ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Unofficial Errata and Commentary for E. T. Jaynes's *Probability Theory: The Logic of Science*

I consider Edwin T. Jaynes's book *Probability Theory: The Logic of Science* to be one of the most important works on the theory of probability of the last century. Unfortunately, Jaynes fell ill and died before he could complete the book. His incomplete manuscript was available on the web at that time, and some entire chapters were missing, parts of other chapters were missing, and others were not in as finished and polished a state as they probably would have been had Jaynes remained in good health long enough to finish the book. Jaynes's colleague G. Larry Bretthorst accepted the responsibility to put the unfinished manuscript into publishable form, and it was published in May 2003, five years after Jaynes's death. However, as Bretthorst writes in the editor's foreword,

> I could have written these latter chapters and filled in the missing pieces, but if I did so, the work would no longer be Jaynes'; rather, it would be a Jaynes-Bretthorst hybrid with no way to tell which material came from which author. In the end, I decided the missing chapters would have to stay missing--the work would remain Jaynes'.

As a result, there remain omissions and some cases of unclear exposition (that contrast sharply with the clear exposition found in the more finished chapters); furthermore, the author is not in a position to issue his own errata to correct various minor errors that remain in the published form of the book.

The purpose of this web page is then to help the readers of *Probability Theory* to get more out of it, by

- collecting and providing a list of (apparent) errata;
- providing additional information and exposition where it may aid comprehension of material in the book; and
- providing commentary on specific chapters or sections.

I would welcome contributions by others to any of these categories. Such contributions (or corrections to my commentary and errata) should be sent to bayes@ksvanhorn.com; please indicate whether you wish to be identified as the contributor, and if so, if you wish your email address to be given. I will forward the errata to the editor, Larry Bretthorst, so that appropriate corrections can be made in any future editions.

There is a mailing list for discussion about Jaynes's book: the etjaynesstudy group at Yahoo groups. If you have questions about Jaynes's book, the list is a good place to ask them. You can also email me at bayes@ksvanhorn.com, and I'll try to answer your question if I can.

My thanks go to the following people who have contributed errata or comments: Nick Cox, Philip Dawid, Anthony Garrett, Tony Kocurko, Naoki Saito, Eliezer S. Yudowsky, and Arnold Zellner. (There may be others I've lost track of; if I've inadvertently left your name off of the preceding list, please let me know.)

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

What's New

6 November 2004: Added some additional info on Zellner's approach to ignorance priors.

8 June 2004:

- A partial solution to exercise 4.1 in Chapter 4, courtesy of Timothy D. Sanders.
- Additional errata in Chapter 2 (eqn. (2.45) and top of p. 33).
- Additional commentary for Chapter 2, showing how to derive eqn. (2.50) from (2.48).

23 November 2003:

- New erratum on p. 158 of Chapter 6, courtesy of Eliezer Yudowsky.
- Updated publication data and added links to discussion following publication of Constructing a logic of plausible inference.
- Added commentary and new errata to Chapter 18.

29 October 2003: New errata and a new section of commentary for Chapter 15.

11 October 2003:

- New errata for Chapter 2, courtesy of Tony Kocurko.
- New errata for the Preface, Chapter 2, Chapter 3, Chapter 7, Chapter 12, Chapter 14, Chapter 16, Chapter 17, Chapter 18, Chapter 22, Appendix B, Appendix C, and References etc., courtesy of Nick Cox.
- New commentary for Chapter 4, Chapter 6, Chapter 7, Chapter 9, Chapter 10, Chapter 17, and Chapter 20, courtesy of Nick Cox.

20 September 2003:

- Added some additional information on computing the parameters of a maximum-entropy distribution.
- Fixed the Jaynes mailing list link.
- New errata in Chapters 1, 2, and 17, especially the latter.
- Additional commentary for Chapter 17.

30 August 2003:

- Intro page: mailing list for discussion of PT:TLOS.
- Added some commentary on Zellner's proposed ignorance prior for an unknown probability of success.

- Added some references on computing the parameters of a maximum-entropy distribution from the desired expected values.

26 August 2003:

- Added a reference to J. M. Garrett's paper under Commentary: Additional treatments of Cox's Theorem.
- Added several pieces of commentary to Chapter 12: Ignorance priors and transformation groups
- Added new errata for Chapter 12 and Chapter 15.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Commentary: Note on exchangeability and de Finetti's Theorem

In several places Jaynes refers to exchangeability and de Finetti's Theorem without defining these; finally, in Chapter 18 section 16 (p. 586 onward) he says a little bit about just what de Finetti's Theorem is. For those readers who are unfamiliar with this topic, Bernardo and Smith's book *Bayesian Theory* has a nice discussion in sections 4.2 and 4.3. Here is a brief, simplified summary of the definitions and theorems given therein, translated into the vocabulary and notation of Jaynes.

Notation. We write x_1^n for $x_1, \dots, x_n$, and x_1^∞ for the infinite sequence $x_1, x_2, \dots$. For simplicity, we assume that the set of possible values for each variable x_i is the same finite set S.

Definition of finite exchangeability. The variables x_1^n are said to be (finitely) exchangeable for a state of information I if, for any constant values X_1^n, we have

$$P(x_1^n = X_1^n \mid I) = P(y_1^n = X_1^n \mid I)$$

whenever y_1^n is a permutation of the variables x_1^n.

Definition of infinite exchangeability. The infinite sequence of variables x_1^∞ is said to be infinitely exchangeable for a state of information I if every finite subsequence of x_1^∞ is exchangeable for I.

Theorem. Let W be the set of probability mass functions over S. If x_1^∞ is an infinitely exchangeable sequence of variables for I, then there exists a probability density f over W such that

$$P(x_1^n = X_1^n \mid I) = \int_{w \in W} f(w) \prod_i w(X_i)$$

for any constant X_1^n. (Note that f may involve delta functions, to assign positive probability to a single specific value in W.)

In other words, we may reason as if there exists some additional variable u such that the variables x_1^n are independently and identically distributed when the value of u is known, and $P(x_i = s \mid u = w, I)$ is $w(s)$.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Preface

- p. xxiii, line 6: Given the context, ``proscribed'' was probably intended to be ``circumscribed'' (meaning ``limited'').
- p. xxiv, note 3, line 4: ``the only property'' should be ``the only properties.''

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Chapter 1: Plausible reasoning

- p. 10, first line: The reference ``(1812)'' should probably be ``Laplace (1812)''.

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 2: The quantitative rules

- p. 31, line after eqn. (2.44): It appears that the reference to (2.25) should be (2.40).
- p. 31, eqn. (2.45): The constraint $0 \le x \le 1$ should be $0 < x \le 1$, to avoid division by 0.
- p. 32, eqn. (2.49): A right parenthesis is missing in the expression `` $S(1-\exp -q$''.
- p. 33, sentence starting at top of page: to account for the restriction $x > 0$, this sentence should probably be rewritten as ``Using continuity, the only solution of this satisfying $\lim_{x \to 0} S(x) = 1$ is...''
- p. 33, second paragraph: ``Again, Aczel (1966) derives the same result without assuming differentiability.'' (This refers to equation (2.58).) I checked out the Aczél reference in preparing a review paper on Cox's Theorem, and nowhere did I find anything like the result of equation (2.58); I can only assume that the result appears in some other work of Aczél. I did find, however, that a similar result appears in Paris's book, *The Uncertain Reasoner's Companion*.
- p. 34, eqn. (2.65), second term after first "=" sign: $\overline{AB}$ should be $\overline{A}\,\overline{B}$.
- p. 34, Exercise 2.2: In two places (lines 2-3 and equation (2.67)) a right parenthesis is missing:

  $$p(C \mid (A_1 + A_2 + \dots + A_n X)$$

  should be

  $$p(C \mid (A_1 + A_2 + \dots + A_n) X).$$

- p. 40, second paragraph: ``The argument we have just given is the first `baby' version of the group invariance principle for assigning plausibilities; it will be extended greatly in Chapter 6, when we consider the general problem of assigning `noninformative priors.' '' I believe that ``Chapter 6'' should actually be ``Chapter 12'' (``Ignorance priors and transformation groups'').
- p. 42, second full paragraph, line six: ``kelvin'' should be ``Kelvin.''

Commentary

Several people have written asking how to derive (2.50) from (2.48); here is the derivation:

1. Rewrite (2.48) as $y = S(x)/(1 - \exp(-q))$.
2. Note that, for all z, $1/(1-z) = 1 + z + O(z^2)$.
3. Rewrite step 1 as

$$y = S(x)(1 + \exp(-q) + O(\exp(-2q))).$$

4. Apply $S()$ to both sides:

$$S(y) = S[S(x) + S(x)\exp(-q) + S(x)O(\exp(-2q))].$$

5. Now do a Taylor series expansion around $y = S(x)$:

$$\begin{aligned} S(y) &= S(S(x)) + S(x)\exp(-q)S'(S(x)) + S(x)O(\exp(-2q))S'(S(x)) \\ &= S(S(x)) + \exp(-q)S(x)S'(S(x)) + O(\exp(-2q)) \end{aligned}$$

since $S(x)S'(S(x))$ has no dependence on q.

Commentary: Additional treatments of Cox's Theorem

The following references not appearing Jaynes' book provide additional perspective on Cox's Theorem:

- J. B. Paris, *The Uncertain Reasoner's Companion: A Mathematical Perspective*, Cambridge University Press, 1994. Chapter 3 proves a version of Cox's Theorem, with great care taken to explicitly list all of the assumptions required.

- K. S. Van Horn, ``Constructing a logic of plausible inference: a guide to Cox's Theorem,'' *International Journal of Approximate Reasoning* **34**, no. 1 (Sept. 2003), pp. 3-24. Reviews Cox's Theorem, explicitly listing the assumptions required and discussing (1) the intuition and reasoning behind these requirements, and (2) the most important objections to these requirements. (Preprint: Postscript, PDF.)
    - G. Shafer, ``Comments on `Constructing a logic of plausible inference: a guide to Cox's Theorem', by Kevin S. Van Horn,'' *IJAR* **35**, no. 1 (Jan. 2004), pp. 97-105. Comments by Glenn Shafer (of Dempster-Shafer belief function theory). (Preprint: PDF, or try Glenn Shafer c.v. and look under ``Other Contributions.'')
    - K. S. Van Horn, ``Response to Shafer's comments,'' *IJAR* **35**, no. 1 (Jan. 2004), pp. 107-110. Van Horn's rejoinder. (Preprint: Postscript, PDF.)
- J. M. Garrett, ``Whence the laws of probability?'', in G. J. Erickson, J. T. Rychert, and C. R. Smith (eds.), *Maximum Entropy and Bayesian Methods*. Boise, Idaho, USA, 1997, Kluwer Academic Publishers. Another derivation of the laws of probability theory, very similar to the Cox derivation but starting from a single logical operation (NAND) instead of two (AND, NOT). (Online, slightly updated: Postscript, PDF)

Note that, like Cox, Jaynes does not explicitly list all of his precise assumptions in one place (although he gives desiderata that motivate them) -- Paris and Van Horn both address this issue.

Commentary: Consistency of Cox's axioms

Section 2.6.2 discusses the question of whether the rules of probability theory are consistent. Jaynes brings up Godel's result that no mathematical system can provide a proof of its own consistency, and later writes that ``These considerations seem to open up the possibility that, by going into a wider field by invoking principles external to probability theory, one might be able to prove the consistency of our rules. At the moment, this appears to us to be an open question.''

Actually, it is not an open question: the rules of probability theory can easily be proven consistent, and the proof can be found in any undergraduate mathematical text discussing set-theoretic probability theory. As I wrote in the above-mentioned review of Cox's Theorem,

> But how do we know that our requirements [Cox's axioms] are not contradictory? How do we know that there is *any* system of plausible reasoning... that satisfies all of our requirements? The set-theoretical approach to probability theory may be taken as an existence proof that our requirements are not contradictory, by taking states of information to be [set-theoretical] probability distributions, and defining [state of information] A, X to be the probability distribution obtained from X by conditioning on the set

> of values for which A is true. In the terminology of mathematical logic, set-theoretical probability theory then becomes the model theory for our logic, a tool to enable us to construct consistent sets of axioms (plausibility assignments from which we derive other plausibility assignments).

Viewing probability theory as an extension of the propositional calculus, Jaynes's ``wider field'' is just the predicate calculus with the axioms of finite set theory and real numbers added.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Chapter 3: Elementary sampling theory

- p. 57, third full paragraph, second line: ``(median) ± (interquartile distance).'' Strictly speaking, this should be ``median and quartiles,'' as what is meant is the interval from the lower quartile to the upper quartile. The interval from (median) $-$ (interquartile distance) to (median) $+$ (interquartile distance) is twice as wide.

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 6: Elementary parameter estimation

- p. 150, eqn. (6.1): `` $h(r \mid NR,n)$'' should be `` $h(r \mid N,R,n)$''.
- p. 158, eqn. (6.44): the right-hand side of the equality should have a factor of $\binom{N}{n+1}^{-1}$ instead of $\binom{N}{n+1}$.
- p. 181, eqns. (6.117) and (6.118): The symbol θ appearing in these equations is not mentioned anywhere else in the section; should it instead be ϕ?

Miscellaneous commentary

- p. 190, section 6.20: The taxicab problem was reviewed by Leo A. Goodman (Serial number analysis. JASA 47: 622-634, 1952). [Contributed by Nick Cox.]

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 7: The central, Gaussian, or normal distribution

- p. 200, eqn. (7.2): There is a spurious comma on the second line of the equation, just before ``$+$.''

- p. 222, eqn. (7.52): I believe the term ``$\alpha C_1$'' should be ``$i\alpha C_1$.''

- p. 222, eqn. (7.53), third line: I believe `` $(y - n\langle x\rangle x)$'' should be `` $(y - n\langle x\rangle)$.''

- p. 222, editor's exercise 7.5, third line from bottom: ``theory'' should be ``theorem.''

- p. 226, second paragraph: ``This simple calculation can be greatly generalized, as indicated by Exercise 7.5. But we note an important proviso to be investigated in Exercise 7.6.'' This should be exercises 7.6 and 7.7 instead of 7.5 and 7.6.

- p. 239, first line: ``Following up the idea in Section 7.2.5''. That should be ``Section 7.25''.

Miscellaneous commentary

- p. 234, section 7.22, first line: Modern scholars appear to spell Halley's first name as ``Edmond''. [Contributed by Nick Cox.]
- p. 240, fourth paragraph: The new historical study suggested has been provided by William H. Kruskal and Stephen M. Stigler (Normative terminology: `normal' in statistics and elsewhere. In Bruce D. Spencer (ed.) *Statistics and public policy*, Oxford University Press, 85-111, 1997; also revised (no subtitle) in Stephen M. Stigler, *Statistics on the table: The history of statistical concepts and methods*, Harvard University Press, Cambridge, MA, 1999). [Contributed by Nick Cox.]

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Chapter 8: Sufficiency, ancillarity, and all that

- p. 254, eqn. (8.29), fourth line: there is a missing comma -- `` $f(t_i\theta)$'' should be `` $f(t_i,\theta)$''.
- p. 256, eqn. (8.46): `` $=v(z_1,\dots,z_n)$'' probably should be inserted after ``$p(Z\mid I)$'', as otherwise the ``$v(z_j)$'' in eqn. (8.47) has no referent.
- p. 262, eqn. (8.65): `` $p(H_0\mid I)$'', in the first term of the right-hand-side of the equation, should be `` $p(H_0\mid DI)$''.
- p. 263, eqn. (8.71): ``$p(H_0I)$'', in the first term of the expression between the square brackets, should be `` $p(H_0\mid I)$''.
- p. 267, footnote 8: There is no reference ``Jaynes (1985e)'' in either the References or Bibliography sections of the book, although the References section contains a ``Jaynes (1985)''.

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 9: Repetitive experiments: probability and frequency

- p. 282, second line after (9.24): ``$\sum_j^m$'' should be `` $\sum_{j=1}^m$''.

- p. 285, last paragraph: ``How many terms $T(n,m)$ are in the sum (9.39)?'' should probably be ``How many choices of $n_1,\ldots,n_m$ are there that sum to n?'' or ``How large is the set U?''

- p. 286, second half: ``$\log(W/n)$'' (both places) should be ``$\log(W)/n$''.

- p. 289, section 9.8, first paragraph: ``From (9.28)and (9.29) we see...'' These equations don't seem to have anything to do with what follows.

- p. 292, equation (9.78): `` $\partial/(\log(Z)\partial\lambda)$'' should be `` $\partial\log(Z)/\partial\lambda$''.

- p. 297, equation (9.94): The preceding text, ``we express (9.88) in decibel units as in Chapter 4:'', is misleading, as ψ_B is not ψ_∞ for some hypothesis $H \in B_m$. To make sense of what follows in this section, use the equality

$$\psi_B = \psi_\infty + n\sum_k f_k \log(f_k).$$

- p. 305, second paragraph: ``where $\psi_i$ depends only on the data and $H_i$ is non-negative over $C$'' should be ``where ψ_i depends only on the data and H_i**, and** is non-negative over C''.

Miscellaneous commentary

- p. 300, section 9.12: This criticism of the (Pearson) chi-squared statistic is both fair and exaggerated: Jaynes is correct, but (1) its sensitivity to small expected frequencies has been well rehearsed in many texts, including elementary treatments; (2) the alternative given here, the likelihood ratio chi-squared statistic, has long been available (since about 1930). [Contributed by Nick Cox.]
- p. 304, first (partial) paragraph: Jaynes is correct, but dependence on published tables is totally avoidable given modern software, and practice is steadily swinging to citing P-values rather than using conventional levels such as 5% and 1%. [Contributed by Nick Cox.]

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 10: Physics of `random experiments'

Miscellaneous commentary

- The book by Eduardo M.R.A. Engel, *A road to randomness in physical systems*, Springer, Berlin, 1992, describes related work in probability and statistics. [Contributed by Nick Cox.]
- p. 315, section 10.2: Neither Ellis nor Venn was the equal of Laplace mathematically, but both did very well in mathematics at Cambridge, an education which included much applied mathematics. Ellis was 1st Wrangler (top in mathematics) in 1840 and Venn was 6th Wrangler in 1857. [Contributed by Nick Cox.]

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 11: Discrete prior probabilities: the entropy principle

- p. 359, equation (11.63): insert minus sign before ``$\partial^2 \log Z(\ldots)$''.
- p. 360, line 3: ``$f_k(x_i)$'' should be ``$f_k(x_i;\alpha)$''.
- p. 360, equation (11.65): to be consistent, ``$f_k(x_i,\alpha)$'' should be ``$f_k(x_i;\alpha)$''.
- p. 360, equation (11.69): insert minus sign before ``$\partial^2 \log Z(\ldots)$''.
- p. 361, equation (11.72), second line: ``$\sum_{i-1}^{n}$'' should be ``$\sum_{i=1}^{n}$''.
- p. 362, equation (11.81): Should ``$\langle g_j f_k \rangle$'' be ``$\langle g_j g_k \rangle$''?
- p. 367, equation (11.92): ``$d$'' should be ``k''.
- p. 368, second-to-last paragraph: **I can't make any sense of this. Can anyone explain this or provide some examples?**

Commentary: Computing parameters of a maxent distribution

Unfortunately, Jaynes doesn't say much about how one finds the specific parameter values λ_i that achieve the desired expectations F_i. When the functions $f_i(x)$ have bounded, nonnegative values, the generalized iterative scaling and improved iterative scaling algorithms, discussed in the following references, can be used:

- J. Darroch and D. Ratcliff, ``Generalized iterative scaling for log-linear models,'' *Ann. Math. Statist.* 43, 1470-1480, 1972.
- S. Della Pietra, V. Della Pietra, and J. Lafferty, ``Inducing features of random fields,'' *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19, number 4, pp. 380-393, 1997. (Available here.)
- A. Berger, ``The improved iterative scaling algorithm: a gentle introduction.'' (Available here.)

These algorithms are most useful when the partition function cannot be efficiently computed.

If the partition function $Z(\lambda)$ can be efficiently computed, then one can find the parameter values that produce the desired expected values F_i by maximizing the function

$$h(\lambda) \equiv -\log Z(\lambda) - \sum_i \lambda_i F_i.$$

Note that $h(\lambda)$ is just the $1/N$ times the log of the likelihood function for the maxent form when the data are such that the average of each $f_i(x)$ is F_i. The reason this works is that

$$\frac{\partial h(\lambda)}{\partial \lambda_i} = E[f_i(x) \mid \lambda] - F_i;$$

at the maximum, the derivatives are zero, so $E[f_i(x) \mid \lambda] = F_i$.

To better understand why maximizing h is useful, let us consider the *discrepancy* $\delta(q \mid p)$ (a.k.a. directed divergence or Kullback-Liebler divergence) between an approximation q to a distribution and the distribution p itself. This is defined as $E[\log(p(x)/q(x))]$, where the expectation is taken over p. The discrepancy is always nonnegative, and equal to zero only if the two distributions are identical. If base-2 logarithms are used, the discrepancy may be thought of as the number of bits of information lost by using the approximation.

We are interested in a particular parameter vector λ_0 that gives $E[f_i(x) \mid \lambda_0] = F_i$ for all i. Our current best guess λ for that parameter vector defines a distribution that may be considered an approximation to the distribution obtained using the unknown λ_0. The discrepancy between these is

$$\begin{aligned}
\delta(p_\lambda \mid p_{\lambda_0}) &= E[\log p(x \mid \lambda_0) - \log p(x \mid \lambda) \mid \lambda_0] \\
&= -\log Z(\lambda_0) - \sum_i \lambda_{0,i} F_i + \log Z(\lambda) + \sum_i \lambda_i F_i \\
&= -h(\lambda) + \text{terms not involving } \lambda
\end{aligned}$$

So increasing $h(\lambda)$ decreases the discrepancy from the desired distribution.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 12: Ignorance priors and transformation groups

- p. 375, equation (12.7): Insert a minus sign in front of the integral.
- p. 378, line 4: ``Harr'' should be ``Haar.''
- p. 378, equation (12.19): `` $\psi(x', \nu', \sigma')$ '' should be `` $\psi(x', \nu', \sigma') dx'$ ''.
- p. 381, second half of page: ``lies in the equations $x' = ax + b$, $\nu' = ax + b$'' should be ``lies in the equations $x' = ax + b$, $\nu' = a\nu + b$''.
- p. 382, equation (12.37): The right-hand-side of the equation is wrong; it should be `` $\exp(-\lambda t)(\lambda t)^n / n!$''.
- p. 384, equation (12.48): the denominator of the left-hand side should be $1 - \theta + a\theta$.
- p. 385, equation (12.51): ``$(n-1)!$'' should be ``$(n-2)!$''.
- p. 386, fourth full paragraph: ``Kendell'' should be ``Kendall.''
- p. 394, third full paragraph, second line: ``James Clark Maxwell'' should be ``James Clerk Maxwell.''

Commentary on 12.4.3: Unknown probability for success

Chapters 11 and 12 I found quite exciting and useful, as construction of reasonable priors is a subject that seems to get short shrift in most books on

Bayesian methods, and the notion of an objective prior, that encodes exactly the information one has at hand and nothing more, is quite appealing.

However, I disagree with Jaynes's construction in 12.4.3 of an ignorance prior for an "unknown probability for success" θ, which he concludes should be an improper prior proportional to $\theta^{-1}(1-\theta)^{-1}$ over the interval $[0,1]$. (This appears to have been first suggested as an ignorance prior by J. Haldane in 1932.) I will argue that Jaynes's rules point to the uniform distribution over $[0,1]$ as the appropriate ignorance prior. I'll begin by critiquing specific passages in 12.4.3.

- p. 383, second full paragraph:

  > For example, in a chemical laboratory we find a jar containing an unknown and unlabeled compound. We are at first completely ignorant as to whether a small sample of this compound will dissolve in water or not. But, having observed that one small sample does dissolve, we infer immediately that all samples of this compound are water soluble, and although this conclusion does not carry quite the force of deductive proof, we feel strongly that the inference was justified. Yet the Bayes-Laplace rule [uniform prior] leads to a negligibly small probability for this being true, and yields only a probability of $2/3$ that the next sample tested will dissolve.

  Critique: This example is irrelevant for evaluating proposed ignorance priors over θ, as this is a situation where we have quite substantial prior information. We know that the relevant information in determining whether a sample of some solid compound will dissolve in water is

  - the chemical identity of the sample,
  - the quantity of sample,
  - the quantity of water, and
  - the temperature.

  All of these are factors we can easily control, and so if we repeat the experiment with the same unknown compound, keeping the other factors the same, we strongly expect to get the same result. That is, this prior information tells us that theta should be (nearly?) 0 or (nearly?) 1, given any particular values for the above four factors.

- p. 383, third full paragraph and onward:

  > [...] There is a conceptual difficulty here, since $f(\theta)d\theta$ is a `probability for a probability'. However, it can be removed by carrying the notion of a split personality to extremes; instead of supposing that $f(theta)$ describes the state of knowledge of any one person, imagine that we have a large population of individuals who hold varying beliefs about the probability

> for success, and that $f(theta)$ describes the distribution of their beliefs.

Critique: This artifice is unnecessary. Following Jaynes's advice to start with the finite and take the infinite only as a well-defined limit, we can begin by considering a case of n trials, and define $\theta = (\#\ \text{successes})/n$.

Our distribution for theta is then a probability of a frequency, not a probability of a probability, and there is no conceptual difficulty. We then take the limit as $n \rightarrow \infty$.

- Continuing:

> Is it possible that, although each individual holds a definite opinion, the population as a whole is completely ignorant of θ? What distribution $f(theta)$ describes a population in a state of total confusion on the issue? [...]
>
> Now suppose that, before the experiment is performed, one more definite piece of evidence E is given simultaneously to all of them. Each individual will change his state of belief according to Bayes' theorem; Mr. X, who had previously held the probability for success to be
>
> $$\theta = p(S \mid X) \qquad (12.42)$$
>
> will change it to
>
> $$\theta' = p(S \mid E, X) = [\text{omitted}] \qquad (12.43)$$
>
> [...] This new evidence thus generates a mapping of the parameter space $0 \leq \theta \leq 1$ onto itself, given from (12.43) by
>
> $$\theta' = \frac{a\theta}{1 - \theta + a\theta} \qquad (12.44)$$
>
> [...] If the population as a whole can learn nothing from this new evidence, then it would seem reasonable to say that the population has been reduced, by conflicting propaganda, to a state of total confusion on the issue. We therefore define the state of `total confusion' or `complete ignorance' by the

> condition that, after the transformation (12.44), the number of individuals who hold beliefs in any given range $\theta_1 < \theta < \theta_2$ is the same as before.

Critique: I find this characterization of complete ignorance to be quite puzzling. I just don't see any reason why this corresponds to any notion of complete ignorance. Furthermore, there are certain possible new pieces of evidence E that *must* change the overall distribution of beliefs -- for example, E might be frequency data for the first N trials, or even a definite statement about the value of θ itself. There is also some ambiguity here. Inference about θ only makes sense in the context of repeated trials; so, does S above really mean S_i (success at i-th trial) for some arbitrary i? If so, we must also assume that E is carefully chosen so that $p(E \mid S_i, X)$ has no dependence on (unobserved values of) i, so that $p(S_i \mid E, X)$ remains independent of i.

- p. 384, sentence following equation (12.43):

> This new evidence thus generates a mapping of the parameter space $0 \leq \theta \leq 1$ onto itself, given from (12.43) by
>
> $$\theta' = \frac{a\theta}{1 - \theta + a\theta} \qquad (12.44)$$
>
> where
>
> $$a = \frac{p(E \mid S, X)}{p(E \mid F, X)}. \qquad (12.45)$$

Critique: It seems to me that Jaynes is here committing an error that he warns against elsewhere: erroneously identifying distinct states of information as the same. In particular, a is a function of the particular individual X, since we are conditioning on different states of information for each individual. In my view, this destroys the entire construction, as we no longer have the transformation (12.44).

Here is my alternate proposal for an ignorance prior, following Jaynes's own advice. We begin with section 12.3, ``Continuous distributions,'' wherein Jaynes writes,

> In the discrete entropy expression

$$H_I^d = -\sum_{i=1}^{n} p_i log[p_i]$$

> we suppose that the discrete points x_i, $i = 1, 2, \ldots, n$, become more and more numerous, in such a way that, in the limit $n \to \infty$,
>
> $$\lim_{n->\infty}(\text{no. of points in } a < x < b)/n = \int_a^b dx\, m(x).$$
>
> If this passage to the limit is sufficiently well-behaved, [...] [t]he discrete probability distribution p_i will go over into a continuous probability $p(x \mid I)$ [...] The `invariant measure' function, $m(x)$ is proportional to the limiting density of discrete points.

Then at the beginning of p. 377, Jaynes writes,

> Except for a constant factor, the measure $m(x)$ is also the prior distribution describing `complete ignorance' of x.

On p. 376, last complete paragraph, Jaynes motivates the introduction of invariance transformations by writing,

> If the parameter space is not the result of any obvious limiting process, what determines the proper measure $m(x)$?

thus strongly implying that if there is an obvious limiting process, this is the preferred method for constructing $m(x)$.

But in this problem there is, in fact, an obvious limiting process -- the one mentioned at the beginning of this commentary. That is, we start by considering a finite case of n trials, define $\theta = (\#\text{ successes})/n$, and define

$$p(x_1, ..., x_n|\theta, I)$$

as in section 3.1 (sampling without replacement). (x_i is 1 if the i-th trial is a success, and 0 otherwise.) Since θ has a finite set of possible values, and ``ignorance'' means we are placing no constraints on the distribution over theta, Chapter 11 tells us that we should use the maximum-entropy distribution for θ, i.e., the uniform distribution over

$$0, 1/n, 2/n, ..., (n-1)/n, 1.$$

In the limit as $n \to \infty$ while k remains fixed we get

$$p(x_1, ..., x_k \mid \theta, I) = \theta^s (1-\theta)^{n-s},$$

where $s = \sum_i x_i$, and the prior over θ turns into a uniform pdf over $[0,1]$.

As a final note, I have some misgivings about even this solution. The problem is that we are not, in fact, completely ignorant about θ. We know of some additional structure to the problem -- that is, we know that θ (in the finite case) is derived from the results of the trials x_i via $\theta = \sum_i x_i / n$. One could argue that we should therefore derive the prior over θ from the ignorance prior over $x_1, \ldots, x_n$. As Jaynes discusses in Chapter 3 (?), in the limit of $n \to \infty$ this amounts to a prior that gives probability 1 to $\theta = 1/2$, and we find that we are incapable of learning--

$$p(x_{k+1} \mid x_1, \ldots, x_k, I) = p(x_{k+1} \mid I) = 1/2.$$

Thus it seems that any nondegenerate prior for θ is, in some sense, informative. At the very least, it tells us that the various trials are subject to some common logical influence.

Commentary on 12.4.3: Other approaches

Arnold Zellner contributed the following references to other priors that have been suggested for the binomial parameter (probability of success):

- *Theory of Probability* (1967), by Sir Harold Jeffreys, pp. 123-125, contains a discussion of various priors for the binomial parameter. He believes that the uniform prior is too flat at the end points and that the improper prior $\theta^{-1}(1-\theta)^{-1}$ goes up too much at the end points, 0 and 1, placing too much probability mass in the vicinity of 0 and 1. Therefore he lumps some probability up at zero and some at 1 with the rest spread uniformly between 0 and 1.
- *Bayesian Analysis in Econometrics and Statistics*, by Arnold Zellner, pp. 117-118, discusses a ``maximal data information'' prior proportional to $\theta^\theta (1-\theta)^{1-\theta}$. This is a bowl-shaped density that is proper and whose value at 0 and 1 is twice its value at 0.5. Elsewhere in the same book he discusses the derivation of ``maximal data information priors'' in more detail.

Zellner's maximal data information prior is defined as that prior which maximizes a quantity G defined as the prior average information in the data pdf, minus the information in the prior pdf. The ``information'' here is intended to be negative the entropy.

Zellner's approach to ignorance priors and Jaynes's approach in PTLOS appear to be incompatible. Jaynes argues that the proper definition of entropy for a continuous distribution involves use of the measure $m(x)$ describing complete ignorance for the sample space, so you must already *have* your ignorance prior in hand before you can even define the entropy/information of a prior pdf. Zellner agrees on the necessity of choosing an information measure $m(x)$ for defining the entropy of a continuous distribution, but considers this to be a separate problem--much like that of choosing a temperature scale (Celsius, Fahrenheit, or Kelvin)--from that of producing a least informative prior density.

See also ``Some Aspects of the History of Bayesian Information Processing'' (to appear, *Journal of Econometrics*), which may be found here.

Commentary on 12.4.4: Bertrand's problem

One may be confused by the fact that integrating θ out of $f(r,\theta)$ (defined in (12.67)) and doing the appropriate change of variables from r to x does not yield (12.68). This is because $f(r,\theta)$ is not, strictly speaking, a pdf in the variables r and θ -- it is an *area* density. The (r,θ) pdf is actually $r\,f(r,\theta)$. (See the first paragraph under ``Rotational invariance,'' where Jaynes writes ``What probability density $f(r,\theta)\,dA = f(r,\theta)\,r\,dr\,d\theta$ should we assign...'')

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Chapter 13: Decision theory, historical background

- p. 403, line after equation (13.9): Insert a minus sign before the summation symbol in the definition of $I(q;p)$.

- p. 403, footnote 1: Arguably, ``predicted nonstorm'' should be ``nonpredicted nonstorm,'' as no storm occurred, contrary to predictions.

- p. 411, equation (13.18): A couple of small technical corrections. `` $\partial L/\partial\beta = 0$'' should be `` $f(x \mid \alpha)\partial L/\partial\beta = 0$''. The sentence that follows the equation is incorrect -- the equation really says that $\beta(x)$ must be a stationary point of L, regardless of x, which leads to the useless solution mentioned later on ($\beta(x) = \alpha$).

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 14: Simple applications of decision theory

- p. 428, equation (14.9): This isn't quite stated correctly. We cannot have $p(D \mid V) = p(D \mid V, Y)$ for all propositions $Y \neq D$; consider, for example, defining $Y = D \wedge V$. A correct statement might require that D be a proposition asserting particular values for model variables d_i, $1 \leq i \leq n$; that V be a proposition asserting a particular value for a model variable v, distinct from the variables d_i; and that Y be a proposition asserting a particular value for another model variable y, distinct from v and the variables d_i.

- p. 429, Theorem: This isn't stated quite correctly. The fact that D is a possible decision, given V, does not imply that $p(V \mid D) \neq 0$. Is $p(V \mid D) \neq 0$ meant as an extra condition? Furthermore, in equation (14.14), given that $P(V \mid D) \neq 0$, the $\Leftarrow$ implication holds, but the $\Rightarrow$ implication holds only if $p(Y \mid V) \neq 0$.

- p. 433, equation (14.32), first line: `` $(VS_1 \mid X)$'' should be `` $p(VS_1 \mid X)$''.

- p. 440, first full paragraph: ``Woodword'' should be ``Woodward.''

- p. 444, first line: the reference to (11.46) should be (11.48).

- p. 447, equation (14.79): I believe the variable r on the right-hand side of the equation should be omitted, to give a numerator of $\exp(-n(\lambda + \mu))$.

- p. 447, equation (14.82): `` $\langle m_1 \rangle_1$ '' should be `` $\langle m_1 \rangle$ ''.
- p. 448, equation (14.83): `` $\langle n_1 \rangle_1 / \langle m_1 \rangle$ '' should be `` $\langle n_1 \rangle / \langle m_1 \rangle^2$ ''.

Commentary on 14.7.3: Solution for Stage 4

Jaynes states,

> ...this new knowledge [a specific order for 40 green widgets], which makes the problem so hard for our common sense, causes no difficulty at all in the mathematics. The previous equations still apply, with the sole difference that the stock S_3 of green widgets is reduced from 50 to 10.

The above seems intuitively plausible, but let's follow Jaynes's advice to always carefully derive results from the basic laws of probability theory, rather than making intuitive leaps. The new information for Stage 4 is a proposition instead of an expected value for the prior distribution to satisfy; the proper procedure then is to take the prior distribution of Stage 3 and condition on $w_{40} \geq 1$ to obtain the Stage 4 distribution. Let's do that now.

Recall that u_r is the number of orders for r red widgets, v_y is the number of orders for y yellow widgets, and w_g is the number of orders for g green widgets. From (14.72), the Stage 3 prior distribution factors into independent distributions for the sets of variables u_r, v_y, and w_g:

$$p(u_1, \ldots; v_1, \ldots; w_1, \ldots) = p_1(u_1, \ldots) p_2(v_1, \ldots) p_3(w_1, \ldots).$$

From (14.69), (14.70), and (14.71), we also see that p_3 factors into independent distributions for each variable w_g:

$$P(w_g = w) = C_0 \exp(-(\lambda_3 g + \mu_3) w), \text{ where } C_0 \text{ is a normalization constant}.$$

Thus, conditioning on $w_{40} \geq 1$ affects only the distribution for w_{40}. Furthermore, the distribution for w_{40} is an exponential distribution, and as such has the easily verified general property that for any $n \geq 0$,

$$P(w_{40} = n + w \mid w_{40} \geq n) = P(w_{40} = w).$$

Using $n = 1$, Jaynes's assertion follows directly.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 15: Paradoxes of probability theory

- p. 460, equation (15.18): Should be `` $p^{(0)} = (1, 0, \ldots, 0)^T$ '' or, equivalently, `` $p_i^{(0)} = \delta_{i,0}$ ''.
- p. 467, equation (15.42): insert a minus sign in front of the argument to the exponential function.
- p. 473, second full paragraph: ``the right-hand sides of (15.58) and (15.61)'' should be ``the **left**-hand sides...''.
- p. 475, equation (15.67): insert a minus sign before the argument to the exponential function.
- p. 481, equation (15.89): $\frac{1}{2}\omega^2$ in the exponential function should be $-\frac{1}{2}\omega^2$.
- p. 481, equations (15.87) and (15.88): The factor $\omega^{n+\gamma-1}$ should be $\omega^{n+\gamma-2}$.

Commentary: The Marginalization Paradox

After spending many hours going over Jaynes's treatment of the Marginalization Paradox, I've come to the conclusion that he got this one wrong: (15.72) is wrong, and (15.70) is the correct formula also for B_1.

Sections 15.8 and 15.9 are a puzzling anomaly, as Jaynes unaccountably breaks a number of the rules he emphasizes so often elsewhere in the book, and this leads him into error. I've written up my conclusions in a separate note (postscript, PDF). In summary, here is what I've shown:

- The paradox arises from an unnoticed divergent integral that shows up when one tries to go from $p(\zeta \mid y,z) = f(\zeta,z)$ to $p(\zeta \mid z) = f(\zeta,z)$; this step is invalid because it requires multiplying by $p(y \mid z)$ and then integrating out y, but it seems to have escaped notice that the improper prior over η results in $p(y \mid z)$ also being improper.
- In the specific case of the change-point problem, if one derives $p(\zeta \mid z)$ for the proper prior $\pi_b(\eta) \propto \eta^a \exp(-b\eta)$, then takes the limit as $b \to 0$ (going to the limiting improper prior $\pi(\eta) \propto \eta^a$), one obtains B_2's answer (15.70), and not B_1's answer (15.72).
- The issue of non-uniform convergence plays an important role in this problem, and as $p(\zeta \mid y,z)$ converges to (15.72), the distribution $p(y \mid z)$ retains significant probability mass in the (ever-smaller) region where $p(\zeta \mid y,z)$ is far from convergence.

It's worth noting, however, that my resolution of the paradox was obtained simply by following the practices Jaynes advocates in PTLOS.

Is it a disaster for Bayesian analysis if we have to abandon the use of improper priors? I don't think so. As Jaynes points out, the really important use of improper priors is as a zero-point for constructing maximum-entropy priors. Furthermore, he shows in one problem after another that even in situations where one might be tempted to say that we are totally ignorant about some parameter, simple common-sense reasoning and application of physical constraints allow us to create a defensible proper prior. There are, in fact, some pretty good reasons (beyond the MP) to stick to proper priors:

- A lot of interesting problems can't be solved analytically, requiring instead the use numerical methods that generally won't work with improper priors. In particular, the use of Markov Chain Monte Carlo (e.g., BUGS) has become increasingly popular over the last decade, and this requires proper priors.
- Model comparison (see Chapter 20) -- one of the more interesting applications of Bayesian methods -- requires proper priors.

Philip Dawid informs me that in 1996 he, Stone, and Zidek also wrote a response to Chapter 15, based on the version of PTLOS available on the Internet at that time; you can find it here as report 172 for 1996.

Some final technical comments:

- One obtains (15.87) via the change of parameters $\omega = R/\sigma$.
- On p. 482 Jaynes talks about applying (15.89) to obtain a posterior over ζ conditional on r. That is, (15.89) is to be used as a likelihood.

Unfortunately, the proportionality in (15.89) retains only factors dependent on r, when instead it needs to retain those factors dependent on ζ or σ (in particular, a factor of $\exp(-n\zeta^2/2)$ is missing.

(This comment comes from DSZ's response, mentioned above.)

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Chapter 16: Orthodox methods: historical background

- p. 494, second full paragraph, last sentence: ``Feinberg'' should be ``Fienberg.''
- p. 495, footnote 1: ``Gossett'' should be ``Gosset.''
- p. 502, second line after eqn. (16.11): ``(median) $\pm$ (interquartile distance).'' Strictly speaking, this should be ``median and quartiles,'' as what is meant is the interval from the lower quartile to the upper quartile. The interval from (median) $-$ (interquartile distance) to (median) $+$ (interquartile distance) is twice as wide.
- p. 503, line 1: ``criteria'' should be ``criterion.''

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

- Miscellaneous Commentary

Chapter 17: Principles and pathology of orthodox statistics

- p. 512, first full paragraph: ``...it seems obvious that, at least for large $n$, this has made things worse instead of better.'' I believe ``large'' should be ``small.''
- p. 513, footnote 2: ``indavertently'' should be ``inadvertently.''
- p. 518, first line after (17.27), and last line on page: ``Schwartz'' should be ``Schwarz.''
- p. 519, eqn. (17.30): $d\alpha$ (in the denominator) should be dx.
- p. 519, eqn. (17.32): $p(x \mid a)$ should be $p(x \mid \alpha)$.
- p. 520, third full paragraph, start of third line: ``Schwartz'' should be ``Schwarz.''
- p. 530, eqn. (17.65): ``$=$'' should be ``$\propto$''.
- p. 538, eqn. (17.93): $\overline{s^2t^2}$ should be $(\overline{s^2})(\overline{t^2})$.
- p. 541, eqn. (17.107): As far as I can tell, this equation is wrong. The final term of the rightmost expression should be

  $$\frac{(\overline{t^2})(\overline{se}) - (\overline{st})(\overline{te})}{(\overline{s^2})(\overline{t^2})}$$

  instead of $\overline{se}/\overline{s^2}$.
- p. 543, first line after (17.118): ``Schwartz'' should be ``Schwarz.''

- p. 544, eqn. (17.125): subscripts of t and t' are missing from the summation symbol $\sum$; the factor $1/N$ for the middle expression should be $1/N^2$; and the rightmost expression should be $\overline{g^2}\sigma^2/N$.
- p. 544, eqn. (17.126): $\sqrt{\overline{g^2}}$ should be $\sqrt{\overline{g^2}/N}$.
- p. 545, eqn. (17.128): The denominator needs to be squared--that is,

$$\text{replace} \qquad \overline{s^2}(1-r^2) \qquad \text{with} \qquad \left(\overline{s^2}(1-r^2)\right)^2.$$

- p. 545, eqn. (17.129): this should read

$$A_0 \pm \sigma\sqrt{\frac{1}{N\overline{s^2}(1-r^2)}}.$$

- p. 545, eqn. (17.130): $\bar{A}$ should be A_0.
- p. 546, eqn. (17.135): T_j, here and in the rest of the section, should be Φ_j. Alternatively, Φ_k in (17.132) should be replaced with T_k.
- p. 546, eqn. (17.136): I think the series should be $A_1, \ldots, A_6, B_1, \ldots, B_6$, as A_0 provides only a constant term, which is subsumed by the trend. Likewise, in (17.137), replace ``$0 \le k \le 6$'' with ``$1 \le k \le 6$''.
- p. 548, middle of page: ``F is the $(N \times n)$ matrix of model functions.'' We called this matrix G on the previous two pages.
- p. 549, eqn. (17.164): This is redundant; it just repeats eqn. (17.161), which appeared half a page earlier.
- p. 551, third full paragraph, third-to-last line: ``student'' should be ``Student.''

Miscellaneous Commentary

- p.517, eqn. (17.23): the normalization constant $1/n!$ is **not** a typo, even though it is the same as the normalization constant in (17.17). That's right: defining $f(n,l) \equiv \exp(-l)l^n$, we have that *both* $\sum_{n=0}^{\infty} f(n,l)$ *and* $\int_0^{\infty} f(n,l)dl$ are equal to $n!$.

- p. 517, eqn. (17.23) and preceding line: Since l is a scale factor, it might seem more reasonable to use an ignorance prior $\propto l^{-1}$ instead of a uniform prior over l. The only effect this has is to replace n with $n-1$ in the posterior $\exp(-l)l^n/n!$ and the corresponding posterior expectation formula (17.23).
- p. 518, footnote 6: This story about Kendall and Jeffreys is probably based on a confusion with G.U. Yule. Kendall was a student at St John's, but he did not become a Fellow. It was Jeffreys and Yule who were both Fellows for many years. [Contributed by Nick Cox.]
- p. 519, eqn. (17.31): Note that the ``change of parameters'' mentioned in the preceding text is *not* a reparameterization of a distribution. Recall that to achieve the minimum variance, there must exist some $q(\alpha)$ such that

  $$\frac{\partial \log p(x \mid \alpha)}{\partial \alpha} = q(\alpha)(\beta(x) - \langle\beta\rangle).$$

  We then define $l(\alpha)$ implicitly via $q(\alpha) = -l'(\alpha)$; the purpose is to make (17.32) come out neatly.
- p. 519, eqn. (17.33): Note that x is a *vector* of observables, not a single observable. Writing x_1^n for the vector $(x_1, \ldots, x_n)$, (17.33) may be written more explicitly as

  $$p(x_1^n \mid \alpha) = \frac{m_n(x_1^n)}{Z_n(l_n)} \exp(-l_n(\alpha)\beta_n(x_1^n));$$

  so to make (17.33) correspond to the results of Chapter 11, we must choose $n = 1$. If the estimators β_i are related by

  $$\beta_n(x_1^n) = 1/n \sum_{i=1}^{n} \beta_1(x_i),$$

  then the distribution that minimizes $\mathrm{var}(\beta_1)$ also minimizes $\ml, assuming the x_i are assigned independent and identical distributions:

$$p(x_1^n \mid \alpha) = \prod_i \frac{m_1(x_i)}{Z_1(l_1)} \exp\left(-l_1 \sum_i \beta_1(x_i)\right)$$

is identical to (17.33) for arbitrary n if we choose $m_n(x_1^n) = \prod_i m_1(x_i)$, $l_n = nl_1$, and $Z_n(l_n) = Z_1(l_1)^n$.

- p. 530, end of section, and p. 549, end of section: A more detailed discussion of the topic can be found in Larry Bretthorst's Ph.D. dissertation, *Bayesian Spectrum Analysis and Parameter Estimation* (Bretthorst got his doctorate under Jaynes.) This was published as Lecture Notes in Statistics 48, but is out of print; the best way to get it now is to download it from Bretthorst's web page on Probability Theory as Extended Logic (near the end of the page).

- p. 531 (Section 17.7, ``The folly of randomization'') Jaynes's example of using Monte Carlo methods to do a simple one-dimensional integral is a bit misleading, in that most uses of MC methods involve high-dimensional integrals and sampling from from distribution $p(x)$ to compute an expectation over that distribution; in such cases, using an n -dimensional grid to numerically integrate is impractical. However, the emergence of techniques such as Latin squares sampling and the use of quasi-random sequences to improve the convergence of MC integration certainly seems to support Jaynes's contention (on p. 532) that ``Whenever there is a randomized way of doing something, there is a nonrandomized way that yields better results from the same data, but requires more thinking.'' I would only add the caveat that the better, nonrandomized way often enough seems to require *much* more thinking, and years of research.

- p. 545, end of section 17.10.5: Chapter 17 is one of those roughed-out chapters that Jaynes never really finished, and it's unclear where he intended to go with this comparison of the Bayesian vs. orthodox estimators using the orthodox criterion of performance. However, let's continue where Jaynes left off. Using the corrected versions of equations (17.129) and (17.130) [see errata above], the expected squared error for the Bayesian estimator is

  $$\frac{\sigma^2}{N\overline{s^2}(1-r^2)},$$

  whereas the expected squared error for the orthodox estimator is

  $$\frac{\sigma^2(1-r^2)}{N\overline{s^2}} + r^4 A_0^2,$$

the extra term coming from the bias of the estimator.

To better understand the behavior of r, let us assume that $\omega = 2\pi/k$ for some integer $k > 2$, and let $N = nk$ (a complete number of cycles) for ease of analysis. It is easily shown that $r^2 \leq 1$. Using

$$\overline{s^2} = \frac{1}{k}\sum_{j=1}^{k} \sin^2(2\pi j/k) = 0.5$$

and

$$\overline{t^2} = \frac{1}{N}\sum_{t=1}^{N} t^2 = \frac{(N+1)(2N+1)}{6}$$

and

$$\overline{st} = \frac{1}{k}\sum_{j=1}^{k} j\sin(2\pi j/k) \approx -\frac{k}{2\pi} \quad \text{for moderate to large } k.$$

we find that

$$r^2 = \frac{72\overline{st}^2}{(N+1)^2(2N+1)^2}.$$

This goes to zero quite rapidly; even for the minimum values of k=3 and n=1, giving N=3, we have $r^2 \approx 0.00765$. As a result, the difference in the expected squared error of the two estimators is usually small. Experimenting with different values for k, n, and the ratio $\rho \equiv \sigma/A_0$, I found substantial advantage for the Bayesian estimator only for small values of n and ρ; in particular, for $n = 1$ I found the expected square error for the orthodox estimator to become much higher than that for the Bayesian estimator as ρ gets smaller than 0.02. For other cases I found only very small differences, often favoring the orthodox estimator.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 18: The A_p distribution and rule of succession

- p. 563, eqn. (18.22): $P(N_p \mid X)$ should be $P(N_n \mid X)$.
- p. 570, eqn. (18.39): $P(n_1 \cdots n_k \mid X)$ should be $P(n_1 \cdots n_K \mid X)$ (K instead of k).
- p. 576, second paragraph, end of line six: ``existance'' should be ``existence.''
- p. 576, second paragraph: ``As we saw earlier in this chapter, even the $+1$ and $+2$ in Laplace's formula turn up when the `frequentist' refines his methods...'' Actually, this is discussed *later* in the chapter -- see eqn. (18.68).
- p. 577, text preceding equation (18.58): The reference to equation (18.55) should probably be (18.56).
- p. 579, last paragraph of section 18.15, line six: ``thoery'' should be ``theory.''
- p. 580, first line: (n/M) should be (n/N).
- p. 580, third line after (18.69): ``Pearson and Clopper'' should be ``Clopper and Pearson.''
- p. 581, third line from bottom: M_δ should be $M\Delta$.
- p. 582, eqn. (18.73): F^1 should be f^1; also, in the second line, first factor, $(n+1)/(N+2)$ should be $(n+1)/(N+3)$.

- p. 582, first three lines after (18.73): F^1 should be f^1 in each instance.

- p. 582, eqn. (18.76): M_n should be $.

- p. 583, eqn. (18.78), second line: M_p should be Mp.

- p. 583, eqn. (18.79): $\overline{M^2}$ should be $\overline{m^2}$ and the factor $(n-(n+1)/(N+2))$ should be $(1-(n+1)/(N+2))$.

- p. 583, eqns. (18.80), (18.81), and (18.82): M_p should be Mp.

- p. 584, second full paragraph, lines 3 and 6: F^1 should be f^1.

- p. 584, second full paragraph, end of line 12, and also line 13: ``uncertainity'' should be ``uncertainty.''

- p. 586, eqn. (18.87): $\binom{N-m}{m}$ should be $\binom{N-n}{m}$.

- p. 587, first line after (18.93): ``If we substitute (18.93)...'' Should this be (18.91)?

Miscellaneous Comments

- p. 554, eqn. (18.1): This definition cannot hold true for arbitrary propositions E; for example, what if E implies A? This kind of problem occurs throughout the chapter. I don't think you can really discuss the A_p distribution properly without explicitly introducing the notion of a sample space and organizing one's information about the sample space as a graphical model in which A has a single parent variable θ, with A_p defined as the proposition $\theta = p$. For those unfamiliar with graphical models / Bayesian networks, I recommend the following book:
  - Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (J. Pearl).

- p. 555, eqn. (18.3): This appears to be at odds with Chapter 12, which advocates the improper Haldane prior (proportional to $p^{-1}(1-p)^{-1}$) as describing the ``completely ignorant population.'' However, that chapter also argues that the Haldane prior applies when one does not even know whether or not both outcomes are possible...and that the uniform prior applies if one *does* know that both outcomes are possible. (I argue in my comments on Chapter 12 that the uniform prior is the correct ignorance prior in general anyway.)

- p. 555, eqn. (18.7): For those who may be confused by this equation, the integrand $p\,(A_p \mid E)$ means $p \cdot (A_p \mid E)$, not the probability density of A_p given E.

- p. 556, third line after (18.9): ``But suppose that, for a given E_b, (18.8) holds independently of what E_a might be; call this `strong irrelevance.' '' If (18.8) holds for any proposition E_a, then in particular it holds for the proposition $E_a \equiv \neg E_b \vee A$; then from (18.8) we have

  $$P(A \mid \neg E_b \vee A) = P(A \mid E_a) = P(A \mid E_a \wedge E_b) = P(A \mid E_b \wedge A) = 1;$$

  then since $\neg E_b$ implies $\neg E_b \vee A$, we also have $P(A \mid \neg E_b) = 1$. Thus, this definition of ``strong irrelevance'' actually ensures that E_b is *highly relevant* to A. As before, this discussion really needs to be rewritten in terms of graphical models to get it right, in particular making use of the notion of d-separation.

- p. 583, eqn. (18.78): To get the second line from the first, use these identities:
  - $E[m^2] = E[m]^2 + V[m]$.
  - For the binomial distribution, $E[m] = Mp$ and $V[m] = Mp(1-p)$.

- p. 586, third full paragraph, first sentence: ``An important theorem of def Finetti (1937) asserts that the converse is also true:...'' What Jaynes says here is not true for finite N; it only holds in the limit as $N \to \infty$. As a counterexample, consider draws without replacement from an urn containing $N = b + w$ balls, with b black and w white. The sequence of draws $x_1, \ldots, x_N$ is exchangeable, but $P(x_1, \ldots, x_N \mid N)$ cannot be generated by any A_p distribution. To see this, note that once we know the values of $x_1, \ldots, x_{N-1}$ we also know the value of x_N with certainty, because we know the total number of balls of each color in the urn.

- p. 586, third full paragraph, second sentence: Even in the limit $N \to \infty$, for this statement to be true in general we must allow $g(p)$ to be a generalized function--that is, we must be able to assign nonzero probability mass to single points using delta functions.

- p. 587, second sentence after (18.89): For this sentence to be true requires that matrix A be nonsingular, where $a_{n,k} \equiv \alpha_k(N, n)$. To see that A is in fact nonsingular, note that $a_{n,k} = 0$ for $k < n$ and $a_{k,k} = 1$.

Then for arbitrary x one can solve for β in $A\beta = x$ by backsubstitution.

Next Up Previous

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Subsections

Chapter 20: Model comparison

Miscellaneous Comments

- p. 613, first paragraph of section 20.5: [Comment by Nick Cox.] The juxtaposition of Wegener and Jeffreys here is accidentally bizarre. Wegener's ideas of continental drift are now seen as a precursor of modern plate tectonics, which is no longer controversial. But for many years a leading objection to those ideas was Jeffreys's insistence that there was no physically possible mechanism for the process postulated (which does differ from what is now accepted). With perfect hindsight, therefore, Jeffreys backed the wrong horse here, despite his other outstanding contributions to geophysics.

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Chapter 22: Introduction to communication theory

- p. 637, last paragraph, third line: ``Michel Ventris'' should be ``Michael Ventris.''

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Appendix B: Mathematical formalities and style

- p. 670, first full paragraph, fourth line, and second full paragraph, last line: ``Schwartz'' should be ``Schwarz.''
- p. 668, first line of second paragraph, also first line of third paragraph: publication year of Lighthill book is 1958, not 1957.
- p. 668, sixth-to-last line: ``Aczel'' should be ``Aczél'' (add accent).
- p. 668, second-to-last line: ``Stieltjes'' should be spelled without an accent.

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

Appendix C: Convolutions and cumulants

- p. 679 (Appendix C), eqns. (C.19) and (C.21). The calligraphic F (which is used in the surrounding text to indicate the Fourier transform operator) should be a simple math italic F (indicating the first moment).
- p. 681, second-to-last line: ``is has only'' should be ``it has only.''

ksvanhorn.com

Home
Bayes Home
Jaynes Errata
Articles
Books
Software
Contact

Next Up Previous

References, Bibliography, and Author index

- p. 683: ``Andrews, D. R.'' should be ``Andrews, D. F.''
- p. 684: Title of Bell entry should be *Men of Mathematics* (``of,'' not ``and.'')
- p. 685: Borel (1924), title: should be ``traité'' and ``probabilités.''
- p. 686: Cournot: ``Theorie'' should be ``Théorie.''
- p. 688: ``Feinberg'' should be ``Fienberg.''
- p. 688: Einstein (1905b): ``contend'' should be ``content.''
- p. 690: Galileo: ``MacMillan'' should be ``Macmillan.''
- p. 692: Hardy: ``MacLearin'' should be ``Maclaurin.''
- p. 692: Harr: ``Harr'' should be ``Haar,'' and journal is wrong -- should be *Annals of Mathematics*.
- p. 697: Year of publication of Lighthill book is 1958, not 1957.
- p. 697: Little and Rubin: ``Little, J. F.'' should be ``Little, R. J. A.''
- p. 698: ``Pearson and Clopper'' should be ``Clopper and Pearson.'' Also, ``confidence in fiducial'' should be ``confidence or fiducial.''
- p. 700: Savage (1961): publisher is University of California Press.
- p. 702: Venn: ``MacMillan'' should be ``Macmillan.''
- p. 701: Stone (1965): ``Harr'' should be ``Haar.''
- p. 705: Ash: should be ``Ash, R. B. (1965)''
- p. 705: Barlow: should be ``Barlow, R.E.''
- p. 705: Barndorf-Nielsen: should be ``Barndorff-Nielsen.''
- p. 705: Barr and Feigenbaum: ``Kaufman'' should be ``Kaufmann.''
- p. 705-706: Bernado: should be ``Bernardo.''
- p. 706: Blanc-Lapierre and Fortet: ``Theorie'' should be ``Théorie,'' and ``Aleatoires'' should be ``Aléatoires.''
- p. 706: Boscovich: ``Geographique'' should be ``Géographique.''
- p. 706: Carnap: ``Routlege'' should be ``Routledge.''
- p. 709: Galton: ``MacMillan'' should be ``Macmillan.''
- p. 709: Gnedenko and Kolmogorov: ``epistomologic'' should be ``epistemologic.''
- p. 712: Jeffreys (1992): ``Vice-Mistress,'' not ``Mistress.''
- p. 713: Kindermann and Snall: ``Snall'' should be ``Snell.''
- p. 713: Legendre: ``méthods'' should be ``méthodes,'' and ``cométes'' should be ``comètes.''
- p. 713: Lindley (1956): journal should be *Ann. Math. Stat.*
- p. 713: Lindley (1971): ``Mathemathics'' should be ``Mathematics.''
- p. 713: Macdonald, second line of commentary: ``Ichthus'' should be

``Icthus.''

- p. 713: Mandelbrot: Title should be *Fractals: form, chance, and dimension.*
- p. 713: Martin and Thompson: ``Thompson'' should be ``Thomson.''
- p. 716: Quaster: should be ``Quastler.''
- p. 716: Reid (1959): author should be A. Rényi.
- p. 716: Robbins (1950 and 1956): ``Mathematics Statistics'' should be ``Mathematical Statistics.''
- p. 717: Siegmann: should be Siegmund.
- p. 718: Stone and Springer-Verlag: ``Springer-Verlag'' should be ``Springer.''
- p. 720: Wilson: Title should be *Entropy in urban and regional modelling.*
- p. 720: Zellner, fourth line of comment: ``validty'' should be ``validity.''
- p. 721: Cantrell should be Cantril.
- p. 721: Dawid, Philip A. should be Dawid, A. Philip.
- p. 722: Gossett, William Sealey should be Gosset, William Sealy.
- p. 722: Hansen: 193 should be 194.
- p. 722: Harr: Harr should be Haar, and 377 should be 378.
- p. 722: Howson: Cowlin should be Colin, and 126 should be 127.
- p. 722: Johnson: Ernes should be Ernest.
- p. 722: Poincarè: should be Poincaré.
- p. 722: Rescher: Nichola should be Nicholas.
- p. 723: Schwartz: 667 should be 668.
- p. 723: Tell: 416 should be 417.
- p. 723: Weierstraz: should be Weierstrasz (or Weierstrass).
- p. 723: Whitehead: North Whitehead should be Alfred North.
- p. 723: Zabel: should be Zabell.
- p. 724: Cramer-Rao: should be Cramér-Rao.
- p. 725: exchangable sequences: should be ``exchangeable sequences.''

Next Up Previous