c]$ of all right-continuous functions with left-hand limits, as per a natural extension of Ex. 9.12. Show that there is finite-dimensional convergence of $Z_m(t)$ to the process $Z(t) = \sigma W(t)/t$, where $W(t)$ is Brownian motion on $[1,c]$. Demonstrate also tightness of Z_m, implying full process convergence $Z_m \to_d Z$ in each $D[1,c]$.

(e) Deduce from this that for each $c > 1$, $M_{m,c} = m^{1/2} \max_{1\leq t\leq c} |U_{[mt]}|$ tends in distribution to $M_c = \max_{1\leq t\leq c} \sigma|W(t)|/t = \sigma \max_{1/c\leq s\leq 1} |W^*(s)|$, where $W^*(s) = sW(1/s)$ is the time-transformed Brownian motion of Ex. 9.11. Show furthermore that $\max_{t\geq c} |W(t)|/t \to_{\rm pr} 0$ as $c \to \infty$.

(f) We are close to establishing that $M_m = m^{1/2} \max_{n\geq m} |\bar{U}_n|$ tends to $M = \sigma W_{\max}$, where $W_{\max} = \max_{0\leq s\leq 1} |W(s)|$ is maximum absolute Brownian motion on the unit interval. We have $M_{m,c} \to_d M_c$ for each c, and $M_c \to_d M$, and see via the two-step method laid out in Ex. 2.32 that a probability bound is needed for $M_m - M_{m,c}$. To this end, we shall in the following point establish that

$$p_m(a) = \Pr(m^{1/2} \max_{n\geq m} |\bar{U}_n| \geq a) \leq 6.75\,\sigma^2/a^2.$$

Show that this implies

$$\Pr(|M_m - M_{m,c}| \geq \delta) \leq \Pr(m^{1/2} \max_{m\geq cm} |\bar{U}_n| \geq \delta) \leq 6.75\,\sigma^2/(c\delta^2),$$

and that this secures the intended $M_m \to_d \sigma W_{\max}$.

(g) Let $q > 1$, and for given m find the k with $q^k \leq m < q^{k+1}$. Writing $\bar{U}_n = S_n/n$ and using the Kolmogorov inequality of Ex. 9.6, show that

$$p_m(a) \leq \sum_{i=k}^{\infty} \Pr(\max_{q^i\leq n<q^{i+1}} |S_n| \geq aq^i/m^{1/2}) \leq \sum_{i=k}^{\infty} \frac{q^{i+1}\sigma^2}{(aq^i/m^{1/2})^2} = \frac{\sigma^2}{a^2}\frac{q^3}{q-1},$$

and establish the bound above.

(h) (xx need polish; round off. perhaps we don't include the $W_{\max}$ distribution directly in Ch10. xx) Going back to N_ε, show via results above that $\varepsilon N_\varepsilon^{1/2} \to_d \sigma W_{\max}$, and hence also $\varepsilon^2 N_\varepsilon \to_d \sigma^2 W_{\max}^2$, as $\varepsilon \to 0$. We know the latter's distribution, as per Ex. 9.52. The 0.95 point in the distribution of $W_{\max}$ is 2.241. Show that an approximate answer to Stout's physician's question (vii.2) is $N_0 \doteq 2.241^2\sigma^2/\varepsilon^2$, with $\sigma^2 = p(1-p)$.

(i) There are various generalisations and extensions of the result derived here for N_ε, the last ε. In particular, Hjort and Fenstad (1992) demonstrate that for classes of estimators for which $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d {\rm N}(0,\sigma^2)$, we also have $\varepsilon^2 N_\varepsilon \to_d \sigma^2 W_{\max}$. Show this, for the case of $\widehat{\theta}_n$ being a smooth function of averages. The setup is $\widehat{\theta}_n = h(\bar{A}_{1,n}, \ldots, \bar{A}_{k,n})$, a smooth function of $\bar{A}_{j,n} = (1/n)\sum_{i=1}^n A_{j,i}$, where $(A_{1,i}, \ldots, A_{k,i})^{\rm t}$ has mean ξ, variance matrix Σ, and $h(\xi) = \theta_0$. Show first that $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d {\rm N}(0,\sigma^2)$, with $\sigma^2 = c^{\rm t}\Sigma c$, in which $c = \partial h(\xi)/\partial a$; this is the delta method and the multidimensional CLT. Then prove $\varepsilon^2 N_\varepsilon \to_d \sigma^2 W_{\max}$. See Hjort and Fenstad (1992); Grønneberg and Hjort (2012) for further results.

Story vii.8 *Boys, girls, and mathematics scores.* PISA (Programme for International Student Assessment) is an international study of the competence level of 15-year olds, in reading, mathematics, and other fields, for member nations of the Organisation for Economic Co-operation and Development (OECD). Here we will not attempt to access more accurate PISA data, but learn how to squeeze out valuable information from very brief and partial information, given in a 2023 Norwegian newspaper story. It concerned results from the mathematics tests, and reported that for the *lower score* box there were 33 percent boys and 30 percent girls, compared to 8 percent boys and 5 percent girls for the *very high score* box. Our operating assumptions are (i) that there are underlying continuous scale normal distributions for the pupils' mathematics skills; and (ii) that the lower score and very high score categories correspond to $X \le 2.50$ and $X \ge 5.50$, on the common scale 1-2-3-4-5-6 used in Norwegian schools.

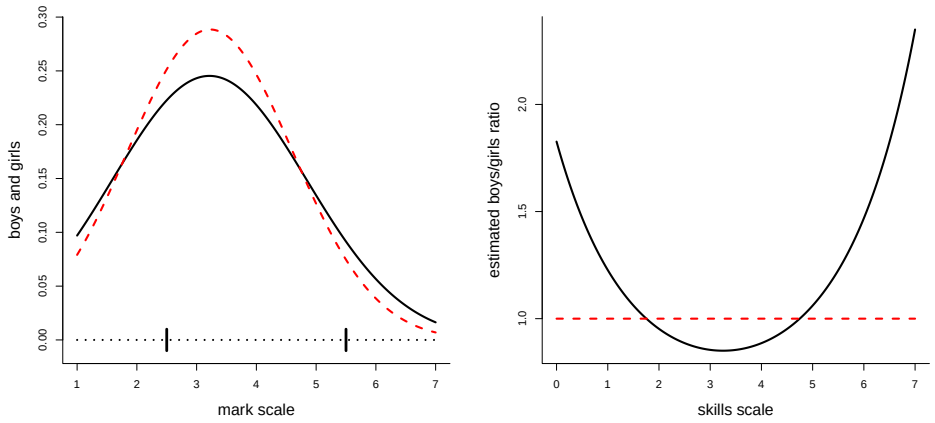


Figure vii.9: *Left panel: estimated math skills densities for Norwegian boys and for girls, age 15, where the Norwegian school mark scale is 1-6. The low score and high score categories are* ≤ 2.50 *and* ≥ 5.50. *Right panel: estimated boys/girls ratio, along the skills scale, under the normality presumption.*

(a) Consider such a math skills distribution, say $X \sim \mathrm{N}(\mu, \sigma^2)$. Under the assumptions made, show that the probabilities of landing in the two categories are $p_1 = \Phi((2.5-\mu)/\sigma)$ and $p_2 = 1 - \Phi((5.5-\mu)/\sigma)$. With estimates for or knowledge of p_1 and p_2, show that

$$(2.5-\mu)/\sigma = a_1 = \Phi^{-1}(p_1), \quad (5.5-\mu)/\sigma = a_2 = \Phi^{-1}(1-p_2).$$

Solve these, with $c = a_2/a_1$, to find

$$\mu = (2.5\,c - 5.5)/(c-1), \quad \sigma = (5.5-\mu)/a_2.$$

(b) With (p_1, p_2) equal to the given $(0.33, 0.08)$ for boys and $(0.30, 0.05)$ for girls, solve the equations, and give a plot of the implied normal densities f_B and f_G, as in Figure vii.9, left panel. Note that $\sigma_B = 1.626$ is substantially bigger than $\sigma_G = 1.383$, whereas the mean parameters are very close.

(c) In order to discuss confidence and significance, assume that the numbers have arisen as $(\widehat{p}_1, \widehat{p}_2) = (N_1/N, N_2/N)$, in a trinomial setup, with (N_1, N_2) the counts in categories

1, 2, based on the full sample size N. Explain how the above corresponds to formulae for estimators $\widehat{\mu}$ and $\widehat{\sigma}$, in terms of $(\widehat{p}_1, \widehat{p}_2)$. Via the delta method and the CLT for multinomials, see Ex. 2.48, show that $\widehat{\sigma} \approx_d \mathrm{N}(\sigma, \tau^2/N)$, with $\tau^2 = c^{\mathrm{t}}\Sigma c$. Here Σ is the trinomial covariance matrix, with $p_1(1-p_1)$ and $p_2(1-p_2)$ on the diagonal and $-p_1p_2$ outside, and c is the gradient of $\sigma(p_1, p_2)$. Using perhaps numerical methods for computing this estimated gradient, compute τ_B and τ_G for the boys and the girls.

(d) For the Norwegian PISA data, behind the numbers $(0.33, 0.08)$ and $(0.30, 0.05)$ above, it is reported that there were about 3000 boys and 3000 girls in this particular study. Give 99 percent intervals for σ_B and σ_G, and also a 99 percent interval for the ratio $\rho = \sigma_B/\sigma_G$. Construct also a confidence curve for ρ. Comment on your findings.

(e) (xx something a bit more, with reference to perhaps more PISA data. connect carefully to variability hypothesis. under the normality presumption, show that the boys-to-girl ratio, as a function of skill level x, takes the form $r(x) = f_B(x)/f_G(x)$. construct a version of the Figure vii.9, right panel. somewhere we also have an instructive thing with $f_2(x)/f_1(x)$ for normal distributions with the same σ. if $f_1 = \mathrm{N}(\mu, \sigma^2)$ and $f_2 = \mathrm{N}(\mu + d, \sigma^2)$, then

$$r(x) = f_2(x)/f_1(x) = \exp\{\tfrac{1}{2}(d/\sigma^2)(x - \tfrac{1}{2}d)\}.$$

find the threshold c with the property that if $x \geq c$, then one expects 10 times as many type 2 compared to type 1. xx)

Story vii.9 *Lean body mass, percent body fat, and correlations.* We consider a dataset pertaining to $n = 37$ Australian rowers, giving in addition to gender the values of x, lean body fat, and y, percent body fat (xx point to ais dataset in data overview xx). The left panel of Figure vii.10 indicates perhaps that the two rowers with smallest x are to be seen as outliers. Below we shall care about robust assessment of the correlation ρ between x and y, and shall learn in the process that the direct correlation might be misleading. These endeavours also have consequences for regressing y on x, e.g. for the purposes of predicting y from x.

(a) Find from the data that the two individuals with smallest x are no. 16 (with 41.54) and no. 30 (with 48.00). Show first that the usually estimated correlation is $\widehat{\rho} = -0.790$, for the full dataset, and a sharper $\widehat{\rho}_r = -0.915$ when these two are removed. Using linear regression for y on x, show that these two are indeed outliers, with too extreme values of $\{y_i - (\widehat{a} + \widehat{b}x_i)\}/\widehat{\sigma}$. Then run linear regression for y on x for the reduced dataset, and argue that none of the 35 thus included are outliers. Construct a version of Figure vii.10, left panel, with the full-data and reduced-data regression lines.

(b) In order to estimate ρ robustly, without necessarily pointing to or indeed knowing about the two rowers with smallest x, fit the 37 (x_i, y_i) points to the binormal distribution, using the robust machinery of BHHJ, see Ex. 5.9. Explain that this for given balance parameter a amounts to minimising the criterion function

$$H_n(\theta) = (2\pi)^{-a}(1+a)^{-1}|\Sigma|^{-a/2} - (1 + 1/a)\, n^{-1} \sum_{i=1}^{n} f(x_i, y_i, \theta)^a + 1/a,$$

where $\theta = (\xi_1, \xi_2, \sigma_1, \sigma_2, \rho)$ and f indicates the binormal density. Carry out such minimisation, for $a = 0.01, \ldots, 0.40$ (the limiting case of $a = 0$ corresponds to ML estimation),

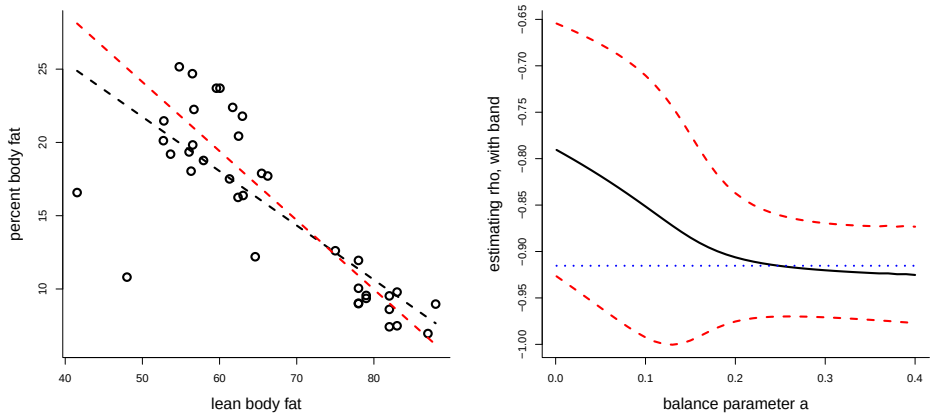


Figure vii.10: *Left panel: lean body fat and percent body fat, for 37 Australian rowers, with two apparent outliers. The two regression lines are for the full dataset (slope* -0.371*) and for the cleaned dataset with the two outliers removed (slope* -0.472*). Right panel: robustly estimated correlation* $\widehat{\rho}(a)$*, along with pointwise 90 percent intervals, computed via the BHHJ methods. The estimated correlation goes from the ML value* -0.790 *to values close the the curated dataset correlation* $\widehat{\rho}_r = -0.915$*, indicated with the horizontal line.*

and plot the resulting $\widehat{\rho}(a)$. Note how this curve starts at the ML value -0.790 and then glides towards values compatible with $\widehat{\rho}_r = -0.915$, without having used any knowledge or speculation about the two alleged outliers, rowers 16 and 30.

(c) Using theory and tools from Ex. 5.19, 5.28, compute the associated approximate standard deviation $\widehat{\tau}(a)/\sqrt{n}$ for $\widehat{\rho}(a)$, and construct a version of Figure vii.10, right panel, with 90 percent intervals. This requires computing the matrices $\widehat{J}_a$, $\widehat{K}_a$, $\widehat{\Sigma}_a = \widehat{J}_a^{-1}\widehat{K}_a\widehat{J}_a^{-1}$ for each a, where K_a estimation benefits from numerical derivation. Compare the usual ML theory, which gives estimate -0.790 and interval $[-0.927, -0.654]$, with the robust BHHJ for balance parameter $a = 0.25$, which gives estimate -0.915 and interval $[-0.970, -0.861]$, and comment.

(d) We now bring gender into the picture, sorting the data into $n_b = 15$ boys and $n_g = 22$ girls, and find a different picture. Construct a version of Figure vii.11, left panel, with regression lines for the two groups. Show that correlations inside the two groups are much smaller in size than for the full dataset, with -0.480 for boys and -0.173 for girls. Carrying out linear regressions, show that all values of $\{y_i - (\widehat{a} + \widehat{b}x_i)\}/\widehat{\sigma}$ are inside the normal range, i.e. there are no outliers per se. Spin this substory as a cautionary tale about correlations.

(e) From a linear regression perspective, individuals 16 (a girl, with $x = 41.54$) and 30 (a boy, with $x = 48.00$) are not necessarily to be judged as outliers, though they are as seen from their respective x distributions. Carry out BHHJ binormal fitting, to compute $\widehat{\rho}(a)$, for the two groups. Construct a version of Figure vii.11, right panel. Note that there for the boys is a sharp discontinuity in the BHHJ estimates, at around $a = 0.079$,

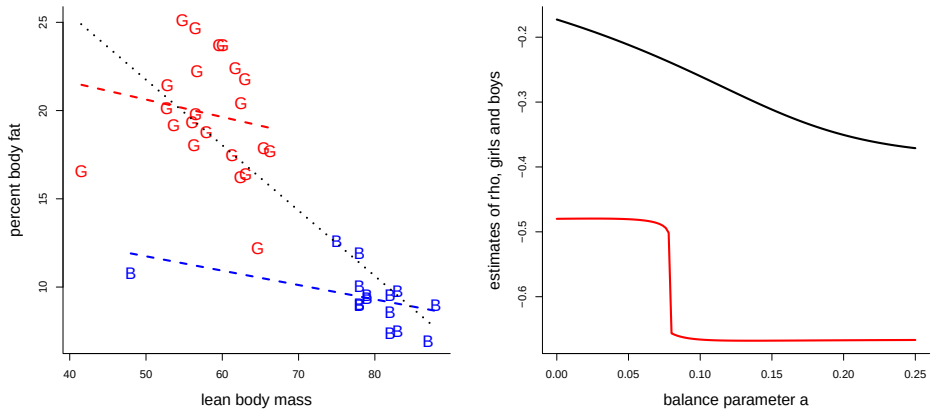


Figure vii.11: *Left panel: the same 37 datapoints as in Figure vii.10, left panel, but now sorted into 15 boys (correlation* -0.480*) and 22 girls (correlation* -0.173*). The two separate regression lines are plotted, in addition to the overall regression line, which does not capture what goes on in the two groups. Right panel: curves of robustly estimated* $\widehat{\rho}(a)$ *for girls and for boys, starting at respectively* -0.173 *and* -0.480*, using binormal fitting with BHHJ.*

even though the minimum criterium values $H_n(\widehat{\theta}(a))$ is continuous. Investigate what is happening and why.

Story vii.10 *Estimating the normal scale: Sir Arthur Eddington vs. Sir Ronald Fisher, 1920.* Suppose $Y_1, \ldots, Y_n$ are i.i.d. from the normal (ξ, σ^2). How should one estimate and carry out inference for the σ? We know the familiar answers, at least if the normality assumption can be trusted, namely to use the minimum sum of squares $Q_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_m$, with $m = n-1$ degrees of freedom. The canonical unbiased estimator for σ^2 is $\widehat{\sigma}^2 = Q_0/(n-1)$, we can test, put up a full confidence distribution, etc. Matters were not so clear a hundred years ago, however, and there is an interesting clash of views and personalities between Sir Arthur Eddington and Sir Ronald Fisher. In his book Eddington (1914), about the structure of the universe, no less, Eddington claimed using sum of absolute deviations, i.e. $n^{-1} \sum_{i=1}^n |Y_i - \bar{Y}|$, is best, whereas Fisher correctly calculated that sum of squares fares better, reported on in Fisher (1920). In interesting historical studies, Stigler (1973, 2006) explains the background and the details, also arguing that this particular problem paved the way for Fisher to both develop *sufficiency* and to become the ultrainfluential mathematical statistician posterity sees him as. We use the opportunity here to go through various details and related follow-up questions.

(a) We start sorting out matters for the case of the mean parameter ξ being known, which we for analysis purposes then can take to be zero. Using results from Ex. 1.30, show that $\mathrm{E}\,|Y_i|^p = c_p \sigma^p$, with $c_p = (\frac{1}{2})^{p/2} \Gamma(p+1)/\Gamma(\frac{1}{2}p+1)$. Here p can be any positive power parameter, not necessarily an integer. Explain that this makes

$$\widehat{\sigma}_p = (M_{n,p}/c_p)^{1/p}, \quad \text{with } M_{n,p} = n^{-1} \sum_{i=1}^n |Y_i|^p,$$

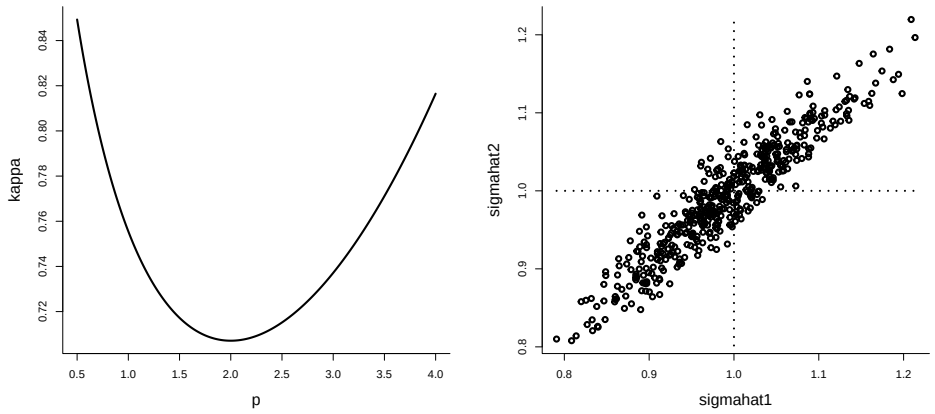


Figure vii.12: *Left panel: limiting standard deviation κ_p, for the $\widehat{\sigma}_p$ estimator of Ex. vii.10; the best value is Fisher's $p = 2$, but the differences are not big. Right panel: 500 simulated realisations of $(\widehat{\sigma}_1, \widehat{\sigma}_2)$, for $n = 100$ and true value $\sigma = 1$. The correlation is 0.936.*

a consistent estimator for σ. Two natural cases would be $\widehat{\sigma}_1 = M_{n,1}/c_1$ and $\widehat{\sigma}_2 = (M_{n,2})^{1/2}$, with $c_1 = (2/\pi)^{1/2}$.

(b) Use the CLT to show that $\sqrt{n}(M_{n,p}/c_p - \sigma^p) \to_d \mathrm{N}(0, \tau_p^2)$, with $\tau_p^2 = (c_{2p}/c_p^2 - 1)\sigma^{2p}$. Then use the delta method to establish that $\sqrt{n}(\widehat{\sigma}_p - \sigma) \to_d \mathrm{N}(0, \kappa_p^2\sigma^2)$, with $\kappa_p^2 = (c_{2p}/c_p^2 - 1)/p^2$. Compute and graph the κ_p function, as with Figure vii.12, left panel, and verify that its minimum is for $p = 2$.

(c) Assume one needs to test $\sigma = \sigma_0$ against $\sigma \neq \sigma_0$. Show that the test statistic $Z_{n,p} = \sqrt{n}(\widehat{\sigma}_p - \sigma_0)/(\kappa_p\widehat{\sigma}_p)$ tends to the standard normal under the null hypothesis. One therefore rejects σ_0 if $|Z_{n,p}| \geq 1.96$, say, the upper 0.025 quantile of the standard normal. For alternatives $\sigma = \sigma_0 + \delta/\sqrt{n}$, use techniques and arguments from Ex. 4.15 to show that the power function converges to $\pi_p(\delta) = \Pr(\chi_1^2(\delta^2/(\sigma_0^2\kappa_p^2)) \geq 1.96^2)$. Compute and graph a few of these, for $\sigma_0 = 1$, say for $p = 1.0, 1.5, 2.0, 2.5, 3.0$. The power curves are close, since the κ_p values are not much bigger than the optimal value κ_2.

(d) Find the joint limit distribution for Eddington's $\widehat{\sigma}_1$ and Fisher's $\widehat{\sigma}_2$, and show that their correlation is high, equal to 0.936. Fisher developed the concept of sufficiency in connection with solving the σ estimation problem; show that $\sum_{i=1}^n Y_i^2$ is sufficient, in the setting above, implying that $p = 2$ is the best value.

(e) The analyses above actually contain the essence also for the more realistic case with unknown ξ, basically since the difference between $n^{-1}\sum_{i=1}^n |Y_i - \widehat{\xi}|^p$ and $n^{-1}\sum_{i=1}^n |Y_i - \xi|^p$ is small enough, for both $p = 1$ and $p = 2$. To see this for $p = 2$ we are helped by simple algebra; show that $n^{-1}\sum_{i=1}^n (Y_i - \xi)^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 + (\bar{Y} - \xi)^2$. Writing $\widehat{\sigma}_2^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ and $\widehat{\sigma}_{2,0}^2 = n^{-1}\sum_{i=1}^n (Y_i - \xi)^2$, show that this implies

$$\sqrt{n}(\widehat{\sigma}_2^2 - \widehat{\sigma}_{2,0}^2) = \sqrt{n}(\bar{Y} - \xi)^2 \sim \sigma^2\chi_1^2/\sqrt{n}.$$

Explain that this leads to $\sqrt{n}(\widehat{\sigma}_2 - \sigma) \to_d \mathrm{N}(0, \kappa_2^2\sigma^2)$, the same limit as reached above, for known ξ.

(f) The parallel case of $p = 1$ is more delicate, however. To learn a bit more, we go via the median M_n, the minimiser of $\sum_{i=1}^n |Y_i - \xi|$ over all ξ. With notation and results from Ex. 5.36, in particular using $D(y)$ equal to 1 for $y \le \xi$ and -1 for $y > \xi$, consider the random convex function

$$\begin{aligned} A_n(s) &= \sum_{i=1}^n \{|Y_i - (\xi + s/\sqrt{n})| - |Y_i - \xi|\} \\ &= \sum_{i=1}^n \{D(Y_i)s/\sqrt{n} + R(Y_i, s/\sqrt{n})\} = U_n s + (f_0/\sigma)s^2 + o_{\rm pr}(1), \end{aligned}$$

where $U_n = \sqrt{n}\bar{D} \to_d U \sim {\rm N}(0,1)$, in terms of $\bar{D} = n^{-1}\sum_{i=1}^n D(Y_i)$, and $f_0/\sigma = \phi(0)/\sigma$ the value of the model density at zero. Use theory developed in Ex. 5.36 to show that under the true value of ξ, we have $Q_n(\xi) = \sum_{i=1}^n (|Y_i - \xi| - |Y_i - M_n|) \to_d \frac{1}{4}U^2\sigma/f_0$. Explain that this means that $\widehat{\sigma}_1 = n^{-1}\sum_{i=1}^n |Y_i - M_n|/c_1$ and $\widehat{\sigma}_{1,0} = n^{-1}\sum_{i=1}^n |Y_i - \xi|/c_1$ have the same limit distributions, with $\sqrt{n}(\widehat{\sigma}_1 - \sigma) \to_d {\rm N}(0, \kappa_1^2\sigma^2)$.

(g) So Fisher wins the argument, $\widehat{\sigma}_2$ being more precise than $\widehat{\sigma}_1$. The difference is not very decisive, however; show that confidence intervals for σ, using the Eddington estimator $\widehat{\sigma}_1$, tend to be 1.068 times wider than for those based on the Fisher estimator $\widehat{\sigma}_2$. Show furthermore that the limiting correlation is as high as 0.936. Create a version of Figure vii.12 (right panel), with simulated realisations of $(\widehat{\sigma}_1, \widehat{\sigma}_2)$ for sample size $n = 100$.

(h) Above we used results for random convex functions to prove that $G_n = n^{-1/2}\sum_{i=1}^n (|Y_i - M_n| - |Y_i - \xi|) \to_{\rm pr} 0$; indeed we learned that $\sqrt{n}G_n$ has a limit distribution, proportional to a χ^2_1. The $G_n \to_{\rm pr} 0$ was a technical necessity for showing that $\sqrt{n}(\widehat{\sigma}_1 - \sigma)$ and the simpler $\sqrt{n}(\widehat{\sigma}_{1,0} - \sigma)$ have the same limit distributions. Eddington's 1914 estimator was in reality $n^{-1}\sum_{i=1}^n |Y_i - \bar{Y}|/c_1$, however, not the median based $n^{-1}\sum_{i=1}^n |Y_i - M_n|/c_1$. To prove that the limit distribution is not affected by the particular choice of centre estimator, consider $G_n = n^{-1/2}\sum_{i=1}^n (|Y_i - \widehat{\xi})| - |Y_i - \xi|)$, with $\widehat{\xi}$ any estimator with a limit distribution for $\sqrt{n}(\widehat{\xi} - \xi)$. For the following, aiming to show $G_n \to_{\rm pr} 0$, let for simplicity $\xi = 0$. Use terminology and details from Ex. 5.36 to show that

$$|y_i - \varepsilon| - |y_i| = D(y_i)\varepsilon + R(y_i, \varepsilon),$$

where $R(y_i, \varepsilon) = 0$ for the many cases where $|y_i| > |\varepsilon|$, and for the remaining cases we have

$$R(y_i, \varepsilon) = \begin{cases} 2(\varepsilon - y_i) & \text{if } 0 \le y_i \le \varepsilon, \text{ if } \varepsilon > 0, \\ 2(y_i - \varepsilon) & \text{if } \varepsilon < y_i \le 0, \text{ if } \varepsilon < 0. \end{cases}$$

Explain that $G_n = \sqrt{n}\bar{D}\widehat{\xi} + n^{-1/2}\sum_{i=1}^n R(Y_i, \widehat{\xi})$, where the first term is $O_{\rm pr}(1/\sqrt{n})$, so it remains to show that also the second term goes to zero. By going through the cases $\widehat{\xi} > 0$ and $\widehat{\xi} < 0$ separately, establish that $n^{-1/2}|\sum_{i=1}^n R(Y_i, \widehat{\xi})| \le 4(C_n/\sqrt{n})|\widehat{\xi}|$, where C_n counts the number of times $|Y_i| \le |\widehat{\xi}|$. Argue that since $\sqrt{n}|\widehat{\xi}|$ has a limit distribution, the remaining detail is to establish that $C_n/n \to_{\rm pr} 0$. Prove this.

(i) Use the theory of influence functions, as with (5.17), to establish that the $\widehat{\sigma}_p$ estimator has influence function ${\rm IF}_p(F, y) = \sigma(1/p)(|y|^p/{\rm E}\,|Y|^p - 1)$. Draw these, for $p = 1$ and $p = 2$, and discuss why this indicates that the Eddington estimator is more robust than the Fisher estimator.

Story vii.11 *Do the data come from f_0 or from f_1?* The Elo rating system, named after Hungarian-American chess player and physics professor Élö Árpád (1903–1992), is used by the World Chess Federation for giving players ratings on a well-defined scale. The system works with databases of pairwise matches, ending with 'A wins', 'draw', 'B wins', and is used also for other sports with similar data outcomes, and even in Large Language Models. In its simplest form, one imagines that when players A and B are about to meet, there are associated random variables X and Y, centred at their current ratings a and b, and that the difference can be expressed as $Z = X - Y = a - b + V$, with V having a continuous distribution symmetric around zero. With H the c.d.f. for V, one furthermore assumes

$$\begin{aligned}
\Pr(\text{A loses}) &= \Pr(Z \le -c) = H(-c-(a-b)),\\
\Pr(\text{draw}) &= \Pr(-c < Z < c) = H(c-(a-b)) - H(-c-(a-b)),\\
\Pr(\text{A wins}) &= \Pr(Z \ge c) = 1 - H(c-(a-b)),
\end{aligned}$$

with c set such that the draw probability $H(c) - H(-c)$ is a reasonable number, like $1/3$, if $a = b$. The system also uses formulae to update a and b after each new match, etc. We do not go into the details here, but note that earlier rating system versions were constructed using a normal distribution for $X - Y$. It was seen that this led to inaccurate predictions for weaker players meeting stronger players, however, and the current Elo system instead uses the logistic distribution, see Ex. 1.61. The difference between a normal and a logistic density, with equal means and variances, is very slight, though. This motivates a study of methods for determining whether data come from density f_0 or density f_1, when perhaps the two are rather close.

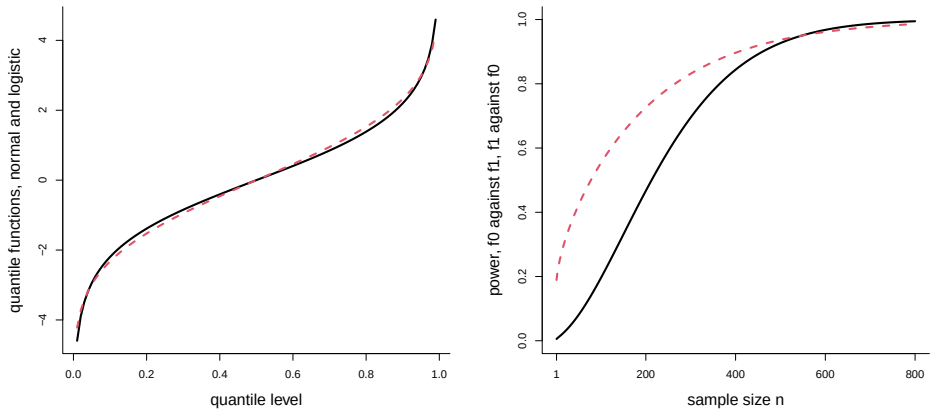


Figure vii.13: *Left panel: the quantile functions $F_0^{-1}(q)$ and $F_1^{-1}(q)$, for the logistic (full curve) and the scaled normal (slanted); they are very close. Right panel: power functions, for testing f_0 against f_1 (full curve, starting lower) and for testing f_1 against f_0 (slanted, starting higher), as a function of sample size, 1 to 800.*

Suppose i.i.d. data $Y_1, Y_2, \ldots$ are observed, these coming from a density f which is either f_0 or f_1, two specified densities on the same domain. We may use the Neyman–Pearson theory of Ch. 4 to test f_0 against f_1, and also vice versa. Below we assume that

the two variances

$$\tau_{0,1}^2 = \mathrm{Var}_{f_0} \log\{f_0(Y)/f_1(Y)\} = \int f_0\{\log(f_0/f_1) - d_{0,1}\}^2 \,\mathrm{d}y,$$
$$\tau_{1,0}^2 = \mathrm{Var}_{f_1} \log\{f_1(Y)/f_0(Y)\} = \int f_1\{\log(f_1/f_0) - d_{1,0}\}^2 \,\mathrm{d}y$$

are finite, involving the two Kullback–Leibler distances

$$d_{0,1} = \mathrm{KL}(f_0, f_1) = \int f_0 \log(f_0/f_1)\,\mathrm{d}y \quad \text{and} \quad d_{1,0} = \mathrm{KL}(f_1, f_0) = \int f_1 \log(f_1/f_0)\,\mathrm{d}y.$$

These distances are positive; see Ex. 5.6 for details.

(a) Let $R_n = \sum_{i=1}^n \log\{f_1(Y_i)/f_0(Y_i)\}$. Show that the optimal test for $H_0\colon f = f_0$ against $f = f_1$ is to reject when R_n is large enough, say $R_n \geq b_n$, with $\Pr_{f_0}(R_n \geq b_n) = 0.05$. Conversely, show that the optimal strategy for testing $f = f_1$ against $f = f_0$ is to reject if R_n is small enough, say $R_n \leq a_n$, with $\Pr_{f_1}(R_n \leq a_n) = 0.05$. The thresholds a_n and b_n may be found from the distribution of R_n under respectively f_0 and f_1 circumstances, e.g. via simulations, for each given n; below we use normal approximations.

(b) Show that

$$R_n/n \to_{\mathrm{pr}} \begin{cases} -d_{0,1} = -\mathrm{KL}(f_0, f_1) & \text{if data are from } f_0, \\ d_{1,0} = \mathrm{KL}(f_1, f_0) & \text{if data are from } f_1. \end{cases}$$

So a plot of R_n, as a function of growing sample size, will end up positive under f_1 but negative under f_0.

(c) Before we come to the more complicated situation of testing the normal against the logistic, analyse the simple setup with $f_0 = \mathrm{N}(0,1)$ and $f_1 = \mathrm{N}(\theta,1)$, where θ is fixed and positive. Show that $R_n = \sum_{i=1}^n (\theta Y_i - \frac{1}{2}\theta^2)$; find the limits of R_n/n under f_0 and f_1; find the precise rejection thresholds a_n and b_n; and find the power functions, i.e. the probability of rejection at f_1 when testing $f = f_0$ and at f_0 when testing $f = f_1$.

(d) Returning to the general setup, explain that with data from f_0, we have $\sqrt{n}(R_n/n + d_{0,1}) \to_d \mathrm{N}(0, \tau_{0,1}^2)$, whereas $\sqrt{n}(R_n/n - d_{1,0}) \to_d \mathrm{N}(0, \tau_{1,0}^2)$ if data are from f_1. From this, explain that good approximations to the rejection thresholds above are (i) to reject f_0 for f_1 if $\sqrt{n}(R_n/n + d_{0,1}) \geq \tau_{0,1} z_0$ and (ii) to reject f_1 for f_0 if $\sqrt{n}(R_n/n - d_{1,0}) \leq -\tau_{1,0} z_0$, with z_0 the relevant upper quantile of the standard normal, like 1.645 for significance testing level 0.05.

(e) Next we investigate the detection power for these two tests. A crucial component here is the sum of distances $\delta = d_{0,1} + d_{1,0}$; the bigger the δ, the easier for the tests. For testing f_0 against f_1, show that

$$\begin{aligned} \Pr_{f_1}(\text{reject } f_0) &= \Pr_{f_1}(\sqrt{n}(R_n/n - d_{1,0} + \delta) \geq \tau_{0,1} z_0) \\ &\doteq \Pr(\tau_{1,0} N \geq \tau_{0,1} z_0 - \sqrt{n}\delta) = \Phi(\sqrt{n}\delta/\tau_{1,0} - (\tau_{0,1}/\tau_{1,0}) z_0), \end{aligned}$$

writing N for a standard normal. Similarly, show that

$$\begin{aligned} \Pr_{f_0}(\text{reject } f_1) &= \Pr_{f_0}(\sqrt{n}(R_n/n + d_{0,1} - \delta) \leq -\tau_{1,0} z_0) \\ &\doteq \Pr(\tau_{0,1} N \leq \tau_{1,0} z_0 + \sqrt{n}\delta) = \Phi(\sqrt{n}\delta/\tau_{0,1} - (\tau_{1,0}/\tau_{0,1}) z_0). \end{aligned}$$

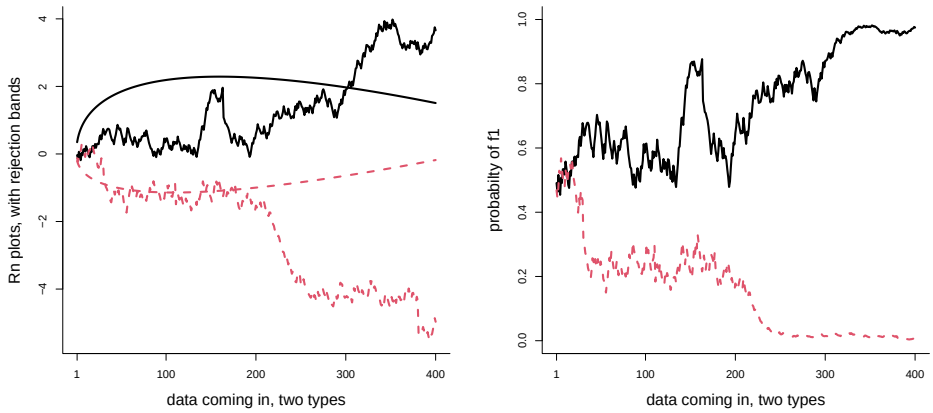


Figure vii.14: *Left panel: plot of R_n, with 400 data points from f_1 (irregular full curve), and with 400 data points from f_0 (irregular slanted curve). Also plotted are the rejection limits, as a function of sample size: R_n needs to be above the full regular curve in order to reject f_0 and claim f_1; and similarly needs to be below the slanted regular curve in order to reject f_1 and claim f_0. Right panel: for the same two datasets, the Bayesian probability of data having come from f_1, as a function of sample size; the upper curve is for f_1 data, the lower for f_0 data.*

(f) Consider now Model 0, with f_0 the logistic density, see Ex. 1.61, and Model 1, with f_1 the $\mathrm{N}(0, \tau^2)$, where we take variance $\tau^2 = \pi^2/3$ to match that of the logistic. The two distributions are rather close, making it a hard challenge to identify which data come from which source. Compute the KL distances and associated variances numerically; one finds $d_{0,1} = 0.01436$, $d_{1,0} = 0.01049$, $\tau_{0,1} = 0.22046$, $\tau_{1,0} = 0.13309$. Construct versions of Figure vii.13, with quantile functions to the left and the two power functions to the right. One needs a fair amount of data from f_1 to have a high probability of claiming that f_0 is not correct, and similarly a fair amount of data from f_0 to have a high probability of caliming that f_1 is not correct. Verify in fact that one needs $n = 365$ data from f_1 to be 80 percent sure of detecting that they are not from f_0, and correspondingly $n = 265$ data from f_0 to be 80 percent sure of detecting that they are not from f_1. Also generate 400 data from f_1, leading to an R_n plot, to see when $f = f_0$ is rejected, and similarly 400 data from f_0, to see when $f = f_1$ is rejected; create a version of Figure vii.15, left panel.

(g) The problem of classifying incoming data as coming from f_0 or f_1 might also be handled in a Bayesian fashion. Suppose f_0 and f_1 are equally likely a priori. Show that

$$\Pr(f_1 \,|\, \text{data}) = \frac{f_1(y_1)\cdots f_1(y_n)}{f_0(y_1)\cdots f_0(y_n) + f_1(y_1)\cdots f_1(y_n)} = \frac{\exp(R_n)}{1+\exp(R_n)}.$$

For the same data points generated above, compute these probabilities, as a function of incoming data, i.e. sample size; construct a version of Figure vii.15, right panel.

(h) Your code for analysing the classifiers and their detection chances ought to be somewhat generic, making it easy to try out other pairs of f_0, f_1. As an illustration, find f_0 and f_1, respectively a $\mathrm{Gam}(a, b)$ and a Weibull (c, d), with the property that the 0.10

and 0.90 quantiles for both are 2.00 and 8.00. Draw the two c.d.f.s in a diagram to see how close they are. Then numerically compute the Kullback–Leibler distances $d_{0,1}$ and $d_{1,0}$, along with the standard deviations $\tau_{0,1}$ and $\tau_{1,0}$. Run the code to learn the power functions, and reflect on how hard it is to see the difference between Gamma and Weibull data.

Story vii.12 *Do the data come from logistic or from probit regression?* In Story vii.11 we used the Elo rating system change, from normal to logistic, as motivation for a careful study of how a statistician may be able to optimally classify incoming data as coming from density f_0 or density f_1, and we learned that this is a tall order, requiring many datapoints, if the two densities are close. For the Elo rating system the challenge is actually significantly harder, as we do not observe the underlying ability differences $Z = X - Y$, merely their thresholded outcomes in the three boxes 'left', 'in the middle', 'right'. This motivates studying how well we may spot the difference between logistic and probit regression (and where we know before starting that the answer will be 'it is hard and requires much data', since the normal and logistic c.d.f.s are close).

(a) Suppose 0-1 regression data (x_i, y_i) are obtained, with $p_i = \Pr(Z_i \le a_0 + b_0 x_i) = \Pr(Y_i = 1 \,|\, x_i)$, with two possibilities for these, in terms of distributions for underlying variables Z_i. We have either Model 0, Z_i following the logistic, corresponding to classical logistic regression and $p_i^{(0)} = H(a_0 + b_0 x_i)$; or Model 1, a scaled normal $(0, \tau^2)$, as with $p_i^{(1)} = \Phi((a_0 + b_0 x_i)/\tau)$, corresponding to probit regression; the scaling is there to have equal variances, $\tau^2 = \pi^2/3$. To make this concrete, suppose (a_0, b_0) are known, and let

$$R_n = \ell_n^{(1)} - \ell_n^{(0)} = \sum_{i=1}^n \Big\{ y_i \log \frac{p_i^{(1)}}{p_i^{(0)}} + (1 - y_i) \log \frac{1 - p_i^{(1)}}{1 - p_i^{(0)}} \Big\}$$

Show that the optimal test for logistic model M_0, against the probit model M_1, is to reject if R_n is big enough, and correspondingly that the optimal test for M_1, against the alternative M_0, is to reject if R_n is small enough.

(b) Suppose the x_i are i.i.d. from some covariate distribution Q. With $p_{\rm true}(x)$ the real $\Pr(Y_i = 1 \,|\, x)$, show that

$$R_n/n \to_{\rm pr} \int \Big[p_{\rm true}(x) \log \frac{p^{(1)}(x)}{p^{(0)}(x)} + \{1 - p_{\rm true}(x)\} \log \frac{1 - p^{(1)}(x)}{1 - p^{(0)}(x)} \Big] \, {\rm d}Q(x).$$

If the data come from the logistic model M_0, show that the limit is negative, say $-d_{0,1}$, and if the data are from the probit model M_1, show that the limit, say $d_{1,0}$, is positive. Argue therefore that for large enough n, the R_n will be negative under M_0 and positive under M_1. To make this concrete, compute these average KL distances $d_{0,1}$ and $d_{1,0}$, for the case of $(a_0, b_0) = (-0.33, 0.66)$, with Q being a normal $\mathrm{N}(0, \tau^2)$; answers: very tiny $d_{0,1} = 0.000742$, $d_{1,0} = 0.000758$.

(c) Similar behaviour can be expected, exhibited, and analysed for the case of unknown parameters in logistic and probit regression. For 0-1 regression data (x_i, y_i), one may fit the logistic model $p_i = H(a + bx_i)$ and the probit normal model $p_i = \Phi((a + bx_i)/\tau)$. The main message will be that it is difficult to spot a clear difference between them; that the fitted regression curves $H(\widehat{a} + \widehat{b}x)$ and $\Phi((\widetilde{a} + \widetilde{b})/\tau)$ will be close; and that one needs a big data volume in order to see that one of the schemes is better than the other. Using

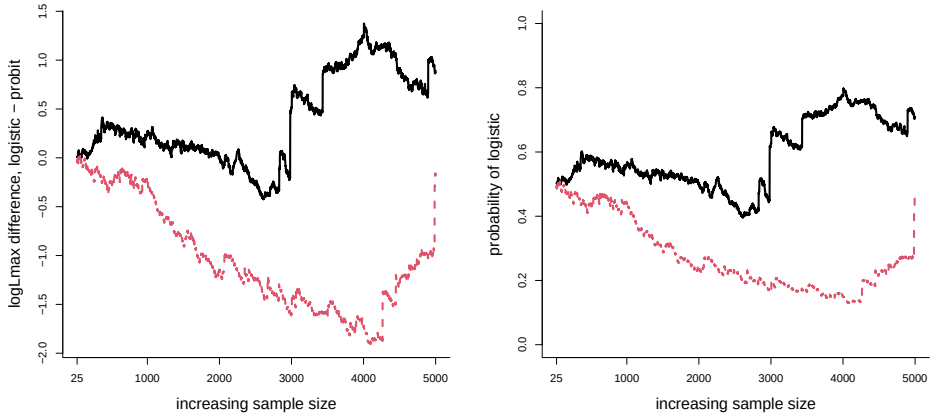


Figure vii.15: *Left panel: the sequence of log-likelihood maxima difference* $\ell_{L,\max} - \ell_{P,\max}$, *as a function of sample size, for a dataset drawn from the logistic model (upper curve, ending positive) and from another dataset drawn from the normal probit model (lower curve, ending negative). Right panel: for the same two datasets, BIC approximation to* $\Pr(\text{logistic} \,|\, \text{data})$, *as a function of sample size.*

L and P subscripts for the logistic and probit models, and in the terminology of Ch. 11, show that

$$\text{aic}_L - \text{aic}_P = 2(\ell_{n,L,\max} - \ell_{n,P,\max}), \quad \text{bic}_L - \text{bic}_P = 2(\ell_{n,L,\max} - \ell_{n,P,\max}),$$

hence equal, here, since the parameter dimension is the same for both models. Run a simulation experiment, taking x_i i.i.d. $\text{N}(0, \tau^2)$, and then creating dataset 0, of big size $n_{\max} = 5000$, with $H(a_0 + b_0 x_i)$, and dataset 1, with $\Phi((a_0 + b_0 x_i)/\tau)$, using $(a_0, b_0) = (-0.33, 0.66)$. Then compute for each $n \le n_{\max}$ the ML estimators in the two models; the log-likelihood maxima $\ell_{n,L,\max}$ and $\ell_{n,P,\max}$; plot the difference $D_n = \ell_{n,L,\max} - \ell_{n,P,\max}$ (which should eventually become positive for logistic generated data but negative for probit generated data); and the associated BIC probability the data stem from the logistic model. Show that the latter is

$$\Pr(\text{logistic} \,|\, \text{data}) = \frac{\exp(\frac{1}{2}\text{bic}_L)}{\exp(\frac{1}{2}\text{bic}_L) + \exp(\frac{1}{2}\text{bic}_P)} = \frac{\exp(D_n)}{1 + \exp(D_n)}.$$

Construct versions of Figure vii.15, left and right panels. You may also plot $(\widehat{a}, \widetilde{a})$ and $(\widehat{b}, \widetilde{b})$ as a function of increasing sample size, to see their similarity, and indeed plot $H(a_0 + b_0 x)$ vs. $\Phi((a_0 + b_0 x)/\tau)$. You may of course also run your code up to an even bigger $n_{\max}$, like 25000, to see how much data it takes for the AIC or BIC to be convinced about which of the two data generating mecanisms is at work.

Notes and pointers

(xx notes and follow-up things for the stories in this chapter. xx)

(xx For Story vii.5, see Stigler (1977). xx)

Part III

Appendix

III.B

Overview of stories and data

A generous number of real datasets are used in this book to illustrate aspects of the methodology being developed, and hence also a correlated generous list of Statistical Stories. Here we provide brief descriptions of each of these real data examples, along with key points to indicate which substantive questions they relate to. Some of these datasets are small, partly meant for simpler demonstrations of certain methods, whereas other are bigger, allowing also more ambitious modelling for reaching inference conclusions. Key words are included to indicate the data sources, the types of model we apply and for what inferential goal, along with pointers to which stories the datasets are analysed.

nils emil notes to themselves, as of 16-January-2025

Not yet well organised, as of 16-January-2025, but we'll get to things and have a clearer overview.

This part of Kioskvelter needs of course serious working-over and polish near the end of the process, and we need to achieve the right level of 'oi!, this might be interesting' and conciseness. Each entry needs (a) brief description of what the data are about, with a source or two; (b) description of which questions arise naturally; (c) briefly worded pointers to which Stories touch these data.

The 2.A list will be updated and polished during autumn 2023. We get the cross-referencing to work for `data:listofstories` etc.

(xx So, rough estimate of magical story number as of 16-January-2025: 18 + 13 + 15 + 8 + 12 + 9 + 11 = 86, but some of those listed 16-January-2025 might not come to fruition. nils pushes mothers babies I and mothers babies II from Demography Eip Medicine to Biology. we're not shying away from StoryX Part I and StoryX Part II to avoid having them too long, etc. our Final Number was meant to be 77 but could even become 88 if several of these are decently short. xx)

List of All Stories

(xx annotated as we proceed. as of 16-January-2025, meant to be helpful for the authors. it will become a better kept list later on, along with brief key words and pointers. the list of these brief dataset descriptions should follow our stories. we also need to monitor the lengths of the stories, though we do accept some shorter ones and some longer ones.

which stories have points more than from (a) to (h)? as of 16-January-2025, there are ok plans for how many stories, in the separate chapters? xx)

I Demography, Epidemiology, Medicine (approx 14)

(1) Cooling of newborns. (nils, 0.90 ferdig)

(2) Overdispersed children. (nils, 0.90 ferdig)

(3) Markovian children. (nils, 0.90 ferdig)

(4) IDU expulsion. (nils, 0.90 ferdig, maa koples til Beta process)

(5) Boys are born slightly bigger than girls. (nils, 0.90 ferdig)

(6) Mothers, babies, birthweights, factors. (nils, 0.90 ferdig)

(7) Mothers, babies, birthweights, logistic or logistic. (nils, 0.90 ferdig)

(8) Mothers, babies, birthweights, birth order. (nils, 0.90 ferdig)

(9) Time to 2nd child after stillborn (nils, 0.90 ferdig)

(10) A third child? (emil, 0.60 ferdig)

(11) PCI and confidence fusion from thirteen studies. (nils, 0.80 ferdig)

(12) Suicide attempt rates for Paroxetine vs. placebo. (nils, 0.85 ferdig)

(13) Pedro og Elvira (emil, 0.50 ferdig)

(14) Onset of menarche (nils, 0.90 ferdig)

(15) Eye retinopathy, FIC for logistic regressions. (nils, 0.85 ferdig)

(16) Cigarette consumption and lung cancer. (nils, 0.85 ferdig)

(17) Norwegian women and men of the future (nils, 0.10 ferdig)

(18) A cure model (emil, 0.20 ferdig)

II Art, History, Literature, Music (approx 11)

(1) Game of Thrones and the Wars of the Roses. (nils, 0.80 ferdig)

(2) Stride towards your bookshelves. (nils, 0.90 ferdig, kan gjoere noe mer)

(3) Tirant lo Blanc: When did Author B take over for Author A? (nils, 0.90 ferdig, maa deales med, linkes til Ch9 monitoring for change)

(4) The children of Odin. (nils, 0.90 ferdig)

(5) How many Abel first-day cover envelopes from 1929? (nils, 0.90 ferdig.)

(6) Markov and Pushkin. (nils, 0.85 ferdig, henter ting fra CLP, med modifikasjoner)

(7) And Quiet Does Not Flow the Don. (nils, 0.90 ferdig)

(8) Republic, Laws, Critias, Philebus, Politicus, Sophist, Timaeus. (nils, 0.85 ferdig, maa stelles med og forkortes og synches med Ch12)

(9) Presidents of the First Republic. (nils, 0.90 ferdig)

(10) Dangerous job assignment: Roman emperor. (nils, 0.90 ferdig)

(11) Lifelengths in Roman Era Egypt, 2100 years ago. (nils, 0.90 ferdig)

(12) Bach, Reger, organ fugues, and Wohltemperierte I und II. (nils, 0.80 ferdig, maa avrundes)

(13) How many piano tuners in Oslo? (nils, 0.75 ferdig)

III Economy, Political Science, Sociology (approx 10)

(1) Power law scaling for academics and support staff. (nils, 0.85 ferdig)

(2) Poisson overdispersion for British mining. (nils, 0.90 ferdig)

(3) Changepoints for British mining. (nils, 0.90 ferdig)

(4) War and Peace and War and Peace, I. (nils, 0.80 ferdig, linkes til boundary parameter things in Ch5, Ch11)

(5) War and Peace and War and Peace, II. (nils, 0.50 ferdig, trenger mer om changepoints, Bayes too)

(6) War and Peace and War and Peace, III. (nils, 0.60 ferdig, men tre angels stories er kanskje en for meget)

(7) Psychiatric disorders and body type. (nils, 0.90 ferdig)

(8) Galton and 111 husbands and wives. (nils, 0.40 ferdig)

(9) Terbeschikkingstelling. (nils, 0.80 ferdig)

(10) Tore Sims Sveriges Riksbank, (nils, 0.90 ferdig)

(11) Does winning make you live longer? (emil, 0.75 ferdig)

(12) Minimum Wages and Employment. (emil, 0.75 ferdig)

(13) How many were killed in Srebrenica, 1991? (nils, 0.80 ferdig)

(14) How many were killed in Guatemala, 1978–1996? (nils, 0.80 ferdig)

(15) Volatility estimation from noisy financial data. (emil, 0.02 ferdig. Har data og vet omtrent hva som skal skrives)

IV Biology, Climate, Ecology (approx 6)

(1) New Haven annual temperatures 1912–1971. (nils, 0.85 ferdig)

(2) Oslo annual temperatures 1901–20024. (nils, 0.85 ferdig)

(3) Where are the snows of yesteryear? (nils, 0.80 ferdig)

(4) Mammals and their bodies and brains. (nils, 0.80 ferdig, need a a little link til Jamtveit story)

(5) Kola temperatures and the Hjort liver index time series 1859-2020. (nils, 0.75 ferdig)

(6) How many Clethrionomys glareoli? (nils, 0.80 ferdig, venter paa Ch7 polishing)

(7) Birds on islands outside Ecuador. (nils, 0.90 ferdig)

(8) Birds on islands, via square-rooting to normal nonlinear regression. (nils, 0.90 ferdig; may be subsumed in previous story)

V Sports (approx 12)

(1) Bolt from heaven. (nils, 0.90 ferdig, venter paa litt Ch7 avrunding)

(2) Will there be a new world record this season? (nils, 0.90 ferdig)

(3) The random angles of golf putting. (nils, 0.90 ferdig)

(4) NBA three point shooting averages. (emil, 0.80 ferdig)

(5) Olympic Unfairness I: Inner and outer for 1000 m speedskating. (nils, 0.80 ferdig, skal ha CD(τ) med punktmasse m.m.

(6) Olympic Unfairness II: From semifinals to finals. (nils, 0.75 ferdig)

(7) Who wins? Computing probabilities as a function match time. (nils, 0.85 ferdig)

(8) The turn-around operation: from 0-2 to 3-2. (nils, 0.85 ferdig, kan kaste inn fotballdata fra Nils-Gerda til sist)

(9) The hot hand in basketball. Myth or not? (emil, 0.05 ferdig. har data og har plottet litt)

(10) When to shoot? An optimal stopping problem in basketball. (emil, 0.02 ferdig. Har et datasett med hvilket dette sporrsmaalet kan besvares)

(11) Winning streaks in chess. (nils-emil, 0.03 ferdig; need book-keeping scripts from Alamaty 2022)

(12) BMI for Olympic speedskaters. (nils, 0.10 ferdig, tas fra II-CC-FF pluss litt mer)

VI Simulated stories (approx 7)

(1) Checking out the CLT. (nils, 0.85 ferdig)

(2) An infinite weighted sum of Bernoullis. (nils, 0.85 ferdig)

(3) Finding magic squares by MCMC. (nils, 0.85 ferdig)

(4) Reconstructing exponential decay beer foam. (nils, 0.85 ferdig)

(5) From falling ill to having recovered. (nils, 0.65 ferdig, vi maa ha noe simulation)

(6) Causal inference and potential outcomes. (emil, 0.33 ferdig, haaper noe kan skrives rundt dette)

(7) Identifying your counterfactual cousin. (nils, 0.80 ferdig)

(8) Recostructing a blurry image. (nils, 0.01 ferdig)

(9) Law, evidence, and Bayes (emil, 0.01 ferdig. Ideen er der. Noe basert paa Calina sin master. Bayesianske nett og noen simuleringer)

VII Miscelannea (approx 6)

(1) The Pearson goodness-of-fit statistic. (nils, 0.85 ferdig, vil ha link til Brownian bridge, og til ML)

(2) Decimals of pi. (nils, 0.75 ferdig, flikker inn dette med aa jukse for hver 1000ende desimal, for aa se om pearson oppdager det, local power)

(3) Random integers via prime numbers. (nils, 0.75 ferdig, gjoer noe mer)

(4) Time-to-failure for machine components. (nils, 0.90 ferdig, taken from Ch6)

(5) Speed of light in 1882, with BHHJ estimation. (nils, 0.85 ferdig)

(6) Stout's physician and the last n. (nils, 0.90 ferdig)

(7) Boys, girls, and mathematics scores (nils, 0.90 ferdig)

(8) Lean body mass, percent body fat, and correlations. (nils, 0.80 ferdig)

(9) Eddington versus Fisher, 1920. (nils, 0.85 ferdig)

(10) Elo rating and how to classify data as f_0 or f_1. (nils, 0.90 ferdig)

(11) Elo rating and logistic versus probit regression. (nils, 0.90 ferdig)

Description of datasets for stories

(xx needs careful work and checking, with crossref. also regarding the order in which we present these datasets. as of 16-January-2025, nils attempts to put these in the order of appearance through the chapters of stories, but this will be modified later. nils tentatively gives numbers to the stories too, perhaps to be take away later. xx)

I (i) The cooling of newborns

[xx Laptook study, point to two smaller papers by Walløe, Thoresen, Hjort, and the Hjort blogpost. Laptook (2017), Walløe et al. (2019a,b). basically: inference for p_1/p_2. return to this in the Bayes chapter, with informative priors on p_1, p_2. basically: $y_0 \sim \text{binom}(m_0, p_0)$ for the noncooled group, 22 of 79 cases; and $y_1 \sim \text{binom}(m_0, p_1)$ for the cooled groups, 19 of 68 cases. xx]

I (ii) Suicide attempt rates, drug vs. placebo

(xx describe data, two Poisson, in to round, from Aursnes et al. (2005, 2006). suicide rates for users of a certain antidepressant, vs. a placebo group. for Story i.12. xx)

I (iii) and (iv) Overdispersed and Markovian children

[xx data from Geißler (1889), on the number of girls in 8-children-families. we demonstrate extra-binomial variability, and assess this degree of overdispersion. see Hjort (2018a). in later chapter also the Markovian model. xx]

y	N	E1	pear1	E2	pear2
0	264	192.325	5.168	255.621	0.524
1	1655	1445.384	5.514	1657.032	-0.050
2	4948	4752.364	2.838	4909.686	0.547
3	8498	8928.902	-4.560	8683.213	-1.988
4	10263	10484.952	-2.168	10024.863	2.378
5	7603	7879.792	-3.118	7735.975	-1.512
6	3951	3701.205	4.106	3896.509	0.873
7	1152	993.421	5.031	1171.238	-0.562
8	161	116.655	4.106	160.865	0.011

(xx In this version of our manuscript we also give the girl-boy data for 195 families worked with in i.3, from Klotz (1972, 1973). In the Real Book we might be content with having only summary statistics. i.e. give our readers $M = 85$, $n - M = 110$, $(N_{0,0}, N_{0,1}, N_{1,0}, N_{1,1}) = (345, 298, 287, 333)$, not bother readers with the full dataset. here 1 is girl and 0 is boy. nei, forresten, nils ombestemmer seg, det tar for meget plass i pdf-en og det hele, so the 195 families can be found here, by nils and emil, but they've been commented away. xx)

I (v) IUD expulsion

(xx describe data, 100 women using IUD, (i, t_i, δ_i), from Peterson (1975). frailty model. check Aalen1978 sjs. used in Story i.4. xx)

I (vi) The 73 French presidents of France 1792–1795

(xx drama following the French Revolution. 73 presidents, some with short lives after elections. two covariates. data from Céline Cunen. Story: ii.9. xx)

Data, for 73 presidents: `id, birth, death, presistart, presiend, v, giro, vip`. Here v is indicator for having experienced a violent death of not (so $\delta = 1 - v$ is indicator for non-censoring, in survival analysis language); giro is indicator for belonging to the Gironde party or not; and vip is a proxy for fame, counted here as the number of languages for which there is wikipedia pages about the president in question (recorded, by Cunen, in October 2017).

I (vii) Boys and girls born in Oslo in 2001-2008

In the larger context of identifying and exploring factors that may influence the chances of babies being born too large, Voldner et al. (2008) examined a certain cohort of 1028 mothers and children (548 boys, 480 girls), all of whom born at Rikshospitalet, Oslo, in the 2001–2008 period. We have had access to the birthweight data in question (via N. Voldner and K.F. Frøslie, personal communication), and use these to illustrate non-parameetric confidence curves for quantiles.

(xx here: too much to show all the 548 + 480 data points, but we point to files, and the figures. xx)

(xx nota bene: will use these data also for other chapters. xx)

I (viii) Lifelengths in Roman era Egypt

In Spiegelberg (1901) the age at death has been recorded for 141 Egyptian mummies, 82 men and 59 women, dating from the Roman period of ancient Egypt from around year 100 B.C. These life-lengths vary from 1 to 96 years, and Pearson (1902) argued that these can be considered a random sample from one of the better-living classes in that society, at

a time when a fairly stable and civil government was in existence. Interestingly, Pearson wrote "in dealing with [these data] I have not ventured to separate the men from the women mortality, the numbers are far too insignificant". That did not stop Claeskens and Hjort (2008b, p. 33–35) from establishing partly different parametric hazard rate models for the two sexes, and where incidentally Gompertz type models were found to be better than e.g. Weibull and gamma models.

(xx then pointers to where we do what. Story i.5 for ratio of quantiles. xx)

```
 1.5  1.8  2.0  2.0  3.0  3.0  3.0  4.0  4.0  4.0  4.0  5.0  5.0  5.0  6.0
10.0 11.0 14.0 14.0 16.0 17.0 17.0 19.0 20.0 20.0 21.0 22.0 22.0 23.0 24.0
24.0 25.0 25.0 25.0 25.0 25.0 26.0 26.0 26.0 26.0 27.0 29.0 30.0 32.0 33.0
33.0 36.0 36.0 37.0 40.0 40.0 40.0 46.0 48.0 48.0 50.0 50.0 50.0 50.0 50.0
52.0 52.0 52.0 55.0 55.0 59.0 60.0 60.0 60.0 60.0 62.0 63.0 65.0 65.0 68.0
68.0 70.5 72.0 72.0 72.0 84.0 90.0  1.5  3.0  3.0  4.0  4.0  4.0  6.0  6.0

 9.0 10.0 11.0 14.0 16.0 17.0 17.0 17.0 18.0 18.0 19.0 19.0 20.0 20.0 20.0
21.0 21.0 21.0 21.0 21.0 21.0 22.0 22.0 23.0 23.0 25.0 25.0 25.0 25.0 26.0
27.0 28.0 29.0 30.0 30.0 33.0 35.0 35.0 35.0 36.0 36.0 40.0 40.0 40.0 50.0
52.0 54.0 55.0 60.0 70.0 96.0
```

I (ix) PCI and confidence fusion from meta-analyses

(xx from Schömig et al. (2008). thirteen two-by-two tables. rare events. xx)

I (xi) Terbeschikkingstelling

(xx from Hjort and Koning (2002), Dutch ombudsman. poisson with changing rates. xx)

I (xii) Brazilian kids

Data from Borgan et al. (2007), come in the three files `prevdat.txt`, `atprevrisk.txt`, and `covariates.txt`. Across these three files, the children are identified by their row number. The first file, `prevdat.txt`, contain sequences of zero and ones, a one indicating the the child on the given row was observed with diarrhoea on the day in question, and zero otherwise. The data file `atprevrisk.txt` contains indicators for noncensoring, 1 indicates that the child was observed on the day in question, 0 indicating that no observation was made (i.e., censoring). The third file contain the covariates:

`sexcat`: Gender of child (0-female , 1-male);

`age.beg`: Age of child in the begining of the study (month);

`dens2`: Number of people/beroom (0-one or two people, 1- >= three people);

`pavcat`: Quality of the street (0- good , 1- bad);

`resagcat`: Type of drinking water-reserve (0-good , 1- bad);

`tipagcat`: Quality of drinking water (0-good, 1-bad);

`valacat`: Presence of holes with dirty water-30m (0-no, 1-yes);

`corrcat`: Presence of small rivers with dirty water-30m (0-no , 1-yes);

`chovecat`: Condition of accomodation during rain (0-good , 1-bad);

`age.mcat`: Age of the mother 0- >= 25 years old, 1- < 25 years old;

`socec2`: Social economic class (0-medium or high, 1- poor);

`no.age5`: Number of other children less and equal 5 (0=none, 1- more than 1).

II (i) The Game of Thrones and the Wars of Roses

[xx via Cunen (2015) blog post and story (bringing her international fame). full dataset: $(t_i, x_{i,1}, x_{i,2}, \delta_i, o_i)$, with t_i the age at death if $\delta_i = 1$ and the last known age for the person if $\delta_i = 0$; $x_{i,1}$ indicator for nobility; $x_{i,2}$ concerns gender, and is 1 for women and 0 for men; indicator for nobility; and o_i is 1 for GoT and 0 for WoR. also with violent-death-or-not, from two worlds. see Story ii.1]. xx]

II (iii) Stride towards your bookshelves

[xx student books oblig story for chapter1: wordlengths. metaanalysis modelling, chapter2, simple analysis in chapter1, and we return to the dataset later, in chapter6. xx]

	xNor	xEng	sdNor	sdEng		xNor	xEng	sdNor	sdEng
1	4.45	2.88	4.53	2.77	33	4.76	3.10	4.10	2.02
2	4.82	2.88	4.87	2.94	34	4.58	2.79	4.03	1.81
3	4.39	2.24	3.94	2.04	35	6.32	4.02	4.49	2.41
4	4.44	2.65	4.02	1.90	36	3.94	1.96	5.30	2.78
5	4.79	3.11	4.56	2.24	37	3.54	1.77	4.73	2.86
6	4.91	3.12	4.05	1.86	38	4.41	2.15	4.16	2.20
7	4.61	2.42	4.64	3.01	39	4.57	2.98	4.28	1.95
8	3.91	1.94	4.48	2.18	40	4.34	2.24	4.31	2.17
9	4.42	2.52	3.94	2.09	41	5.12	3.16	4.33	2.26
10	4.86	2.61	5.21	2.86	42	4.47	2.69	3.81	2.00
11	6.18	4.62	5.43	2.89	43	4.29	2.33	4.08	1.95
12	5.04	3.47	4.97	2.86	44	3.87	2.06	4.72	2.26
13	3.99	2.71	4.33	2.52	45	3.96	2.19	4.20	2.49
14	4.35	2.64	4.82	2.36	46	4.43	2.40	4.74	2.53
15	4.56	2.93	5.23	3.35	47	5.01	2.88	4.47	2.35
16	5.27	3.10	5.16	3.20	48	4.10	1.98	3.69	1.69
17	4.88	3.07	4.17	2.25	49	4.98	2.99	4.15	2.06
18	4.33	2.42	4.58	2.60	50	4.17	2.21	4.15	2.43
19	3.90	1.80	4.80	2.61	51	5.26	3.80	4.92	2.32
20	4.19	2.04	4.37	2.35	52	4.13	2.25	3.74	1.92
21	4.41	2.43	4.20	2.11	53	4.65	3.07	4.57	2.55
22	4.10	2.06	4.35	1.73	54	5.68	3.92	4.79	2.38
23	3.84	1.80	4.92	2.55	55	4.60	2.44	4.61	2.42
24	3.82	1.83	3.88	1.77	56	4.18	2.10	4.88	2.78
25	4.61	3.18	3.96	2.41	57	4.55	3.18	4.32	1.87
26	4.44	2.60	4.24	2.03	58	4.40	2.74	4.39	2.15
27	4.26	2.44	3.99	2.31	59	4.81	2.82	4.76	2.42
28	5.11	3.10	4.45	2.55	60	5.11	3.26	4.68	2.40
29	4.38	2.74	4.33	2.44	61	4.29	2.48	4.38	2.57
30	4.18	2.65	4.97	2.40	62	4.27	2.35	4.07	2.16
31	4.41	2.42	3.56	1.70	63	4.84	2.73	4.21	2.29
32	5.52	3.14	4.83	2.67	64	4.17	2.30	4.07	1.81

As part of the obligatory exercises work for an introductory course on statistical methodology at the Department of Mathematics, University of Oslo, we made our students carry out the following task. Each student was told to stride towards her or his bookshelves, to pick one book in Norwegian and one in English, then record the lengths of the first 100 words on page 51. The books could be novels, collections of short stories, poetry, or prose in general, but not technical material (as with matematics or statistics); and the students were instructed to use page 52 if page 51 didn't have enough words. Do Fløgstad, Kjærstad, Solstad tend to use words with more or less the same lengths as do Miller, Lessing, Munro? What about Askeladden and Winnie Pooh?

The students were asked to summarise information and to compare their own two datasets in terms of means and standard deviations. This was expected to involve tests for equality of means and of variances, confidence intervals for differences, perhaps comments on skewnesses, etc. But the experiment also gave us an interesting combined data set, where we recorded the empirical mean and standard deviation for each dataset, for the two languages, for each student. In other words, we have summary statistics data $(x_i, \widehat{\kappa}_{1,i}, y_i, \widehat{\kappa}_{2,i})$ for $i = 1, \ldots, n$, for the $n = 64$ students, with

$$\begin{aligned} x_i &= \text{average word-length for 100 Norwegian words for student } i, \\ y_i &= \text{average word-length for 100 English words for student } i, \end{aligned} \tag{B.1}$$

along with

$$\begin{aligned} \widehat{\kappa}_{1,i} &= \text{standard deviation for the 100 Norwegian word-lengths for student } i, \\ \widehat{\kappa}_{2,i} &= \text{standard deviation for the 100 English word-lengths for student } i, \end{aligned} \tag{B.2}$$

Figure [xx nemlig xx] displays the word-length averages (x_i, y_i), with grand averages $\bar{x} = 4.549$ for Norwegian and $\bar{y} = 4.436$ for English words (so these are both averages over 6400 words).

There are at least four notable aspects of these data, each rather non-trivial when it comes to their precise assessments and analyses. We point to and briefly explain these aspects or features now, and come back to precise inference methods and results in later sections.

The first and second aspects we wish to call attention to and analyse is that the overall means are reasonably similar, for Norwegian and English, but that the overall variability measures, related to the sizes of the $(\widehat{\kappa}_{1,i}, \widehat{\kappa}_{2,i})$ of (B.2), are significantly different.

The third feature is that there is positive correlation between word-lengths; if x_i is small, for student i, then y_i also tends to be small, etc. Indeed we find $\rho = \text{corr}(x, y)$ estimated in the usual fashion as 0.283. This is a deflated correlation, however, in that there is an undiluted correlation, say $\rho_0 = \text{corr}(x_0, y_0)$, in which we would be more interested, where $(x_{0,i}, y_{0,i})$ are the real word-length averages for student i, across all sentences in all books on her or his bookshelves. This ρ_0, if we may estimate it, despite the (x_0, y_0) not being visible, gives a better measure of the extent to which long-worded books in Norwegian tend to be coupled with long-worded books in English (and similarly, for short-worded texts), for a given student (and her or his bookshelves). To arrive at estimates and inference for the ρ_0 we need to accept that the directly observed (x, y) are proxies for (x_0, y_0), not the real thing, and we need to model them. We shall learn that the real ρ_0 is notably bigger than ρ.

Finally the fourth aspect worth serious examination is the level of differences in average word-lengths, across students. There is sufficient variation from student to student, as we shall see, that neither the (x_i, y_i) of (B.1) nor the underlying and not visible $(x_{i,0}, y_{i,0})$ alluded to above can be seen as samples from the same homogeneous normal distribution. The statistical task becomes how to model and assess the level of variability, among the $x_{i,0}$ and the $y_{i,0}$.

II (iv) Platon: from Republic to Laws

(xx to come. $2^5 = 32$ different clausulae, the five last syllables in Platon's sentences, corpus for corpus. here '0' and '1' indicate 'short' and 'long' (or 'light' and 'stressed'),

for these last five syllables. data from Cox and Brandwood (1959). this is about ordering the last five works, Crit, Phil, Pol, Soph, Tim, on the timeline from corpus A, Republic ($n_A = 3778$ sentences), to corpus B, Laws ($n_B = 3783$ sentences). Sample sizes for Crit, Phil, Pol, Soph, Tim are smaller, 150, 958, 770, 919, 762 sentences. Story ii.8 in Ch 4. xx)

```
0 0 0 0 0  1.1  2.4  3.3  2.5  1.7  2.8  2.4

1 0 0 0 0  1.6  3.8  2.0  2.8  2.5  3.6  3.9
0 1 0 0 0  1.7  1.9  2.0  2.1  3.1  3.4  6.0
0 0 1 0 0  1.9  2.6  1.3  2.6  2.6  2.6  1.8
0 0 0 1 0  2.1  3.0  6.7  4.0  3.3  2.4  3.4
0 0 0 0 1  2.0  3.8  4.0  4.8  2.9  2.5  3.5

1 1 0 0 0  2.1  2.7  3.3  4.3  3.3  3.3  3.4
1 0 1 0 0  2.2  1.8  2.0  1.5  2.3  4.0  3.4
1 0 0 1 0  2.8  0.6  1.3  0.7  0.4  2.1  1.7
1 0 0 0 1  4.6  8.8  6.0  6.5  4.0  2.3  3.3
0 1 1 0 0  3.3  3.4  2.7  6.7  5.3  3.3  3.4
0 1 0 1 0  2.6  1.0  2.7  0.6  0.9  1.6  2.2
0 1 0 0 1  4.6  1.1  2.0  0.7  1.0  3.0  2.7
0 0 1 1 0  2.6  1.5  2.7  3.1  3.1  3.0  3.0
0 0 1 0 1  4.4  3.0  3.3  1.9  3.0  3.0  2.2
0 0 0 1 1  2.5  5.7  6.7  5.4  4.4  5.1  3.9

1 1 1 0 0  2.9  4.2  2.7  5.5  6.9  5.2  3.0
1 1 0 1 0  3.0  1.4  2.0  0.7  2.7  2.6  3.3
1 1 0 0 1  3.4  1.0  0.7  0.4  0.7  2.3  3.3
1 0 1 1 0  2.0  2.3  2.0  1.2  3.4  3.7  3.3
1 0 1 0 1  6.4  2.4  1.3  2.8  1.8  2.1  3.0
1 0 0 1 1  4.2  0.6  4.7  0.7  0.8  3.0  2.8
0 0 1 1 1  2.8  2.9  1.3  2.6  4.6  3.4  3.0
0 1 0 1 1  4.2  1.2  2.7  1.3  1.0  1.3  3.3
0 1 1 0 1  4.8  8.2  5.3  5.3  4.5  4.6  3.0
0 1 1 1 0  2.4  1.9  3.3  3.3  2.5  2.5  2.2

0 1 1 1 1  3.5  4.1  2.0  3.3  3.8  2.9  2.4
1 0 1 1 1  4.0  3.7  4.7  3.3  4.9  3.5  3.0
1 1 0 1 1  4.1  2.1  6.0  2.3  2.1  4.1  6.4
1 1 1 0 1  4.1  8.8  2.0  9.0  6.8  4.7  3.8
1 1 1 1 0  2.0  3.0  3.3  2.9  2.9  2.6  2.2

1 1 1 1 1  4.2  5.2  4.0  4.9  7.3  3.4  1.8
```

II (v) Tirant lo Blanc: the world's first novel

(xx the 425 chapters. sentence lengths as well as relative frequencies of words of lengths 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. too much to give all data here, btu we give a figure and point to a data file. xx) [xx describe data. used in Cunen et al. (2018). word length frequencies, $(\widehat{p}_1, \ldots, \widehat{p}_{10})$, for 420 chapters. when did Author B take over for Author A? xx]

II (vi) Sons of Odin

Odin had six sons (though sources are not entirely clear on the matter): Thor, Balder, Vitharr, Váli (cf. the Eddic poems and the Snorri Edda), Heimdallr, Bragi (cf. Snorri's

kennings). In Story [xx nemlig xx] we construct a confidence distribution for the number of all of Odin's children. In some of the Snorri kennings there are also references to Týr and Höd as sons of Odin (and yet other names are mentioned in the somewhat apocryphical Skáldskaparmál). [xx pointer to brief stories about this, one CD, one Bayes, which prior fits the CD answer. xx]

II (vii) Markov and Pushkin

(xx describe data, versions of which used also in Hjort and Varin (2008); Schweder and Hjort (2016). vowels and consonants in Pushkin's classic 1833 epic poem Евгений Онегин. point to Markov (1906, 1913). analysed in Story ii.6. So *Heedless of the proud world's enjoyment* becomes

```
0 1 1 0 0 1 0 0  1 0  0 0 1  0 0 1 1 0  0 1 0 0 0  0  1 0 0 1 1 0 1 0 0
```

i t.d. xx)

II (viii) And Quiet Does Not Flow the Don

[xx describe data, Hjort (2007), mention Kjetsaa et al., Kjetsaa et al. (1984); Lessing (1997). Is it Sholokhov, or is it Kruikov? Grandest scandal in all of Nobel Prize history? analysed in Story ii.7. data are sentence lengths, sorted into window 1-5, 6-10, 11-15, and so on, for three text corpora: Sh (known to be Sholokhov), Kr (known to be Kriukov), TD (the apple of discord, Tikhij Don).

from	to	Sh	Kr	TD
1	5	799	714	684
6	10	1408	1046	1212
11	15	875	787	826
16	20	492	528	480
21	25	285	317	244
26	30	144	165	121
31	35	78	78	75
36	40	37	44	48
41	45	32	28	31
46	50	13	11	16
51	55	8	8	12
56	60	8	5	3
61	65	4	5	8

II (ix) Bach and the others

A fugue, whether for a piano, an organ, or an ensemble of instruments, starts with the principal fugue theme itself, before it is imitated and varied, perhaps in complex ways, in other voices; typical Bach fugues have from three to five voices. Rydén (2020) has studied fugue subjects in organ compositions of Bach and other composers, and has defined certain features, say $x_1, \ldots, x_6$, which can be worked out for each given organ fugue. In brief, x_1 is the length; x_2 the range; x_3 the number of unique notes; x_4 the start interval; x_5 the number of unique intervals betweem successive notes; and x_6 the max interval. These are described in a bit more detail in Story ii.12, where we investigate differences between the fugues of J.S. Bach (1685–1750) and Max Reger (1873–1916), using data from J. Rydén (personal communciation).

One of the present authors has also played through all 24 preludes and fugues from Bach's Das Wohltemperierte Klavier, part I (from the Anhalt-Köthen period, c. 1722), carefully recording $x_1, \ldots, x_6$ for each fugue, and similarly for all the 24 preludes and fugues from part II (from Bach's rather different Leipzig years, c. 1742). As part of Story ii.12 we investigate whether there are any notable differences between these Bach collections, or between the Clavier fugues and those for organ. (xx decide later whether we go for two consecutive stories. xx)

WTK I, c. 1722							WTK II, c. 1743						
		x1	x2	x3	x4	x5	x6			x1	x2	x3	x4	x5	x6
1	C	14	9	6	2	4	7	1	C	29	9	6	2	6	9
2	c	20	11	7	1	7	9	2	c	9	7	5	3	5	7
3	Ciss	20	15	11	2	10	12	3	Ciss	12	7	5	4	5	7
4	ciss	5	5	4	1	3	5	4	ciss	22	12	8	1	3	7
5	D	13	9	6	2	4	9	5	D	9	10	6	0	6	7
6	d	20	9	7	2	4	9	6	d	24	12	9	2	3	5
7	Diss	24	14	10	3	7	10	7	Diss	20	9	6	7	6	7
8	diss	13	8	6	7	5	7	8	diss	13	8	6	0	3	5
9	E	21	12	7	2	4	5	9	E	6	5	4	2	3	3
10	e	26	13	12	3	7	9	10	e	42	13	9	2	8	10
11	F	21	10	7	2	3	8	11	F	21	13	8	1	6	7
12	f	11	12	9	1	5	7	12	f	28	11	9	7	5	8
13	Fiss	16	10	7	5	4	5	13	Fiss	24	14	8	1	7	12
14	fiss	18	7	7	2	3	3	14	fiss	15	12	7	4	7	8
15	G	31	14	8	2	5	10	15	G	42	17	11	3	7	7
16	g	12	8	7	1	4	8	16	g	24	8	7	4	5	5
17	Giss	7	9	5	7	5	9	17	Giss	24	14	7	3	7	12
18	giss	12	12	9	1	5	9	18	giss	27	13	8	2	7	12
19	A	15	13	10	5	6	8	19	A	19	9	6	2	5	5
20	a	31	13	8	1	6	9	20	a	12	13	10	4	7	11
21	B	38	15	8	2	8	9	21	B	24	14	9	2	5	5
22	b	10	13	6	6	4	13	22	b	27	9	7	2	5	6
23	H	14	10	7	1	3	7	23	H	12	12	7	4	6	9
24	h	21	15	12	3	6	9	24	h	27	13	8	4	5	12

III (i) Power law scaling for academics and support staff

(xx from Jamtveit et al. (2009), Jamtveit et al. (2018). xx)

III (ii) Statistical Sightings of Better Angels

[xx Correlates of War data. several stories. Story I, iii.4, for Ch. 3: about the $w_i = x_i - x_{i-1}$. Story II, iii.5, and also Story III, iii.6, for Chs. 2: estimates and intervals for $\rho_q = \mu_{L,q}/\mu_{R,q}$, ratio of quantiles, using direct large-sample theory. here we take Korea 1950 as change-point, with $n_L = 60$ wars to the left and $n_R = 35$ wars to the right. story III, for Ch. 7: the ρ_q again, but now via II-CC-FF, more precise nonparametric confidence. story IV, finding the change-point, in Ch. 9. Pinker (2011), Cunen et al. (2020a), Gleditsch (2020) volume on the work of Lewis Fry Richardson. we examine the between-war times $w_i = x_i - x_{i-1}$, fit both the simple exponential, as per Poisson process assumption, and a two-parameter model extension. later briefly in Bayes chapter. xx]

(xx here is the full data matrix, for wars $i = 1, \ldots, 95$, with onset time x_i, the between-onset times $w_i = x_i - x_{i-1}$, the battle deaths count z_i, and its logarithm $\log z_i$. from com26 of nilswork21. xx)

i	x	w	z	log z	i	x	w	z	log z
1	1823.269	--	1000	6.908	49	1932.458	0.489	92661	11.437
2	1828.322	5.053	130000	11.775	50	1934.222	1.764	2100	7.650
3	1846.319	17.997	19283	9.867	51	1935.758	1.536	20000	9.903
4	1848.247	1.928	7527	8.926	52	1937.519	1.761	1000000	13.816
5	1848.278	0.031	6000	8.700	53	1938.581	1.061	1726	7.454
6	1849.333	1.056	2600	7.863	54	1939.364	0.783	28000	10.240
7	1851.553	2.219	1300	7.170	55	1939.917	0.553	151798	11.930
8	1853.814	2.261	264200	12.484	56	1940.919	1.003	1400	7.244
9	1856.819	3.006	2000	7.601	57	1941.492	0.572	16634907	16.627
10	1859.331	2.511	22500	10.021	58	1947.822	6.331	3500	8.161
11	1859.811	0.481	10000	9.210	59	1948.375	0.553	8000	8.987
12	1860.697	0.886	1000	6.908	60	1950.825	2.450	910084	13.721
13	1860.792	0.094	1000	6.908	61	1954.675	3.850	2370	7.771
14	1862.294	1.503	20000	9.903	62	1956.836	2.161	3221	8.077
15	1863.894	1.600	1000	6.908	63	1956.844	0.008	2426	7.794
16	1864.086	0.192	4481	8.408	64	1958.111	1.267	1122	7.023
17	1865.181	1.094	310000	12.644	65	1958.647	0.536	1800	7.496
18	1865.736	0.556	1000	6.908	66	1962.806	4.158	1853	7.525
19	1866.472	0.736	44100	10.694	67	1965.103	2.297	1021442	13.837
20	1870.553	4.081	204313	12.227	68	1965.597	0.494	7061	8.862
21	1876.242	5.689	4000	8.294	69	1967.431	1.833	19600	9.883
22	1877.317	1.075	285000	12.560	70	1968.036	0.606	13866	9.537
23	1879.122	1.806	13868	9.537	71	1969.183	1.147	5368	8.588
24	1882.531	3.408	10079	9.218	72	1969.539	0.356	1900	7.550
25	1884.458	1.928	12100	9.401	73	1970.231	0.692	6525	8.783
26	1885.244	0.786	1000	6.908	74	1971.925	1.694	11223	9.326
27	1894.569	9.325	15000	9.616	75	1973.767	1.842	14439	9.578
28	1897.125	2.556	2000	7.601	76	1974.556	0.789	1500	7.313
29	1898.311	1.186	3685	8.212	77	1975.814	1.258	2700	7.901
30	1900.464	2.153	3003	8.007	78	1977.564	1.750	10500	9.259
31	1900.547	0.083	4000	8.294	79	1977.733	0.169	8000	8.987
32	1904.106	3.558	151831	11.931	80	1978.828	1.094	3000	8.006
33	1906.408	2.303	1000	6.908	81	1979.131	0.303	21000	9.952
34	1907.136	0.728	1000	6.908	82	1980.728	1.597	1250000	14.039
35	1909.519	2.383	10000	9.210	83	1982.236	1.508	1001	6.909
36	1911.747	2.228	20000	9.903	84	1982.308	0.072	1655	7.412
37	1912.797	1.050	82000	11.314	85	1986.875	4.567	8000	8.987
38	1913.542	0.744	60500	11.010	86	1987.014	0.139	4000	8.294
39	1914.581	1.039	8578031	15.965	87	1990.589	3.575	41466	10.633
40	1918.894	4.314	11750	9.372	88	1992.269	1.681	5240	8.564
41	1918.936	0.042	13246	9.491	89	1993.100	0.831	14000	9.547
42	1919.122	0.186	100000	11.513	90	1995.025	1.925	1500	7.313
43	1919.294	0.172	11000	9.306	91	1998.350	3.325	120000	11.695
44	1919.347	0.053	50000	10.820	92	1999.233	0.883	5002	8.518
45	1919.836	0.489	40000	10.597	93	1999.356	0.122	1172	7.066
46	1920.542	0.706	1000	6.908	94	2001.875	2.519	4002	8.295
47	1929.631	9.089	3200	8.071	95	2003.219	1.344	7173	8.878
48	1931.969	2.339	60000	11.002					

III (v) The assortative mating according to temper

In his classic *Natural Inheritance*, Galton (1889) gives a fascinating and entertaining analysis of 'temper', coupled with the general theme of inheritance, and courageously classifies husbands and wives as 'bad-tempered' or 'good-tempered'. We use these data in Story iii.8 to not merely test for independence of these character traits, but to provide confidence distributions and curves for relevant parameters. Are bad-tempered men better at finding good-tempered women than the good-tempered men are?

```
              wife:
              good  bad
husband: good  24    27
         bad   34    26
```

Galton, 1887: "We can hardly, too, help speculating uneasily upon the terms that our own relatives would select as most appropriate to our particular selves."

III (vi) Sims and Sveriges Riksbank

(xx perhaps, the Tore story on Sims. xx)

III (vii) Does winning make you live longer?

(xx emil story with RDD. xx)

III (viii) Minimum wages and employment

(xx emil story. xx)

III (ix) How many were killed in Guatemala?

(xx Lum et al. (2013) and Urdal. xx)

III (x) Volatility and estimation from noisy data

(xx emil story. xx)

IV (i) New Haven annual temperatures 1912–1971

(xx annual temperatures at New Haven, Connecticut, for the years 1912 to 1971, as follows. xx)

```
 9.94 11.28  9.67 10.61  9.67  8.83  9.89 10.50  9.61 11.06 10.44  9.78
 9.61 10.33  9.11 10.39 10.50 10.33 10.83 11.56 11.00 10.61  9.89 10.11
10.22 10.89 11.00 10.50  9.33 10.94 10.56 10.33 10.94 10.83 11.17 10.72
10.56 12.22 10.78 11.50 11.72 12.56 11.11 11.11 10.50 11.44 10.11 11.44
10.89 11.06 10.28 10.50 10.94 10.78 10.94 10.44 11.06 11.00 11.06 11.67
```

IV (ii) Skiing days at Bjørnholt 1897–2012

The table gives the number of skiing days at the location Bjørnholt in Nordmarka, a tramride and a skiing hour away downtown Oslo, from 1896 to 2022. A skiing day is defined as there being at least 25 cm snow on the ground. Note the 'hole' in the time series, with no counting of skiing days for the time window 1938 to 1954. Data from `rimfrost.no`. See Heger (2011), Cunen et al. (2018).

```
1896   25 1916  155 1936  151 1972  120 1992   39  2012   95
1897  147 1917  139 1937  108 1973  114 1993   70  2013  108
1898  169 1918  115 1954   37 1974  146 1994  119  2014   81
1899  135 1919  175 1955  135 1975  129 1995  123  2015   97
1900  135 1920  141 1956  111 1976  150 1996  116  2016   84
1901  137 1921  110 1957   82 1977  165 1997   69  2017   55
1902  131 1922  129 1958  125 1978  152 1998   39  2018  115
1903  147 1923  164 1959  143 1979  182 1999  121  2019  125
1904  185 1924  135 1960  141 1980  156 2000   57  2020   34
1905  127 1925   91 1961  141 1981  157 2001  150  2021  110
```

1906	116	1926	130	1962	152	1982	148	2002	129	2022	96
1907	165	1927	163	1963	150	1983	128	2003	113		
1908	170	1928	127	1964	92	1984	121	2004	108		
1909	176	1929	75	1965	150	1985	138	2005	37		
1910	182	1930	122	1966	162	1986	139	2006	120		
1911	149	1931	140	1967	142	1987	127	2007	58		
1912	174	1932	65	1968	125	1988	141	2008	131		
1913	125	1933	55	1969	153	1989	5	2009	101		
1914	128	1934	110	1970	191	1990	27	2010	144		
1915	190	1935	117	1971	178	1991	97	2011	117		

IV (iii) British coal-mining disasters, 1851 to 1962

[xx for chapter1: when data length is stretched, at which time point is it too much for the homogeneous Poisson assumption? in later chapters too. story us treated in Cunen et al. (2018), but then with a change-point perspective. xx]

1851	4	1871	5	1891	2	1911	0	1931	3	1951	1
1852	5	1872	3	1892	1	1912	1	1932	3	1952	0
1853	4	1873	1	1893	1	1913	1	1933	1	1953	0
1854	1	1874	4	1894	1	1914	1	1934	1	1954	0
1855	0	1875	4	1895	1	1915	0	1935	2	1955	0
1856	4	1876	1	1896	3	1916	1	1936	1	1956	0
1857	3	1877	5	1897	0	1917	0	1937	1	1957	1
1858	4	1878	5	1898	0	1918	1	1938	1	1958	0
1859	0	1879	3	1899	1	1919	0	1939	1	1959	0
1860	6	1880	4	1900	0	1920	0	1940	2	1960	1
1861	3	1881	2	1901	1	1921	0	1941	4	1961	0
1862	3	1882	5	1902	1	1922	2	1942	2	1962	1
1863	4	1883	2	1903	0	1923	1	1943	0		
1864	0	1884	2	1904	0	1924	0	1944	0		
1865	2	1885	3	1905	3	1925	0	1945	0		
1866	6	1886	4	1906	1	1926	0	1946	1		
1867	3	1887	2	1907	0	1927	1	1947	4		
1868	3	1888	1	1908	3	1928	1	1948	0		
1869	5	1889	3	1909	2	1929	0	1949	0		
1870	4	1890	2	1910	2	1930	2	1950	0		

IV (iv) Mammals and their bodies and brains

[xx 56 mammals, table gives average weight of body and average weight of brain, in kg. *you* are no. 25. how special are you? (log-body, log-brain) follow a binormal, more or less. for Story iv.4. xx]

1	0.480	0.0155	20	10.000	0.1150	39	60.000	0.0810
2	0.019	0.0003	21	3.300	0.0256	40	3.600	0.0210
3	600.000	0.4230	22	0.200	0.0050	41	0.320	0.0019
4	14.000	0.0700	23	85.000	0.3250	42	0.743	0.0200
5	14.800	0.0982	24	2.625	0.0123	43	0.075	0.0012
6	33.500	0.1150	25	62.000	1.3200	44	0.148	0.0012
7	0.728	0.0055	26	6654.000	5.7120	45	0.122	0.0030
8	0.420	0.0064	27	6.800	0.1790	46	0.920	0.0057
9	0.060	0.0010	28	0.120	0.0010	47	0.101	0.0040
10	1.000	0.0066	29	0.022	0.0004	48	0.048	0.0003
11	0.005	0.0001	30	0.010	0.0002	49	86.250	0.1800
12	3.500	0.0108	31	1.400	0.0125	50	4.500	0.0250

13	2.950	0.0123	32	2.500	0.0121	51	207.501	0.1690
14	1.700	0.0063	33	55.500	0.1750	52	0.900	0.0026
15	2547.000	4.6030	34	52.200	0.4400	53	0.104	0.0025
16	0.023	0.0003	35	100.000	0.1570	54	2.000	0.0175
17	521.000	0.6550	36	25.235	0.1800	55	3.380	0.0445
18	187.000	0.4190	37	0.550	0.0024	56	4.230	0.0504
19	0.770	0.0035	38	1.620	0.0114			

IV (v) Kola temperatures and The Hjort liver index time series 1859–2020

[xx The Hjort liver index. comes in two forms, the hli_{bulk} and $\text{hli}_{\text{perfish}}$. from Hermansen et al. (2016). one story on the five-parameter model with gamma marginals, Story iv.5, with ML. xx]

IV (vi) How many Clethrionomys glareoli?

(xx fill in basic reference. xx)

Bolt from heaven

Bolt (2013) is an interesting book, but the author restrains most of his attention to himself and his own achievements, so to find relevant data for our measure of surprise analyses for track and field events in Story v.1 we have needed to track down and organise our own files. Our analyses in particular utilise all sub-10.00 100 metre races from 2000 to 2008, and on the technical side involve Bartletting the deviance function derived from the extreme value distribution.

(xx Bolt did 9.72 in June 2008. how surprising ought we to have been? data: during the eight seasons 2000 to 2007, there were 20, 27, 29, 14, 38, 21, 29, 17 races clocked at 10.00-or-better, a total of 195 such. Story v.1. xx)

9.86	9.87	9.91	9.93	9.94	9.95	9.96	9.97	9.97	9.97	9.98	9.98
9.98	9.99	9.99	10.00	10.00	10.00	10.00	10.00	9.82	9.84	9.85	9.88
9.88	9.90	9.90	9.90	9.91	9.92	9.94	9.95	9.96	9.96	9.96	9.96
9.97	9.97	9.97	9.98	9.98	9.98	9.99	9.99	9.99	10.00	10.00	9.78
9.87	9.89	9.91	9.93	9.94	9.94	9.94	9.94	9.94	9.95	9.95	9.96
9.97	9.97	9.97	9.98	9.98	9.98	9.98	9.98	9.98	9.99	9.99	9.99
9.99	9.99	10.00	10.00	9.93	9.94	9.95	9.97	9.97	9.97	9.97	9.98
9.99	9.99	10.00	10.00	10.00	10.00	9.85	9.86	9.87	9.87	9.88	9.89
9.90	9.90	9.91	9.91	9.91	9.92	9.93	9.93	9.93	9.93	9.93	9.93
9.94	9.94	9.95	9.95	9.96	9.96	9.97	9.97	9.97	9.97	9.98	9.98
9.99	9.99	9.99	9.99	10.00	10.00	10.00	10.00	9.77	9.84	9.85	9.88
9.89	9.94	9.96	9.96	9.98	9.99	9.99	9.99	9.99	9.99	9.99	9.99
10.00	10.00	10.00	10.00	10.00	9.77	9.84	9.85	9.85	9.85	9.86	9.86
9.86	9.87	9.88	9.88	9.88	9.89	9.91	9.91	9.92	9.92	9.95	9.95
9.96	9.96	9.97	9.97	9.98	9.99	9.99	9.99	9.99	10.00	9.74	9.78
9.83	9.84	9.85	9.90	9.91	9.93	9.94	9.95	9.96	9.97	9.97	9.98
9.98	9.99	10.00									

Height, weight, BMI for Olympic speedskaters

(xx speedskaters taking part in Olympics, 1274 men, from 1952 to 2018, 907 women, from 1960 to 2018. we have height, weight, and hence BMI data for these, from files gathered by NLH and fellow speedskating history enthusiasts Jeroen Heijmans and Arild Gjerde. see Story v.11, with overall confidence curves for the 0.2, 0.5, 0.8 quantiles of the

BMI distributions; there is also a more complex story for the II-CC-FF story of medians, partly from Cunen and Hjort (2022). xx)

(xx the 1274 + 907 BMI scoes are available at the kioskvelter website. mean and standard deviation are 23.706, 1.445 for the males, and 22.050, 1.412 for the females. The data ranges are from 18.365 to 28.965 for the men and 17.647 to 26.675 for the women. Figure v.17 shows how clearly the bulk of women's BMI distribution is situated to the left of the men's. xx)

Inner and outer

[xx describe data, inner and outer starts for World Championships 1000 m speedskating. manually collected and curated by NLH over the years, via ISU protocols. analysed in Story v.6. xx]

From semifinals to finals

(xx describe data. type A A B B B A, skiers coming from Semifinal A or Semifinal B, their ranks in the final, for lots of events, Olympics 2022, 2018, 2014, 2010, World Championships 2021, 2019, 2017, 2015, 2013, 2011, and various FIS World Cup events. analysed in Story v.7. xx)

Golf putting

(xx fill in. Gelman and Nolan (2002), Schweder and Hjort (2016, Ch. 14). xx)

Who wins: Real Time Real Excitement Plots

(xx fill in. data I, that particular Nor-Den match; data II, some 75 matches and their scores, for correlated Poissons. xx)

Abelian envelopes

How many Abel commemorative envelopes were issued, in 1929? Prior to 2011, five such envelopes were known in the international community of stamp collectors, with so-called R numbers 280, 304, 308, 310, 328. We derive in Story ii.5 a confidence distribution for the number N of envelopes originally issued. (xx on the Bayesian version in Ch. 6. xx) Story ii.5 concerns reaching Bayesian inference for this unknown number of envelopes, based in these five R numbers. Story ii.5. are also about how these inerence summaries, the posterior for the Bayesian and the CD for the frequentist, can be updated, in the light of three more 1902 envelopes discovered in 2012, carrying R numbers 314, 334, 389. The data, along with the facsimile shown in Figure ii.9, are from colleagues Y. Reichelt and N.V. Johansen of the authors. (xx nils moves the facisimile to the story itself. xx)

Mixed effects for doughnuts

The table below provides data for eight different fats used for six consecutive working days in connection with doughnut mixing; specifically, $y_{i,j}$ given there displays the grams of fat absorbed for fat $i = 1, \ldots, r = 8$ across days $j = 1, \ldots, s = 6$. The data are from Scheffé (1959, p. 137); cf. also McCloskey (1943). (xx then point to where we have a little story about this, with mixed effects, relative influence of days, etc. xx)

	Mon	Tue	Wed	Thu	Fri	Sat	means
1	164	177	168	156	172	195	172.00
2	172	197	167	161	180	190	177.83

3	177	184	187	169	179	197	182.17
4	178	196	177	181	184	191	184.50
5	163	177	144	165	166	178	165.50
6	163	193	176	172	176	178	176.33
7	150	179	146	141	169	183	161.33
8	164	169	155	149	170	167	162.33
means	166.38	184.00	165.00	161.75	174.50	184.88	172.75

State-wise cigarette consumption and cancers

For each of 44 US states (actually, 43 states and the District of Columbia), the table below, dating from 1960, gives `cig`, number of cigarettes smoked (hundreds per capita); `blad`, deaths per 100k population from bladder cancer; `lung`, deaths per 100k population from lung cancer; `kid`, deaths per 100k population from kidney cancer; `leuk`, deaths per 100k population from leukemia. (xx used in Ch. 6. can also be used for regression stories in other chapters. the same x acts on four regressions. outliers? what happens with the verbatim, first line? xx)

	cig	blad	lung	kid	leuk
AL	18.20	2.90	17.05	1.59	6.15
AZ	25.82	3.52	19.80	2.75	6.61
AR	18.24	2.99	15.98	2.02	6.94
CA	28.60	4.46	22.07	2.66	7.06
CT	31.10	5.11	22.83	3.35	7.20
DE	33.60	4.78	24.55	3.36	6.45
DC	40.46	5.60	27.27	3.13	7.08
FL	28.27	4.46	23.57	2.41	6.07
ID	20.10	3.08	13.58	2.46	6.62
IL	27.91	4.75	22.80	2.95	7.27
IN	26.18	4.09	20.30	2.81	7.00
IO	22.12	4.23	16.59	2.90	7.69
KS	21.84	2.91	16.84	2.88	7.42
KY	23.44	2.86	17.71	2.13	6.41
LA	21.58	4.65	25.45	2.30	6.71
ME	28.92	4.79	20.94	3.22	6.24
MD	25.91	5.21	26.48	2.85	6.81
MA	26.92	4.69	22.04	3.03	6.89
MI	24.96	5.27	22.72	2.97	6.91
MN	22.06	3.72	14.20	3.54	8.28
MS	16.08	3.06	15.60	1.77	6.08
MO	27.56	4.04	20.98	2.55	6.82
MT	23.75	3.95	19.50	3.43	6.90
NB	23.32	3.72	16.70	2.92	7.80
NE	42.40	6.54	23.03	2.85	6.67
NJ	28.64	5.98	25.95	3.12	7.12
NM	21.16	2.90	14.59	2.52	5.95
NY	29.14	5.30	25.02	3.10	7.23
ND	19.96	2.89	12.12	3.62	6.99
OH	26.38	4.47	21.89	2.95	7.38
OK	23.44	2.93	19.45	2.45	7.46
PE	23.78	4.89	12.11	2.75	6.83
RI	29.18	4.99	23.68	2.84	6.35
SC	18.06	3.25	17.45	2.05	5.82
SD	20.94	3.64	14.11	3.11	8.15
TE	20.08	2.94	17.60	2.18	6.59
TX	22.57	3.21	20.74	2.69	7.02
UT	14.00	3.31	12.01	2.20	6.71
VT	25.89	4.63	21.22	3.17	6.56

```
WA 21.17 4.04 20.34 2.78 7.48
WI 21.25 5.14 20.55 2.34 6.73
WV 22.86 4.78 15.53 3.28 7.38
WY 28.04 3.20 15.92 2.66 5.78
AK 30.34 3.46 25.88 4.32 4.90
```

Mothers and babies

(xx describe the data, from Hosmer and Lemeshow (1999), 189 newborns, their weights, along with information about their mothers, including weight before pregnancy, age, an indicator for smoking, and also information about three ethnic groups. used for nonparametric confidence intervals for quantiles, and nils plans to have a fic. also something for comparing distribution of weight at birth for smoking and non-smoking mothers. xx)

Longevity returns to political office

[xx describe the data set of Barfort et al. (2020), that is analysed in Story iii.11. xx]

Onset of menarche

(xx clean and polish text, with pointer to exercise. xx) Data pertain to 3918 Warszawa girls and onset of menarche. The data have three columns, giving respectively (i) age x (or rather the midpoint in the appropriate age interval); (ii) the number m of girls in this age window; (iii) the number y among the of girls in this age window who have had their first menstruation. Thus y_j in age window j is seen as a realisation of a binom(m_j, p_j) distribution, with $p_j = p(x_j)$ the probability of onset having taken place at age x_j or earlier.

x	m	y	x	m	y	x	m	y
9.21	376	0	12.33	93	29	14.58	120	113
10.21	200	0	12.58	100	39	14.83	102	95
10.58	93	0	12.83	108	51	15.08	122	117
10.83	120	2	13.08	99	47	15.33	111	107
11.08	90	2	13.33	106	67	15.58	94	92
11.33	88	5	13.58	105	81	15.83	114	112
11.58	105	10	13.83	117	88	17.58	1049	1049
11.83	111	17	14.08	98	79			
12.08	100	16	14.33	97	90			

Decimals of π

For our Story vii.2 we have first gathered the first 6,537,216 digits of π and then via some coding found the first 223,157 values of V_{10}, the lengths of cycles required to have seen all decimals $0, 1, \ldots, 9$. These are gathered in our file `V10counts`.

Australian rowers

(xx to be edited and clarified. for Story vii.9. data from `www.statsci.org/data/oz/ais.txt`. Telford, R. D. and Cunningham, R. B. (1991) Sex, sport, and body-size dependency of hematology in highly trained athletes. Medicine and Science in Sports and Exercise, 23(7):788-794. gender is 11 for girls and 12 for boys. x is lean body mass, y is percent body fat. do we know or merely guess that no 16 and no 30 are coxswains? xx)

	gender	x	y	bmi
1	11	66.24	17.71	25.436
2	11	57.92	18.77	22.630
3	11	56.52	19.83	21.856
4	11	54.78	25.16	22.270
5	11	56.31	18.04	21.275
6	11	62.96	21.79	23.470
7	11	56.68	22.25	23.190
8	11	62.39	16.25	23.174
9	11	63.05	16.38	24.536
10	11	56.05	19.35	22.955
11	11	53.65	19.20	19.763
12	11	65.45	17.89	23.363
13	11	64.62	12.20	22.666
14	11	60.05	23.70	24.236
15	11	56.48	24.69	24.212
16	11	41.54	16.58	20.464
17	11	52.78	21.47	20.810
18	11	52.72	20.12	20.168
19	11	61.29	17.51	23.060
20	11	59.59	23.70	24.402
21	11	61.70	22.39	23.974
22	11	62.46	20.43	22.617
23	12	78.00	9.00	23.566
24	12	75.00	12.61	25.839
25	12	78.00	9.03	24.057
26	12	87.00	6.96	23.850
27	12	78.00	10.05	25.090
28	12	79.00	9.56	23.844
29	12	79.00	9.36	25.314
30	12	48.00	10.81	19.690
31	12	82.00	8.61	26.069
32	12	82.00	9.53	25.503
33	12	82.00	7.42	23.688
34	12	83.00	9.79	26.795
35	12	88.00	8.97	25.615
36	12	83.00	7.49	25.055
37	12	78.00	11.95	24.928

Other Stuff, and references to be used

(xx we mention and point to things here, not yet on board as of 16-January-2025, but probably to be factored in. we also point to references that we sooner or later will have on board. xx)

Aalen (1992), Aalen and Gjessing (2004), Cunen et al. (2020c), Frigessi and Hjort (2002), Gjessing et al. (2003), Hjort (1990b), Hjort (1990a), Hjort and Stoltenberg (2021), ... when we mention Breiman (2001), we also point to the Special Issue of *Observational Studies* (vol. 7, 2021), e.g. Bickel's comments, in `muse.jhu.edu/issue/45147`. also Gelman et al. (2022).

References

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.

Aït-Sahalia, Y. and Jacod, J. (2014). *High-Frequency Financial Econometrics*. Princeton University Press, Princeton.

Aldous, D. J. and Eagleson, G. K. (1978). On mixing and stability of limit theorems. *The Annals of Probability*, pages 325–331.

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., Tvete, 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., Tvete, 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.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*.

Barfort, S., Klemmensen, R., and Larsen, E. G. (2020). Longevity returns to political office. *Political Science Research and Methods*, 9:658–664.

Bartolucci, F. and Lupparelli, M. (2008). Focused Information Criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistis*, 9:658–664.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a densithy power divergence. *Biometrika*, 85:549–559.

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.

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, Kondon.

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.

Borgan, Ø. and Keilman, N. (2019). Do Japanese and Italian women live longer than women in Scandinavia? *European Journal of Population*, 35:87–99.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360.

Breiman, L. (2001). Statistical modeling: The two cultures [with comments and a rejoinder by the author]. *Statistical Science*, 16:199–231.

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.

Candès, E. J., Lei, L., and Ren, Z. (2021). Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*.

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 George, E. I. (1992). Explaining the Gibbs Sampler. *American Statistician*, 46:167–174.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118:e2107794118.

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.

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 worlks 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. (2024). 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.

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.

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.

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.

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.

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 Analaysis of Contingency Tables*. Chapman and Hall/CRC, New YOrk.

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 discontiniuous phenomena. *Journal of Nonparametric Statistics*, 14:1–5.

Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

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. (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.

Ghosh, M. (2002). Basu's theorem with applications: a personalistic review. *Sankhya*, 35:721–741. Special issue in memory of D. Basu.

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.

Glad, I. K., Hjort, N. L., and Ushakov, N. G. (2003). Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, 30:415–427.

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.

Gran, J. M. and Stensrud, M. J. (2022). Hva er forventet levealder? *Tidsskrift for Den norske legeforening*, page 245.

Grønneberg, S. and Hjort, N. L. (2012). On the errors committed by sequences of estimator functionals. *Mathematical Methods of Statistics*, 20:327–346.

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.

Hall, P. G. (1983). Large-sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, 11:1156–1174.

Hanche-Olsen, H. and Holden, H. (2010). The Kolmogorov-Riesz compactness theorem. *Expositiones Mathematicase*, 28:385–394.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition*. Springer, New York.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 9:97–109.

Haug, K. K. (2019). Focused model selection for Markov chain models, with an application to armed conflict data. Technical report, University of Oslo. Master Thesis.

Heger, A. (2011). Jeg og jordkloden. *Dagsavisen*, Dec. 16.

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.

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, N. L. (1986a). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinaavian 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. *Scandinaavian 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 Statisical Review*, xx:355–387.

Hjort, N. L. (1994). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association*, 89:665–675.

Hjort, N. L. (2007). And quiet does not flow the Don: Statistical analysis of a quarrel between Nobel laureates. In Østreng, W., editor, *Concilience*, 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. 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. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23:882–904.

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 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 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.

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 Mémin, J. (1981). Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité. In *Séminaire de Probabilités (Strasbourg), tome 15*, pages 529–546. Springer.

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, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R. Second Edition.* Springer, New York.

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.

Jones, M. C. (1991). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters*, 12:51–56.

Jones, M. C., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, 88:865–873.

Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, 27:951–981.

Jullum, M. and Hjort, N. L. (2019). What price semiparametric Cox regression? *Lifetime Data Analysis*, 25:406–438.

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.

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. *Ophtalmology*, 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.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer, New York.

Kusolitsch, N. (2010). Why the theorem of Scheffé should be rather called a theorem of Riesz. *Periodica Mathematica Hungarica*, 61:225–229.

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.

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. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.

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.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15:211–225.

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.

Loader, C. (1996). Local likelihood density estimation. *Annals of Statistics*, 67:1602–1618.

Lum, K., Price, M. E., and Banks, D. (2013). Applications of multiple systems estimation in human rights research. *American Statistician*, 24:191–200.

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.

Marron, S. and Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics*, 20:712–736.

McCloskey, R. (1943). *Homer Price*. Scholastic Inc., New York.

McCullagh, P. (2002). What is a statistical model? [with discussion]. *Annals of Statistics*, 30:1225–1310.

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.

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.

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.

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.

Price, R. M. and Bonett, D. G. (2001). Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation*, 68:xx–xx.

Price, R. M. and Bonett, D. G. (2002). Distribution-free confidence intervals for difference and ratio of medians. *Journal of Statistical Computation and Simulation*, 72:xx–xx.

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.

Romano, J. P. and Siegel, A. F. (1986). *Counterexamples in Probability and Statistics*. Wadsworth & Brooks/Cole, Belmont.

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, Beijin.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimation. *Scandinavian Journal of Statistics*, 9:65–78.

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.

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.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

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. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, London.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9.

Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, 1:185–202.

Shiryaev, A. N. (1996). *Probability. Second edition*. Springer, Berlin.

Shumway, R. H. and Stoffer, D. S. (2016). *Time Seires 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.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

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, Puplic 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:xx–xx.

Stoltenberg, E. A. and Hjort, N. L. (2021). Models and inference for on-off data via clipped Ornsten–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.

Swensen, A. R. (1983). A note on convergence of distributions of conditional moments. *Scandinavian Journal of Statistics*, 10:41–44.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

Tversky, A. and Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. *Chance*, 2:16–21.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

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 Obstetricia 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.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media, Berlin/Heidelberg.

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.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

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.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D., Riffe, T., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., C, W., and M, B. (2021). Methods protocol for the Human Mortality Database. University of California, Berkeley, US, and Max Planck Institute for Demographic Research, Rostock, Germany. `https://www.mortality.org/` [Version 6. Last revised January 26, 2021].

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

James, LeBron, 595

Kahnemann, Daniel, 595

Love, Kevin, 595

Subject index

Assignment mechanism, 615

Cold hand hypothesis, 595
Confounder, 614

Hot hand, 595

Instrumental variable, 616

observational study, 615

Probit model, 463

Randomised experiement, 615

Stable unit treatment value assumption, 616