Hanns Ludwig Harney

Bayesian Inference

Data Evaluation and Decisions

Second Edition

Springer

Bayesian Inference

Hanns Ludwig Harney

Bayesian Inference

Data Evaluation and Decisions

Second Edition

Springer

Hanns Ludwig Harney
Max Planck Institute for Nuclear Physics
Heidelberg
Germany

ISBN 978-3-319-41642-7 ISBN 978-3-319-41644-1 (eBook)
DOI 10.1007/978-3-319-41644-1

Library of Congress Control Number: 2016945790

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

The present book, although theoretical, deals with experience. It questions how to draw conclusions from random events. Combining ideas from Bayes and Laplace with concepts of modern physics, we answer some aspects of this question.

The book combines features of a textbook and a monograph. Arguments are presented as explicitely as possible with the aid of of appendices containing lengthy derivations. There are numerous examples and illustrations, often taken from physics research. Problems are posed and their solutions provided.

The theory presented in the book is conservative in that the most widely-known Gaussian methods of error estimation remain untouched. At the same time, some material is unconventional. The non-informative prior is considered the basis of statistical inference and a unique definition is given and defended. Not only does the prior allow one to find the posterior distribution, it also provides the measure one needs to construct error intervals and make decisions.

The example of binomial distribution — sketched on the book-cover — represents 300 years of statistics research. It was the first clearly formulated statistical model and the first example of statistical inference. We hope to convince the reader this subject is not yet closed.

Heidelberg, Germany Hanns Ludwig Harney

Acknowledgement

I thank my wife Barbara Teloo as well as Marie and Justus Teloo for their encouragement.

The help of R. Harney, C. von Savigny, H. Bernshausen and J. Teloo in drawing the figures is gratefully acknowledged. Discussions with O.-A. Al-Hujaj, C.I. Barbosa, V.E. Bunakov, A. Richter, B. Dietz-Pilatus, A. Müller, Ch. Fuhrmann and H.A. Weidenmüller helped to clarify many of the questions treated on the following pages.

Contents

About the Author

Hanns Ludwig Harney born in 1939, professor at the University of Heidelberg. He has contributed to experimental and theoretical physics within the Max-Planck Institute for Nuclear Physics at Heidelberg. His interest was focussed on symmetries, such as isospin and its violation, as well as chaos, observed as Ericson fluctuations. Since the 1990's, the symmetry properties of common probability distributions lead him to a reformulation of Bayesian inference.

Chapter 1
Knowledge and Logic

Science does not prove anything. Science infers statements about reality. Sometimes the statements are of stunning precision; sometimes they are rather vague. Science never reaches exact results. Mathematics provides proofs but it is devoid of reality. The present book shows in mathematical terms how to express uncertain experience in scientific statements.

Every observation leads to randomly fluctuating results. Therefore the conclusions drawn from them must be accompanied by an estimate of their truth, usually expressed as a probability. Such a conclusion typically has the form: "The quantity ξ inferred from the present experiment has the value $\alpha \pm \sigma$." An experiment never yields certainty about the true value of ξ. Rather the result is characterised by an interval in which the true value should lie. It does not even lie with certainty in that interval. A more precise interpretation of the above interval is: "The quantity ξ is in the interval $[\alpha - \sigma, \alpha + \sigma]$ with the probability $K = 0.68$." Trying to be even more precise one would say: "We assign a Gaussian distribution to the parameter ξ. The distribution is centered at α and has the standard deviation σ. The shortest interval containing ξ with the probability $K = 0.68$ is then $\alpha \pm \sigma$." In simplified language, the standard deviation of the assumed Gaussian distribution is called "the error" of the result, although "the" error of the result cannot be specified. One is free to choose the probability K and thus the length of the error interval.

The present book generalises the well-known rules of Gaussian error assignments to cases where the Gaussian model does not apply. Of course, the Gaussian model is treated too. But the book is animated by the question: How to estimate the error interval when the data follow a distribution other than Gaussian, for example, a Poissonian one? This requires us to answer the following general questions: What is, in any case, the definition of an error interval? How do we understand probability? How should the observed events x be related to the parameter ξ of interest?

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_1

1.1 Knowledge

The parameter that one wants to know is never measured directly and immediately. The true length of a stick is hidden behind the random fluctuations of the value that one reads on a meter. The true position of a spectral line is hidden in the line width that one observes with the spectrograph. The fluctuations have different causes in these two cases but they cannot be avoided. One does not observe the interesting parameter ξ. Rather, one observes events x that have a distribution p depending on ξ. Data analysis means to infer ξ from the event x, usually on the basis of a formally specified distribution p that depends parametrically on ξ. This parameter is also called the hypothesis that conditions the distribution of x. The connection between x and ξ is given by $p(x|\xi)$, in words, "the distribution p of x, given ξ." It must depend on ξ in such a way that different hypotheses entail different distributions of x so that one can learn from x about ξ.

Inferring ξ is incomplete induction. It is induction because it is based on observation, as opposed to logical deduction based on first principles. It is incomplete because it never specifies the true value. Note that even an experiment that produces a huge amount of data does not yield all possible data, and its repetition would produce a different event and thus a different estimate of ξ. For this reason, no experiment yields the true value of ξ, and inference of ξ means to assign a distribution to ξ. One assumes, of course, that all the events are in fact conditioned by one and the same true value of ξ. Thus the a posteriori distribution $P(\xi|x)$ assigned to ξ represents the limited knowledge about ξ.

There has been a long debate as to whether this procedure - Bayesian inference - is justified and is covered by the notion of probability. The key question was: can one consider probability not only as the relative frequency of events but also as a value of truth assigned to the statement, "ξ lies within the interval $\mathcal{I}$"? We take it for granted that the answer is "Yes", and consider the debate as historical.

For the founders of statistical inference, Bayes[1] [1] and Laplace[2] [2, 5], the notion of probability has carried both concepts: the probability attached to a statement[3] ξ can mean the relative frequency of its occurrence or the state of knowledge about ξ.

This "or" is not exclusive; it is not an "either or". It allows statements ξ that cannot be subjected to a "quality test," revealing how often they come true. Such a test is possible for the statement, "The probability that the coin falls with head upward is 1/2." However, the statement, "It is very probable that it rains tomorrow," is not amenable to the frequency interpretation, not because the qualitative value "very probable" is vague but because "tomorrow" always exists only once. So the

[1] Thomas Bayes, 1702–1761, English mathematician and Anglican clergyman. In a posthumously published treatise, he formulated for the first time a solution to the problem of statistical inference.

[2] Pierre Simon Marquis de Laplace, 1749–1827, French mathematician and physicist. He contributed to celestial and general mechanics. His work *Mécanique céleste* has been considered to rival Newton's *Principia*. He invented spherical harmonics and formulated and applied Bayes' theorem independently of him.

[3] We do not distinguish between the quantity ξ and a statement about the value of ξ.

latter statement can only be interpreted as evaluating the available knowledge. A fine description of these different interpretations has been given by Cox [6, 7]. See also Chaps. 13 and 14 of Howson and Urbach [8]. A taste of the above-mentioned debate is given by the polemics in [9].

We speak of probability in connection with statements that do not allow the frequency interpretation. However, we require a mathematical model that quantifies the probability attached to a statement.

We think that the distinction is only an apparent one, because the interpretation of probability as a value of the available knowledge cannot be avoided. This can be seen from the following examples.

Somebody buys a car. The salesman claims to be 95 % sure that the car will run the first 100 000 km without even minor trouble. This praise states his knowledge about or his belief in the quality of the product in question. However, there could be a statistical quality test which would turn this personal belief into a frequency of breakdown. But even if the praise by the salesman becomes objective in this sense, it becomes a personal belief for the interested client into the quality of his or her car, the one car that he or she decides to buy.

Let us try to translate this into the language of measurement. The quantity ξ was measured as $\alpha \pm \sigma$. Hence, with 68 % probability it is in that interval. Setting aside systematic errors, one can offer a frequency interpretation to this statement: if one were to repeat the measurement, say 100 times, the result should fall into the interval with the frequency of 68 %. This is right but does not well describe the situation. If one actually had 100 more measurements, one would reasonably use them to state one final result of considerably higher precision than the first one. How to do this is described in Chap. 2. The final result would again be a single one [11, 12].

There does not seem to be a clear distinction between the cases which allow the frequency interpretation of probability and the cases which allow only its interpretation as a value of knowledge. The latter one is the broader one, and we accept it here. But we insist on mathematically formulated distributions. Some of them are presented in Chaps. 4 and 5.

As a consequence, it may be practical but it is not necessary to distinguish the statistical from the systematic error of an experiment. The statistical error is the consequence of the finite amount of data and can in principle be demonstrated by repeating the experiment. The systematic error results from parameters that are not precisely known, although they are not fluctuating randomly. These two types of error correspond rather well to the above two interpretations of probability. Accepting them both as possible interpretations of a unique concept, one can combine both errors into a single one. This is indeed done for the graphical representation of a result or its use in related experiments; see Sect. 4.2.1 of the *Review of Particle Physics* [13].

1.2 Logic

Because probability can be interpreted as the value of the available knowledge, it can also be considered the implementation of non-Aristotelian logic into scientific communication [14]. Jaynes has simply termed it the logic of science [15, 16]. It even serves everyday communication as certain weather forecasts show. In philosophy, it is called the logic of temporal statements [17]: "temporal" because the value of truth estimates the future confirmation of the statement. Without this relation to time, a statement must be either true or false.

The probability attached to the statement ξ can be considered the value of truth assigned to ξ. The continuum of probability values is then a manifold of values of truth situated between "false" and "true". This introduces the *tertium* which is excluded in Aristotelian logic by the principle *tertium non datur*, which says that a third qualification - other than "true" and "false" - is not available.

Logical operations in a situation where other qualifications are available must be done in such a way that they are consistent with Aristotelian logic in the following sense: probabilities must be quantified. The calculus of probability must be part of mathematics. Mathematics is based on Aristotelian logic. The rules of mathematical logic can be laid down in terms of symbolic logic. Therefore the rules of handling probabilities (i.e. continuous values of truth) must be consistent with symbolic logic.

From this consideration, there follow certain conditions which must be observed when values of truth are assigned to statements like, "from ξ follows x". This value of truth is the conditional probability $p(x|\xi)$, that is, the probability to find x when ξ is given. Cox [6] showed in 1946 that consistency between non-Aristotelian and mathematical logic requires the following two rules.

1. Let ξ and x be statements that possibly imply each other. The values of truth μ, p, and w of the statements "ξ holds", "x follows from ξ", and "$x \wedge \xi$ holds", respectively, must be defined such that the product rule

$$w(x \wedge \xi) = p(x \mid \xi)\, \mu(\xi) \tag{1.1}$$

holds. Here, the operator $\wedge$ means the logical "and". This relation implies Bayes' theorem. Cox's result is not simply a consequence of the principle that the probability to obtain $x \wedge \xi$ is the product of the probability to obtain x and the probability to obtain ξ. This "principle" is obvious only if x and ξ are statistically independent variables. But this is not the case because $p(x \mid \xi)$ states that the distribution of x is conditioned by ξ. Cox's result implies the assumption that there is a convincing definition of the prior distribution $\mu(\xi)$, independent of x. The present book shows that there is such a definition provided that x is conditioned by ξ via a symmetry property of p which we call form invariance.

2. Conditional distributions - such as $p(x|\xi)$ and $P(\xi|x)$ - must be proper and normalised so that

$$\int d\xi\, P(\xi|x) = 1\,. \tag{1.2}$$

Here, the integral without indication of the limits of integration extends over the entire domain of definition of ξ. This rule is necessary in order to assign a probability to a negation. The probability of the assertion "ξ is not in the interval $[\xi_<, \xi_>]$" is the integral over the complement of the interval $[\xi_<, \xi_>]$. The integral over the complement exists because P is required to be proper. The assignment of unit probability to the statement that ξ is somewhere in its domain of definition, is a convention. Not only the posterior P but also conditional distributions must be normalised. Equation (1.2) holds analogously for the model $p(x|\xi)$. Without this requirement, the dependence of $p(x|\xi)$ on the parameter ξ would not be clearly defined. One could multiply it with any nonnegative function of ξ without changing the distribution of x. Hence, inferring ξ from the event x is possible only if (1.2) holds. Nevertheless, in the present book, distributions are admitted that cannot be normalised, provided that they are unconditional. Such distributions are called improper. One cannot assign a value of truth to a negation on a quantity with an improper distribution. Even in that case, however, one can assign a value of truth to a statement that contains the logical "and". We return to this in Sect. 2.1.

The joint distribution of the multiple event $x_1 \wedge x_2 \wedge \cdots \wedge x_N$ is discussed often. The interested reader should derive it from the logical rule (1.1) under the assumption that x_k follows the distribution $p(x_k|\xi)$ (The solution can be found in Sect. A.1). The logical operator $\wedge$ is never written in what follows. Instead, the multiple event is called $x_1, \dots, x_N$ or simply $x = (x_1, \dots, x_N)$.

An immediate consequence of the rules (1.1, 1.2) is Bayes' theorem discussed in Chap. 2. This theorem specifies the posterior probability P of x, given ξ. By the same token, the error interval of ξ is given. It is the smallest interval in which ξ lies with probability K. We call it the Bayesian interval $\mathcal{B}(K)$. To find the smallest interval, two things must exist: a measure in the space of ξ and a "most likely" value of ξ. A measure allows assigning a length or a volume to a subset of the domain of definition of ξ. The measure is identified with the "prior distribution" μ appearing in Eq. (1.1).

1.3 Ignorance

Into the definition of $P(\xi|x)$ enters a distribution $\mu(\xi)$ which is independent of the event x. This distribution can be interpreted as a description of ignorance about ξ, and is called the a priori distribution. All methods of inference described in the present book rely on Bayes' theorem and a definition of μ.

The definition starts from a symmetry principle. In Chaps. 6, 8, and 11, models $p(x|\xi)$ are considered that connect the parameter ξ with the event x by way of a group of transformations. This symmetry is called form invariance. The invariant measure of the symmetry group, which we explain in Chap. 6, is the prior distribution μ. This procedure is inspired by the ideas of Hartigan [18], Stein [19], and Jaynes [21].

The invariant measure is not necessarily a proper distribution; see Sect. 2.5. It can be obtained, without any analysis of the group, as a functional of the model p. The functional is known as Jeffreys' rule [22]. Here, it is introduced in Chap. 9.

By accepting the interpretation of probability as a value of truth, we include the "subjective" or "personal" interpretations presented in [23–27]. However, we do not go so far as to leave the prior distribution at the disposal of the person or the community analyzing given data. This is done in Chap. 14 of Howson and Urbach [8] and in the work by D'Agostini [28–30]. Instead, we adhere to a formal general definition of the prior distribution in order to avoid arbitrariness.

Form-invariant distributions offer more than a plausible definition of the prior distribution. Form invariance helps to clarify the dependence of parameters on each other. This allows us to devise a scheme where one parameter ξ_1 is inferred independently of the other parameters $\xi_2, \ldots, \xi_N$ in the sense that ξ_1 refers to an aspect of the event x that is separate from the aspects described by the other parameters; see Chap. 12. This scheme is useful because the extraction of ξ_1 is often linked to and dependent on other parameters that must be included in the model even though they are not interesting. A Gaussian model involves two parameters: its central value and its variance. Often the interest is focused on the central value only.

Form invariance is usually considered to occur so rarely that one cannot found the definition of the prior distribution on it. See Sect. 6.9 of [31] and [32]. Chapter 11 of the present book shows that there are more form-invariant distributions than were previously believed.

Still, Bayesian inference cannot be restricted to form-invariant distributions. When this symmetry is lacking, one considers the square root of the probability $p(x|\xi)$ - that is, the amplitude a_x - as a component of a vector that depends parametrically on ξ. This is the parametric representation of a surface in a vector space. The measure on the surface is the prior distribution. To understand this, one needs some differential geometry [33–35], which is explained in Chap. 9. The differential geometrical measure is again given by Jeffreys' rule [22].

Differential geometry by itself cannot establish Jeffreys' rule as the generally valid measure. One must show that the surface $a(\xi)$ is to be considered in the space of the amplitudes, not of the probabilities or of a function other than the square root of the probabilities. This, however, is indicated by the form-invariant models.

Beyond the observed event x, information on ξ is often available that should be incorporated into Bayesian inference and that will let the Bayesian interval shrink. The order of magnitude of ξ is usually known. A fly is neither as small as a microbe nor as large as an elephant. One knows this before measuring a fly. Such information can be built into the prior distribution which thereby changes from the ignorance prior μ to an informed prior μ^{inf}. An informed prior may simply be the posterior of a preceding experiment. It may also be generated by entropy maximisation, given previous information. Jaynes [36, 37] has transferred this method from thermodynamics into the analysis of data. This idea has found much interest and has led to a series of conferences [38–51] and many publications [52]. We take this method as well known and do not treat it in the present book. Note, however, that entropy maximisation cannot replace the definition of the ignorance prior μ. According to Jaynes [21], the method uses μ.

1.4 Decisions

Bayesian inference chooses from the family of distributions $p(x|\xi)$ the one that best reproduces the observed event x. This does not mean that any one of the distributions is satisfactory. How does one decide whether the model $p(x|\xi)$ is satisfactory in the sense that it contains a distribution consistent with the available data?

When x follows a Gaussian distribution, this is decided by the chi-squared criterion described in Chap. 13. It turns out that to make the decision, one needs a measure in the space of ξ. Again we identify this measure with the prior distribution μ.

Hence, the definition of a measure is essential for practically all conclusions from statistical data. One needs a measure - the prior distribution - in order to infer a parameter and to construct an error interval; see Chap. 2. One needs a measure in order to decide whether the given value of a parameter is probable or rather improbable; see Chap. 3. One needs a measure in order to decide whether a given set of events is compatible with a predicted distribution; see Chap. 13.

References

1. T. Bayes, An essay towards solving a problem in the doctrine of chances. Phil. Trans. Roy. Soc. **53**, 330–418 (1763). Reprinted in Biometrika 45, 293–315 (1958) and in ‘Studies in the History of Statistics and Probability’ E.S. Pearson and M.G. Kendall eds., C.Griffin and Co. Ltd., London 1970 and in ’Two papers by Bayes with commentaries’ W.E. Deming ed., Hafner Publishing Co., New York, 1963
2. P.S. de Laplace, Mémoire sur la probabilité des causes par les événements. Mém. de math. et phys. présentés à l’Acad. roy. des sci. **6**, 621–656 (1774). Reprinted in [3], vol. 8, pp. 27–65. An English translation can be found in [4]
3. P.S. de Laplace, *Œuvres complè tes de Laplace* (Gauthier-Villars, Paris, 1886–1912). 14 volumes
4. S.M. Stigler, Laplace’s 1774 memoir on inverse probability. Statist. Sci. **1**, 359–378 (1986)
5. P.S. de Laplace. *A Philosophical Essay on Probabilities* (Dover, New York, 1951). Original title Essay philosophique sur les probabilités
6. R.T. Cox, Probability, frequency and reasonable expectation. Am. J. Phys. **14**, 1–13 (1946)
7. R.T. Cox, *The Algebra of Probable Inference* (Johns Hopkins Press, Baltimore, 1961)
8. C. Howson, P. Urbach, *Scientific Reasoning: The Bayesian Approach*, 2nd edn. (Open Court, Chicago, 1993)
9. E.T. Jaynes, Confidence intervals vs. Bayesian intervals, in *Papers on Probability, Statistics and Statistical Physics*, Synthese Library, ed. by R.D. Rosenkrantz (Reidel, Dordrecht, 1983), pp. 149–209. The original article is in [10]
10. W.L. Harper, C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (Reidel, Dordrecht, 1976)
11. E. Schrödinger, The foundation of the theory of probability I. Proc. R. Irish Acad. **51A**, 51–66 (1947). Reprinted in Gesammelte Abhandlungen, Vienna 1984, vol. 1, pp. 463–478
12. E. Schrödinger, The foundation of the theory of probability II. Proc. R. Irish Acad. **51A**, 141–146 (1947). Reprinted in Gesammelte Abhandlungen, Vienna 1984, vol. 1, pp. 479–484
13. Particle Data Group, The review of particle physics. Eur. Phys. J. C **15**, 1–878 (2000)
14. H. Reichenbach, *Wahrscheinlichkeitslehre* (Vieweg, Braunschweig, 1994)
15. E.T. Jaynes, Probability theory as logic, in Fougère [44], pp. 1–16

16. E.T. Jaynes. Probability theory: The logic of science. Unfinished book, see http://omega.albany.edu:8008/JaynesBook.html
17. C.F. von Weizsäcker, *Die Einheit der Natur* (Hanser, Munich, 1971). See page 241
18. J. Hartigan, Invariant prior distributions. Ann. Math. Statist. **35**, 836–845 (1964)
19. C.M. Stein, Approximation of improper prior measures by proper probability measures, in Neyman and Le Cam [20], pp. 217–240
20. J. Neyman, L.M. Le Cam (eds.), *Bernoulli, Bayes, (Laplace. Proceedings of an International Research Seminar. Statistical Laboratory* (Springer, New York, 1965)
21. E.T. Jaynes, Prior probabilities. IEEE Trans. Syst. Sci. Cybern. **SSC–4**(3), 227–241 (1968)
22. H. Jeffreys, *Theory of Probability* (Oxford University Press, Oxford, 1939). 2nd edition 1948; 3rd edition 1961, here Jeffreys' rule is found in III §3.10
23. J.L. Savage, *The Foundations of Statistics* (Wiley, New York, 1954). There is a second edition by Dover, New York 1972
24. J.L. Savage, Elicitation of personal probabilities and expectations. J. Am. Statist. Assoc. **66**, 783–801 (1971)
25. D.V. Lindley, A statistical paradox. Biometrika **44**, 187–192 (1957)
26. Bruno de Finetti, *Probability, Induction and Statistics: The Art of Guessing* (Wiley, New York, 1972)
27. B. de Finetti, *Theory of Probability* (Wiley, London, 1974)
28. G. D'Agostini, Probability and measurement uncertainty in physics — a Bayesian primer (1995). Lecture Notes, preprint DESY 95-242, arXiv:hep-ph/9512295v2
29. Giulio D'Agostini. Jeffreys priors versus experienced physicist priors. Arguments against objective Bayesian theory. arXiv:physics/9811045 and http://www-zeus.roma1.infn.it/~agostini/
30. G. D'Agostini, Bayesian reasoning versus conventional statistics in high energy physics, in von der Linden et al. [51], pp. 157–170
31. O. James, *Berger Statistical Decision Theory and Bayesian Analysis*, 2nd edn., Springer Series in Statistics (Springer, New York, 1985)
32. C.P. Robert, *The Bayesian Choice*, 2nd edn. (Springer, Berlin, 2001)
33. S. Amari, *Differential Geometrical Methods in Statistics*, vol. 28, Lecture Notes in Statistics (Springer, Heidelberg, 1985)
34. C.C. Rodriguez, Objective Bayesianism and geometry, in Fougère [44], pp. 31–39
35. C.C. Rodriguez, Are we cruising a hypothesis space? in von der Linden et al. [51], pp. 131–139
36. E.T. Jaynes, Information theory and statistical mechanics I. Phys. Rev. **106**, 620–630 (1957)
37. E.T. Jaynes, Information theory and statistical mechanics II. Phys. Rev. **108**, 171–190 (1957)
38. C.R. Smith, W.T. Grandy Jr., (eds.), *Maximum-Entropy and Bayesian Methods in Inverse Problems [Proceedings of Two Workshops, Laramie 1981 and 1982]* (Reidel, Dordrecht, 1985)
39. C.R. Smith, G.J. Erickson (eds.), *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems, Laramie, 1983* (Kluwer, Dordrecht, 1987)
40. J.H. Justice (ed.), *Maximum Entropy and Bayesian Methods in Applied Statistics, Calgary, 1984* (Cambridge University Press, Cambridge, 1986)
41. G.J. Erickson, C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods in Science and Engineering, Laramie, 1987*, vol. 1 (Kluwer, Dordrecht, 1988). Foundations
42. G.J. Erickson, C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods in Science and Engineering, Seattle, Wyoming 1987*, vol. 2. Applications (Kluwer, Dordrecht, 1988)
43. J. Skilling (ed.), *Maximum Entropy and Bayesian Methods, Cambridge, 1988* (Kluwer, Dordrecht, 1989)
44. P.F. Fougère (ed.), *Maximum Entropy and Bayesian Methods, Dartmouth, 1989* (Kluwer, Dordrecht, 1990)
45. W.T. Grandy Jr., L.H. Schick (eds.), *Maximum Entropy and Bayesian Methods, Laramie, 1990* (Kluwer, Dordrecht, 1991)
46. C.R. Smith, G.J. Erickson, P.O. Neudorfer (eds.), *Maximum Entropy and Bayesian Methods, Seattle 1991* (Kluwer, Dordrecht, 1992)

47. A. Mohammad-Djafari, G. Demoment (eds.), *Maximum Entropy and Bayesian Methods, Paris 1992* (Kluwer, Dordrecht, 1993)
48. G.R. Heidbreder (ed.), *Maximum Entropy and Bayesian Methods, Santa Barbara 1993* (Kluwer, Dordrecht, 1996)
49. J. Skilling, S. Sibisi (eds.), *Maximum Entropy and Bayesian Methods, Cambridge, 1994* (Kluwer, Dordrecht, 1996)
50. K.M. Hanson, R.N. Silver (eds.), *Maximum Entropy and Bayesian Methods, Santa Fe, 1995* (Kluwer, Dordrecht, 1996)
51. W. von der Linden, V. Dose, R. Fischer, R. Preuss (eds.), *Maximum Entropy and Bayesian Methods, Garching, 1998* (Kluwer, Dordrecht, 1999)
52. B. Buck, V.A. Macaulay (eds.), *Maximum Entropy in Action. A Collection of Expository Essays* (Clarendon Press, Oxford, 1991)

Chapter 2
Bayes' Theorem

In Sect. 2.1, Bayes' theorem is derived. The prior distribution that it contains must be defined so that it transforms as a density. Transformations of densities and functions are discussed in Sect. 2.2. A symmetry argument can define the prior. This is described in Sects. 2.3 and 2.4. Prior distributions are not necessarily proper. In Sect. 2.5, we comment on improper distributions because it is unusual to admit any of them. In Sect. 2.6, we discuss the information conveyed by a proper distribution. Several problems are suggested to be solved by the reader. Their solutions are given in Sect. A.2.

2.1 Derivation of the Theorem

The logical connection "x and ξ" is equivalent to "ξ and x". Therefore, the distribution w of (1.1) can also be factorised in the form

$$w(x \wedge \xi) = P(\xi \mid x)\, m(x)\,. \tag{2.1}$$

This relation does not mean that, for a given w, the factorisation must be obvious. This relation means that whenever two of the three distributions w, P, m are defined, the third one is given by this equation. Combining (1.1) with (2.1) yields

$$p(x|\xi)\, \mu(\xi) = P(\xi|x)\, m(x)\,. \tag{2.2}$$

Both of the statements x and ξ refer to numerical variables, so that the precise form of the statements is, "The event has the coordinate x" and "The hypothesis has the coordinate ξ". Nobody really speaks that way. One simplifies to speak of the event x and the hypothesis ξ.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_2

In Chap. 1 we have given reasons for ascribing probabilities to both an event that statistically fluctuates and the hypothesis that is unknown. Indeed, Sect. 2.2 combines distributions of x and ξ. For practical reasons - not for reasons of logic - we introduce a notational difference. Events are always denoted by Latin letters and hypothesis parameters by Greek letters. We write $p(x|\xi)$ for a distribution of x, conditioned by ξ. This distribution is also called the "statistical model" or simply the "model". The unconditioned distribution μ is called the prior distribution of the hypothesis. We write capitals, especially P, for the distribution of the parameter conditioned by the event. It is called the posterior. The distributions of ξ are derived in the present book.

Conditional distributions must be proper. For every ξ one has

$$\int dx\, p(x|\xi) = 1\,, \tag{2.3}$$

and for every x, one requires

$$\int d\xi\, P(\xi|x) = 1\,. \tag{2.4}$$

Integrals that do not show the limits of integration, extend over the full range of definition of the integration variable. The integral over x is to be read as a sum if x is a discrete variable. The hypothesis shall always be continuous.

By use of the normalisation Eq. (2.4), one obtains

$$m(x) = \int d\xi\, p(x \mid \xi)\mu(\xi) \tag{2.5}$$

from (2.2). We require the existence of this integral. The posterior distribution then is

$$P(\xi|x) = \frac{p(x|\xi)\mu(\xi)}{\int d\xi'\, p(x|\xi')\mu(\xi')}\,. \tag{2.6}$$

This relation is Bayes' theorem. It relates the distribution of x, given ξ, to the distribution of ξ, given x. Thus it allows one to infer the parameter ξ from the observation x provided the prior μ is defined and the integral (2.5) exists.

Bayes' theorem suggests that one can attribute μ to the parameter ξ "before" any observation x is available. This is the reason for the name. The prior describes ignorance about ξ. Laplace [1, 4, 5] set $\mu(\xi) \equiv \text{const}$. Bayes, hesitatingly, made the same ansatz, knowing that he had no sufficient argument. We show that ignorance cannot be represented in an absolute way. Its representation depends on the context. The context is given by the model p. In Chaps. 6, 9, and 11, we show that it allows one to define μ. The existence of the integral (2.5) is a consequence of the form invariance of p; see Chap. 6.

The requirements (2.3) and (2.4) entail that $w(x \wedge \xi)$ can be integrated over x or ξ. However, w may be improper, that is, cannot be integrated over x and ξ, because we admit μ to be improper. The prior may, for example, be constant all over the real axis.

The form (2.6) of Bayes' theorem does not depend on the mathematical dimension of x or ξ. They may be one-dimensional or multidimensional variables. The event x may be continuous or discrete. We assume the parameter ξ to be continuous.

The rule (2.1) cannot be used to factorise every given distribution of a multidimensional variable [6, 7]. Let $\xi = (\xi_1, \xi_2)$ be a two-dimensional parameter. The rule (2.1) could lead one to factorise $P(\xi|x)$ into a distribution of ξ_1 conditioned by ξ_2, x, and an unconditioned distribution of ξ_1. However, the condition specifies the implication: given x, the distribution of ξ follows. If ξ_2 were a condition for ξ_1 then the value of ξ_2 would, together with x, imply the distribution of x_1. Yet a given value of ξ_2 does not imply the distribution of ξ_1. We cannot separate Eq. (2.1) from our starting point (1.1) which specifies that ξ_1 and ξ_2 imply the distribution of the events; they do not imply each other.

2.2 Transformations

The prior μ cannot be universally defined as $\mu \equiv$ const because the uniform distribution is not invariant under reparameterisations. By a reparameterisation, we mean the transition from ξ to another variable η through a transformation T that is,

$$\eta = T\xi \ . \tag{2.7}$$

A transformation is an invertible mapping. Inasmuch as ξ is continuous, $\mu(\xi)$ is a density. The transformation to $\mu_T(\eta)$ is made such that probabilities remain the same, not densities. This means

$$\mu_T(\eta)\mathrm{d}\eta = \mu(\xi)d\xi \tag{2.8}$$

or

$$\mu_T(\eta) = \mu(\xi)\left|\frac{d\xi}{d\eta}\right| \ . \tag{2.9}$$

The absolute value $|\dots|$ appears because the probability densities are never negative. If ξ is a multidimensional variable, the derivative in (2.9) must be replaced by the Jacobian determinant

$$\left|\frac{\partial\xi}{\partial\eta}\right| \equiv \left|\det\frac{\partial\xi}{\partial\eta}\right| \ . \tag{2.10}$$

Of course, Eq. (2.8) can be interpreted not only as (2.9) but also as

$$\mu(\xi) = \mu_T(\eta)\left|\frac{\partial\eta}{\partial\xi}\right| \ . \tag{2.11}$$

The Eqs. (2.9) and (2.11) are consistent with each other because

$$\det \frac{\partial \xi}{\partial \eta} = \left(\det \frac{\partial \eta}{\partial \xi} \right)^{-1} . \tag{2.12}$$

Equation (2.9) means that a transformation changes the uniform distribution into a nonuniform one. Transforming, for example, the real variable ξ to $\eta = \xi^2$, one obtains $\mu_T(\eta) \propto \eta^{-1/2}$. As another example, consider $\mu(\xi_1, \xi_2) \equiv \text{const}$, depending on the Cartesian coordinates ξ_1, ξ_2. If μ is transformed to polar coordinates where the radius is $\eta = (\xi_1^2 + \xi_2^2)^{1/2}$ and the angle is ϕ, one obtains $\mu_T(\eta, \phi) \propto \eta$. Hence, one cannot universally represent ignorance by the uniform distribution.

In contrast to a density, a function $f(\xi)$ transforms as

$$f_T(\eta) = f(\xi) . \tag{2.13}$$

Here, the values of f and f_T at corresponding points ξ and η are equal. Therefore, the constant function is invariant under all transformations; the uniform density is not.

Bayes' theorem transforms properly under reparameterisations of ξ or x. The interested reader should convince himself or herself of this fact.

If ξ is one-dimensional, one can always find a transformation such that the prior becomes uniform. This transformation is

$$\eta = \int^{\xi} d\xi' \, \mu(\xi') \tag{2.14}$$

for any lower limit of the integration. The proof is left to the reader.

If ξ is multidimensional it is not always possible to transform ξ such that the prior becomes uniform.

Bayesian inference was forgotten or even fell into disrepute in the century that followed Laplace. The distribution of ignorance remained undefined; see the descriptions of history by [8, 9, 11–13]. In the twentieth century, Bayes' theorem was rediscovered [14–21]. The original article was reprinted [22], a series of conferences on its application have taken place [23–29], and current textbooks on statistics mention Bayes' theorem [30].

The definition of the prior, however, remained controversial. One attempt to solve the dilemma is to declare that - for continuous ξ - the prior cannot be objectively defined. It then becomes an unavoidably subjective element in the interpretation of data. Although certain logical rules have to be respected [31–36], it is up to the experienced researcher to make an ansatz for μ within a framework of historically grown conventions.

Scientific judgments are possibly conditioned by their cultural context. We think that this is not relevant for the Bayesian prior distribution. Otherwise one could, strictly speaking, draw any conclusion from a given event. We adhere to a formal definition [16] of μ.

2.3 The Concept of Form Invariance

In order to get an idea of the theory developed in Chaps. 6, 8, and 11, let us consider a model that depends on the difference between x and ξ only, that is,

$$p(x|\xi) = w(x - \xi)\,, \tag{2.15}$$

such as the Gaussian

$$p(x|\xi) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left(- \frac{(x - \xi)^2}{2\sigma^2} \right) \tag{2.16}$$

centered at ξ. The quantity σ (called the standard deviation) is considered to be given and not to be inferred. For this reason, it is not listed among the hypotheses of the model p.

The distribution centred at $\xi = 0$ is given in Fig. 2.1. The standard deviation σ characterises its width. It is not the width at half maximum; the value of 2σ is the width of the central interval that contains 68 % of the events. This means that the shaded area in the figure has the size of 0.68. The distribution (2.16) is normalised to unity.

From (2.15), x and ξ have the same physical dimension (as, e.g., a length), and the event x seems a good estimate for the parameter ξ. We can say more than that: such a model suggests that the distribution of ξ is centred at x and has the same form as the distribution of x. This amounts to the surmise

$$P(\xi|x) = p(x|\xi) \tag{2.17}$$

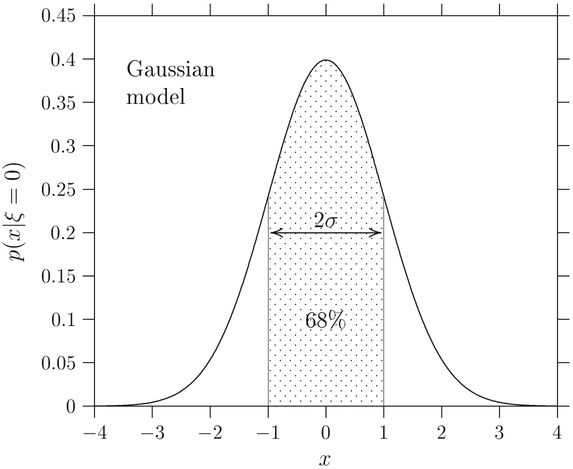


Fig. 2.1 The Gaussian distribution (2.16) centred at $\xi = 0$. The standard deviation σ has been chosen equal to unity

which holds if

$$\mu(\xi) \equiv \text{const}\,. \tag{2.18}$$

Indeed, this is true by the theory of Chap. 6. The idea is as follows. The model (2.16) depends on the difference $x-\xi$. Thus x and ξ have the same physical dimension; they are defined on the same scale. This requires that a given difference $x-\xi$ must mean an interval of one and the same length whatever the value of ξ. Thus the measure

$$\mu = \frac{\text{distance}}{\text{difference of parameters}}$$

must be uniform. Now the model (2.16) connects x and ξ via a translation. The parameter ξ labels a translation of the possible values of x. Translating x and ξ by the same value leaves (2.16) unchanged. This defines a symmetry between x and ξ because the set of all possible translations is a mathematical group. The invariant measure μ of the group converts a difference of parameters into a distance. The invariant measure of the group of translations is constant. The only density that is invariant under the translations is the uniform density (see Chaps. 6 and 7). We call the symmetry between event and parameter "form invariance". This notion is taken from the present example. The form of the distribution of x remains the same for all ξ. For more details see Chap. 6.

Let ξ be transformed to $T\xi = \eta$ in the model p of (2.15). Translational symmetry continues to exist with respect to ξ, not with respect to η (unless T is itself a translation). Thus the symmetry picks up the parameterisation in which the prior is uniform.

We do not restrict ourselves to models that are form invariant. This symmetry, however, provides the paradigmatic cases from which one can learn about the tools and the phenomena in statistical inference.

The models (2.15) yield the simplest possible application of Bayes' theorem. Equation (2.17) together with (2.16) leads to the well-known Gaussian error intervals; compare Sect. 3.2. Therefore, this example shows that the arguments of the present book do not contradict, but instead generalise Gaussian error estimation. This is also borne out by the fact that the error deduced from N observations is $\propto N^{-1/2}$, as we calculate now.

2.4 Many Events

Instead of only one, we now consider N observations $(x_1, \ldots, x_N)$. All x_k are drawn from the same distribution $p(x|\xi)$. They are conditioned by the same value of ξ, but they are statistically independent. This means that their joint distribution is the product

$$p(x_1, \ldots, x_N|\xi) = \prod_{k=1}^{N} p(x_k|\xi) \tag{2.19}$$

of the distributions of the x_k (see Problem A.1.1). One can also say that all the x_k have been drawn from the same ensemble. The prior should not depend on the number N of events. This is indeed the case as we prove in Chaps. 6 and 9. Bayesian inference from N events then yields

$$P(\xi|x_1, ..., x_N) = \frac{\mu(\xi) \prod_{k=1}^{N} p(x_k|\xi)}{\int d\xi' \, \mu(\xi') \prod_{l=1}^{N} p(x_k|\xi')} \,. \tag{2.20}$$

One can interpret this formula as an "iteration" of Bayes' theorem. The posterior P can be obtained by introducing the distribution $P_{N-1}(\xi|x_1, \dots, x_{N-1})$, obtained from $N-1$ events, as an "informed prior" into the analysis of the Nth event. The reader is asked to show this in detail.

Let us use the Gaussian model (2.20) to exemplify Bayesian inference from N events. In order to write down (2.19), we introduce the notation $\langle \dots \rangle$ for the average

$$\langle f(x) \rangle = N^{-1} \sum_{k=1}^{N} f(x_k) \tag{2.21}$$

of a function $f(x_k)$. This means, in particular, that

$$\langle x \rangle = N^{-1} \sum_{k=1}^{N} x_k \,. \tag{2.22}$$

With this notation, one finds

$$\begin{aligned} p(x_1, ..., x_N \mid \xi) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{N}{2\sigma^2} (\langle x \rangle - \xi)^2\right) \\ \times \exp\left(-\frac{N}{2\sigma^2} \left[\langle x^2 \rangle - \langle x \rangle^2\right]\right) \,. \end{aligned} \tag{2.23}$$

The reader should verify this result. With the prior distribution (2.18), one obtains the posterior

$$P(\xi \mid x_1, ..., x_N) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{N}{2\sigma^2} (\xi - \langle x \rangle)^2\right) \,, \tag{2.24}$$

a Gaussian centred at the average $\langle x \rangle$ of the events. Thus the multidimensional event $(x_1, \dots x_N)$ leads to a posterior of the same symmetry and very much the same form as the one-dimensional event, provided that one replaces x by $\langle x \rangle$. This is a consequence of the form invariance of the model (2.16). The quantity $\langle x \rangle$ is called

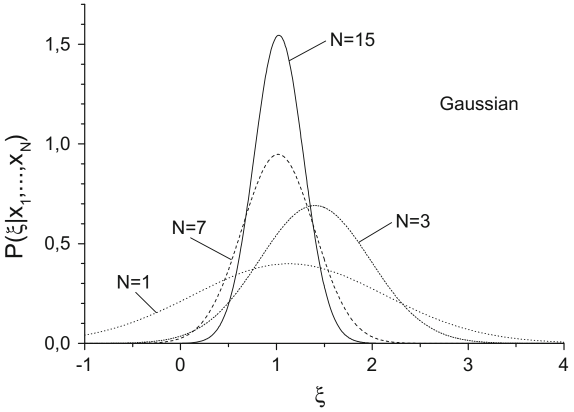


Fig. 2.2 Posterior distribution of the Gaussian model (2.16) with N observations. See text

the estimator of ξ. Its general definition is given in Sect. 3.3. It differs, in general, from the average over the events.

The posterior (2.24) has the standard deviation $N^{-1/2}\sigma$. This is the usual Gaussian error estimate for N observations. It yields the error interval

$$\xi = \langle x \rangle \pm N^{-1/2}\sigma \,. \tag{2.25}$$

This is the Bayesian interval in which ξ is found with the probability of 68 % (see Chap. 3).

In Fig. 2.2, the parameter ξ is inferred from an increasing number N of events. The model (2.16) is used with $\sigma = 1$. We assume that the experimenter knows the precision of her measuring device. The events have been drawn with the help of a random generator with the true value of ξ set to unity. In reality, of course, one does not know the true value; one infers it. Note that the posterior distributions narrow down with increasing N but also jump back and forth. The process of approaching the true value is itself random.

In the next three figures, the exponential model

$$p(t|\tau) = \tau^{-1} \exp(-t/\tau) \tag{2.26}$$

is used to show qualitatively that the posterior approaches a Gaussian function for large N even if the model is not Gaussian. The exponential model is described in Sect. 4.2. The prior is

$$\mu(\tau) \propto \tau^{-1} \tag{2.27}$$

as we show in Sects. 3.2 and 7.2. By use of a random number generator, the events t were drawn [37] from the exponential distribution with $\tau = 3$ given in Fig. 2.3. The posteriors for $N = 3$ and $N = 10$ events are represented in Fig. 2.4. The curves are not symmetric under reflection. For N=50,100,300, the posteriors in Fig. 2.5 become more and more symmetric and, in fact, approximately Gaussian.

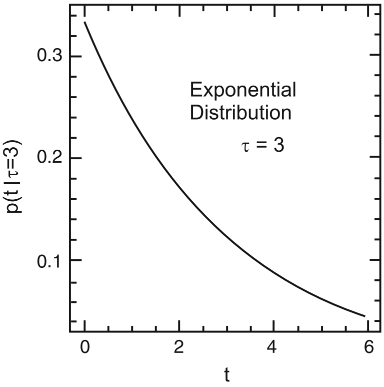


Fig. 2.3 The exponential distribution (2.26) with the parameter $\tau = 3$. See text

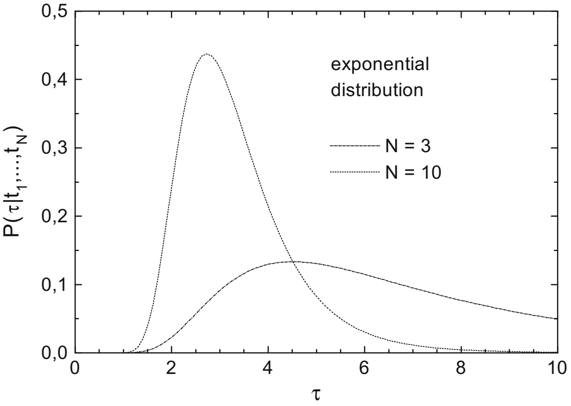


Fig. 2.4 Posteriors of the exponential model from N observations. See text

If one reparameterises τ such that the prior becomes uniform (i.e. according to Eq. (2.14)) the Gaussian approximation becomes valid at smaller values of N.

The model p should be such that the posterior approaches a Gaussian for large N. As a counterexample, we discuss an ill-defined model. Consider the rectangular distribution

$$p(x|\xi) = \begin{cases} 0 \text{ for } & x - \xi < -1/2 \\ 1 \text{ for } & -1/2 < x - \xi < 1/2 \\ 0 \text{ for } & x - \xi > 1/2\,. \end{cases} \tag{2.28}$$

We have set the true value of ξ equal to unity and let a random number generator draw $N = 8$ events from the distribution given in the upper part of Fig. 2.6. The events are marked as dots on the abscissa of the lower part. The model is of the type (2.15). Therefore the prior is uniform, and the posterior is

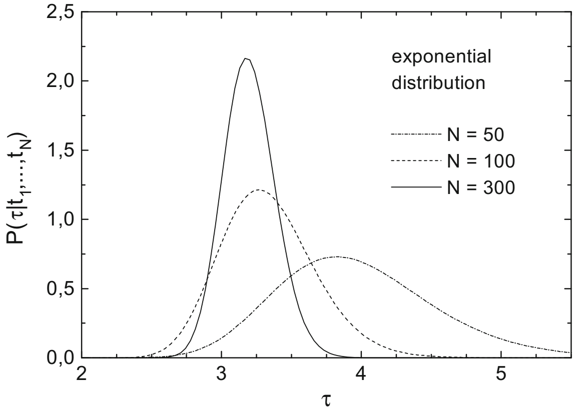


Fig. 2.5 Posteriors of the exponential model from *N* observations. See text

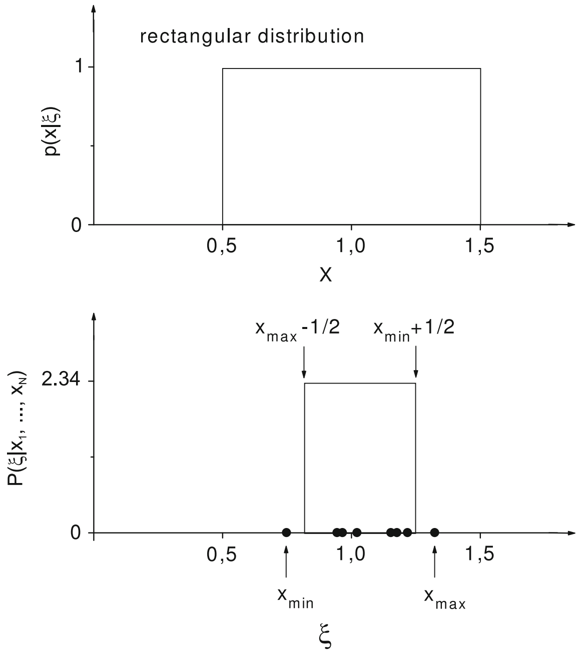


Fig. 2.6 The rectangular distribution and its posterior. See text

$$P(\xi|x_1 \ldots x_N) = \begin{cases} 0 & \text{for} \quad \xi < x_{max} - 1/2 \\ (1 - x_{max} + x_{min})^{-1} & \text{for } x_{max} - 1/2 < \xi < x_{min} + 1/2 \\ 0 & \text{for} \quad \xi > x_{min} + 1/2 \,. \end{cases} \tag{2.29}$$

Here, x_{min}, x_{max} are the lowest and highest events, respectively. The posterior from the eight events is displayed in the lower part of the figure. It is rectangular, and never approaches a Gaussian.

A reasonable feature of the posterior (2.29) is the fact that, as w increases, it will be more and more concentrated around its maximum. This is so because the interval $(x_{max} - 1/2, x_{min} + 1/2)$, where P differs from zero, shrinks. This happens because x_{max} and x_{min} will be found closer and closer to their limits 1/2 and 3/2, respectively. This concentration of the distribution can be interpreted and quantified via the Shannon information [38] conveyed by P; see Sect. 2.6 below.

An increase of information is expected. Nevertheless, the model (2.28) is unreasonable. *The* shortest interval containing ξ with a given probability K does not exist; that is, the Bayesian interval, introduced in Sect. 1.2, is not well defined. Instead, there is a manifold of intervals containing ξ with the probability K. We require the existence of the Bayesian interval. This is assured if the model $p(x|\xi)$ is a regular function of ξ and possesses a unique absolute maximum. The expression (2.28) has neither property.

Both the uniform prior of the Gaussian (2.16) and the prior (2.27) of the exponential model are improper. This deserves a discussion. In the following section, we comment on improper distributions.

2.5 Improper Distributions

In general, one requires probability distributions to be normalised. Every axiomatic foundation of probability theory asks for proper distributions. See Sect. 1.1 of the book by Lee [39] or Sect. 4.8 in Kendall's handbook [18]. However, we do not do so and admit improper distributions, where the integral

$$\int d\xi \mu(\xi)$$

does not exist. The central value ξ of the Gaussian model (2.16) is defined everywhere on the real axis. The prior is uniform and thus improper.

Nevertheless such distributions are useful. One can calculate the probability

$$w_1 = \int_{\mathcal{I}_1} d\xi \, \mu(\xi) \,, \tag{2.30}$$

that ξ is contained in the compact interval $\mathcal{I}_1$ and compare it with the probability that ξ is in another compact interval $\mathcal{I}_2$. However, one cannot assign a probability to the statement, "ξ is not in $\mathcal{I}_1$." This requires that the integral over the complement of $\mathcal{I}_1$ exists. If one wants to assign a probability to every logically possible statement, one must require the normalisation (2.3), (2.4) for every distribution.

In the derivation of Bayes' theorem, no negation is used. Therefore, no inconsistencies appear if improper priors are admitted. They are defined only up to an arbitrary factor. By requiring the normalisation of the posterior, this arbitrariness drops out

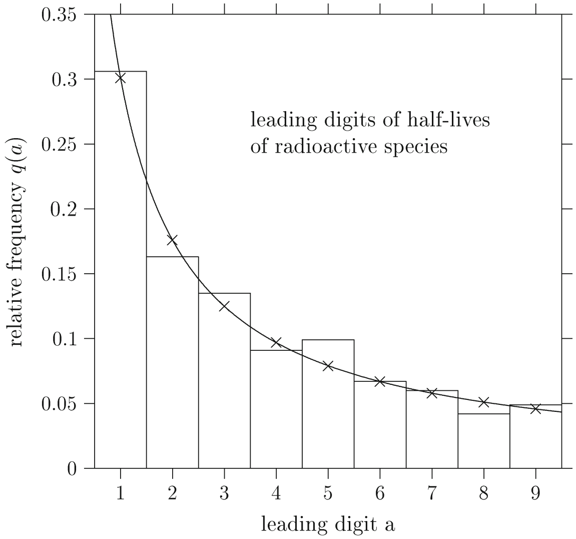


Fig. 2.7 The distribution of the leading significant digits of measured half-lives of radioactive species. The prediction (2.32) is shown by the *crosses*. The histogram represents the measured data

of Bayes' theorem. More generally, we require that all conditional distributions are proper.

The improper distribution

$$\mu(\xi) \propto \xi^{-1} \tag{2.31}$$

has caused some astonishment. It answers the question: "What is the distribution of all measured half-lives of radioactive nuclides?" This seems surprising [13, 40, 41]. However, (2.31) is the only possible answer, if the question is meaningful. The question does not say in which units the half-lives shall be considered: seconds, years? If we take this omission seriously, the answer must be invariant under any change of scale. We show in Chaps. 6 and 7 that the only distribution with this property is (2.31).

One can test (2.31) against the data by evaluating the distribution q of the leading digit a . From (2.31), it follows that

$$q(a) = \log \frac{a+1}{a} \ . \tag{2.32}$$

The reader should convince himself of this. In Fig. 2.7, this distribution is compared [13] with the leading digits of a set of 1203 measured half-lives. The agreement is impressive. Whether it really confirms (2.31) cannot be answered precisely because the distribution (2.32) contains no parameter to optimise the agreement with the data.

In the following section, we define Shannon's information mentioned above.

2.6 Shannon Information

The information conveyed by the statement, "x has the distribution $w(x)$," has been defined by Shannon [38] to be

$$S = \int \mathrm{d}x \, w(x) \ln w(x) \, . \tag{2.33}$$

This expression exists for every proper distribution $w(x)$. It is related to Boltzmann's definition of entropy.[1]

The expression S is most easily understood in the case where x takes discrete values x_k, and the integration in (2.33) amounts to a summation. The distribution w is normalised to unity. Then $w(x_k)$ is a nonnegative number between 0 and 1. For a given k, the product $w(x_k) \ln w(x_k)$ is minimal at $w(x_k) = 1/2$; it tends to zero for $w \to 0$ and for $w \to 1$. Therefore the expression (2.33) is usually negative. It becomes minimal if $w = 1/2$ everywhere. It increases if w comes closer to unity at some places and closer to zero at other ones. Hence, S measures the degree of concentration of the distribution of x. The maximum value of S is $S = 0$ which is reached when $w(x_k)$ equals unity for one k and zero for all other ones. Then the information is complete because one is sure about any coming event.

If x is a continuous variable then the information can never be complete. The distribution can be ever more concentrated, and the information can grow indefinitely. The variable ξ is continuous in the context of the present book. Thus the Shannon information on ξ can grow indefinitely.

In the next Chap. 3, we show that decisions require a measure and the existence of the "Bayesian interval" which serves as an error interval.

References

1. P.S. de Laplace, Mémoire sur la probabilité des causes par les événements. Mém. de math. et phys. présentés à l'Acad. roy. des Sci. **6**, 621–656, 1774. Reprinted in [2], vol. 8, pp. 27–65. An English translation can be found in [3]
2. P.S. de Laplace, *Œuvres complètes de Laplace* (Gauthier-Villars, Paris, 1886–1912). 14 volumes
3. S.M. Stigler, Laplace's 1774 memoir on inverse probability. Statist. Sci. **1**, 359–378 (1986)
4. P.S. de Laplace, *Théorie analytique des probabilités* (Courcier, Paris, 1812). Second edition 1814; third edition 1820. Reprinted as vol. 7 of [2]
5. P.S. de Laplace, *A Philosophical Essay on Probabilities* (Dover, New York, 1951). Original title Essay philosophique sur les probabilités

[1] Ludwig Boltzmann, 1844–1906, Austrian physicist. He is one of the founders of the kinetic theory of gases. He based thermodynamics on statistical mechanics. Here, we refer to Boltzmann's so-called H theorem. It says that the expression $H = \int \mathrm{d}^3 x p(x,t) \ln p(x,t)$ does not increase with increasing time t, if $p(x,t)$ is the distribution in space of the particles of a gas at the time t; see [42].

6. T. Podobnik, T. Živko, On consistent and calibrated inference about the parameters of sampling distributions, August 2005. arXiv:physics/0508017
7. T. Podobnik, T. Živko, On probabilistic parametric inference. J. Stat. Plan. Inference **142**, 3152–3166 (2012)
8. S. Stigler, *American Contributions to Mathematical Statistics in the Nineteenth Century* (Arno, New York, 1980). Two volumes
9. E.T. Jaynes, Bayesian methods: general background. in Justice [10], pp. 1–25
10. J.H. Justice (ed.), *Maximum Entropy and Bayesian Methods in Applied Statistics, Calgary, 1984* (Cambridge University Press, Cambridge, 1986)
11. S.M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900* (Harvard University Press, Cambridge, 1986)
12. C. Howson, P. Urbach, *Scientific Reasoning: The Bayesian Approach*, 2nd edn. (Open Court, Chicago, 1993)
13. S. Steenstrup, Experiments, prior probabilities, and experimental prior probabilities. Am. J. Phys. **52**, 1146–1147 (1984)
14. H. Jeffreys, *Theory of Probability* (Oxford University Press, Oxford, 1939). 2nd edition 1948; 3rd edition 1961, here Jeffreys' rule is found in III §3.10
15. J. Neyman, L.M. Le Cam (eds.), in *Bernoulli, Bayes (Laplace. Proceedings of an International Research Seminar. Statistical Laboratory)* (Springer, New York, 1965)
16. E.T. Jaynes, Prior probabilities. IEEE Trans. Syst. Sci. Cybern. **SSC–4**(3), 227–241 (1968)
17. P.W. Anderson, The Reverend Thomas Bayes, needles in haystacks, and the fifth force. Phys. Today **45**, 9–11 (1992)
18. A. O'Hagan, *Bayesian Inference*, vol. 2B, Kendall's Advanced Theory of Statistics (Arnold, London, 1994)
19. N.G. Polson, G.C. Tiao (eds.), *Bayesian Inference* (Cambridge University Press, Cambridge 1995). Two volumes
20. A. Zellner, *An Introduction to Bayesian Inference in Econometrics* (Wiley, New York, 1971). Reprinted in 1996
21. A. Zellner, A Bayesian era, in Bernardo et al. [24], pp. 509–516
22. T. Bayes, An essay towards solving a problem in the doctrine of chances. Phil. Trans. Roy. Soc. **53**, 330–418 (1763). Reprinted in Biometrika 45, 293–315 (1958) and in 'Studies in the History of Statistics and Probability' E.S. Pearson and M.G. Kendall eds., C.Griffin and Co. Ltd., London 1970 and in 'Two papers by Bayes with commentaries' W.E. Deming ed., Hafner Publishing Co., New York, 1963
23. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M Smith (eds.), in *Bayesian Statistics. Proceedings of the International Meeting held at Valencia on May 28–2 June 1979* (Valencia University Press, Valencia, 1980)
24. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M Smith (eds.), in *Bayesian Statistics 2. Proceedings of the Second International Meeting held at Valencia on September 6–10, 1983* (North-Holland, Amsterdam, 1985)
25. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M Smith (eds.), in *Bayesian Statistics 3. Proceedings of the Third International Meeting Held at Valencia on June 1–5, 1987* (Oxford University Press, Oxford, 1988)
26. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (eds.) in *Bayesian Statistics 4. Proceedings of the Fourth International Meeting held at Valencia on April 15–20, 1991* (Oxford University Press, Oxford, 1992)
27. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (eds.), in *Bayesian Statistics 5. Proceedings of the Fifth International Meeting held at Valencia on June 5–9, 1994* (Clarendon Press, Oxford, 1996)
28. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (eds.) in *Bayesian Statistics 6. Proceedings of the Sixth Valencia International Meeting June 6–10, 1998* (Clarendon Press, Oxford, 1999)
29. A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics: Essays in Honour of H. Jeffreys*, vol. 1, Studies in Bayesian Econometrics (North Holland, Amsterdam, 1980)

30. M. Wirtz, C. Nachtigall, *Wahrscheinlichkeitsrechnung und Inferenzstatistik*, 3rd edn. (Juventa, Weinheim, 2004), p. 84
31. R.T. Cox, Probability, frequency and reasonable expectation. Am. J. Phys. **14**, 1–13 (1946)
32. J.L. Savage, *The Foundations of Statistics* (Wiley, New York, 1954). There is a second edition by Dover, New York 1972
33. D.V. Lindley, A statistical paradox. Biometrika **44**, 187–192 (1957)
34. J.L. Savage, Elicitation of personal probabilities and expectations. J. Am. Statist. Assoc. **66**, 783–801 (1971)
35. B. de Finetti, *Probability Induction and Statistics: The Art of Guessing* (Wiley, New York, 1972)
36. B. de Finetti, *Theory of Probability* (Wiley, New York, 1974)
37. O.A. Al-Hujaj, Objektive Bayessche Statistik. Theorie und Anwendung. Master's thesis, Fakultät für Physik und Astronomie der Universität Heidelberg, Max-Planck-Institut für Kernphysik, D-69029 Heidelberg (1997)
38. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.*, XXVII, 379–423, 622–656 (1948)
39. P.M. Lee, *Bayesian Statistics: An Introduction*, 2nd edn. (Arnold, London, 1997)
40. F. Benford, The law of anomalous numbers. Proc. Am. Philos. Soc. **78**, 551–572 (1938)
41. D.S. Lemons, On the numbers of things and the distribution of first digits. Am. J. Phys. **54**, 816–817 (1986)
42. P. Ehrenfest, T. Ehrenfest, Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. Physikalische Zeitschrift **8**, 311–314 (1907)

Chapter 3
Probable and Improbable Data

In Sect. 3.1, the Bayesian interval is defined. It contains the probable values of a parameter ξ and serves as the "error interval" of ξ. It is the basis of decisions because it allows distinguishing between probable and improbable data. It requires a measure $\mu(\xi)$ to be defined in the space of ξ. Examples are discussed in Sect. 3.2. The construction of the Bayesian interval is described in Sect. 3.3. In Sect. 3.4 we formulate a condition for the existence of the Bayesian interval. Its existence is necessary in order to infer ξ. The solutions of the problems suggested to the reader are given in Sect. A.3.

3.1 The Bayesian Interval

The conclusions that we draw from perceptions are based on the assignment of probabilities. A text that one reads may be spoiled; one still grasps its message with a reasonably high probability. If the text is ruined, the message becomes ambiguous. Waiting for a friend who is late, one may initially assume that he or she has the usual problems with traffic. When too much time elapses, one assumes that the friend will not come at all. A radioactive substance that emits on average one particle per second should not allow a break of one hour during which no radiation is registered. If that happens, one seeks a flaw in the apparatus. Although the time between two events is random, a break of one hour seems extremely improbable in this case.

What is an improbable event? The examples show that its definition is the basis of decisions. If there is a theory predicting ξ^{pre}, and ξ^{pre} turns out to be improbable, one rejects the theory. Let $Q(\xi)$ be a proper distribution of the real parameter ξ. The value ξ^{pre} is improbable if it is outside an interval $\mathcal{B}(K)$ containing the probable events. One is free to choose the probability K with which the probable events fall into $\mathcal{B}$. It has the property

$$\int_{\mathcal{B}} d\xi\; Q(\xi) = K \; . \tag{3.1}$$

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_3

There is a manifold of areas that fulfil this equation. We require that the smallest one exists and is unique. This is $\mathcal{B}(K)$. We call it *the* Bayesian interval or more generally *the* Bayesian area [1].

The construction of the smallest area requires the definition of the length or more generally the volume

$$V = \int_{\mathcal{I}} d\xi \, \mu(\xi) \tag{3.2}$$

of an area $\mathcal{I}$. A measure μ is needed in order to define the volume such that it is independent of reparameterisations. The interested reader should show that V is not changed by the transformation (2.7). Only in combination with a measure is the “smallest interval” a meaningful notion. Decisions and error intervals require a measure. The measure is identified with the prior distribution $\mu(\xi)$ defined systematically in Chaps. 6, 9, and 11.

Why do we consider the smallest interval $\mathcal{B}(K)$ to be *the* error interval? There are innumerable areas in which ξ is found with probability K. For the Gaussian distribution (2.16), (2.17) they may extend to infinity. An error interval that extends to infinity does not convey much information about ξ. An error interval larger than $\mathcal{B}$ yields less information than does $\mathcal{B}$.

Note that $1 - K$ is the probability that ξ is outside $\mathcal{B}(K)$. To reject a theory because ξ^{pre} falls outside the Bayesian area is erroneous with probability $1 - K$. Hence, K should be chosen reasonably close to unity. One cannot choose it equal to unity without losing the possibility to decide. Every decision remains a risk. The reader should discuss why this is so!

3.2 Examples

3.2.1 The Central Value of a Gaussian

Let ξ be the central value of the Gaussian (2.16). The measure μ is uniform and the distribution of ξ is the posterior P of (2.17), which is again the Gaussian (2.16). Owing to the reflection symmetry of this function, the Bayesian interval is centred at the observed x and can be described as $\mathcal{B}(K) = [x - \Delta\xi(K), x + \Delta\xi(K)]$. The notation $\Delta\xi$ shall say that this is an interval in the space of the parameter ξ. The observed event x does not have an error. It is what it is. The parameter ξ that one infers has an error.

One usually chooses $\Delta\xi$ to be a simple multiple of the standard deviation σ of the Gaussian. When $K = 0.68$ then $\Delta\xi = \sigma$. The value of $\Delta\xi$ is given for several values of K in Table 3.1. The value of $\Delta\xi$ is displayed in Fig. 3.1 in units of σ.

The ubiquitous occurrence of the Gaussian distribution in measurement uncertainty is explained in Chap. 4. The results of the present subsection reproduce the well-known and widely-used Gaussian error assignments.

Table 3.1 Bayesian intervals for the Gaussian model

K	$\Delta\xi$
0.6827	$1.000\,\sigma$
0.8000	$1.282\,\sigma$
0.9000	$1.645\,\sigma$
0.9500	$1.386\,\sigma$
0.9545	$2.000\,\sigma$
0.9973	$3.000\,\sigma$
$1 - 6.3 \times 10^{-5}$	$4.000\,\sigma$

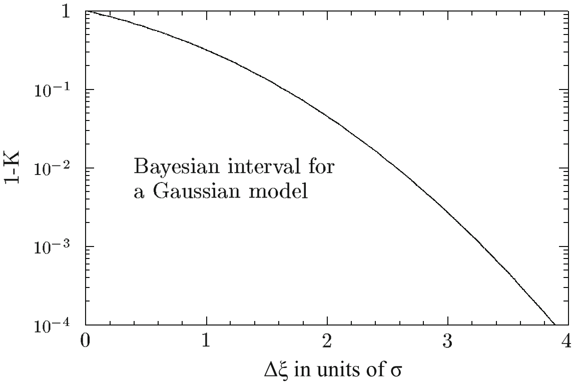


Fig. 3.1 The Bayesian interval for the Gaussian model. The relation between $1 - K$ and the half-length $\Delta\xi$ of the Bayesian interval is given in units of σ

3.2.2 *The Standard Deviation of a Gaussian*

Suppose that the central value of the Gaussian is known to be zero and that the standard deviation σ is to be inferred from the event x; that is, the model is

$$p(x|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad ; \quad x \text{ real}, \ \sigma > 0 \ . \tag{3.3}$$

An example of this model occurred in experiments on parity violation in highly excited nuclear states [2, 3]. The parity mixing matrix elements were measured for many states close to the neutron threshold [4]; see Fig. 3.2. These matrix elements fluctuate statistically with the mean value zero. The parameter of interest - which yields information on the strength of the parity violating force as well as the reaction mechanism - is the standard deviation of the matrix elements. We encounter this example again in Chap. 10, where the central value of the Gaussian is inferred in addition.

In the case of the model (3.3), the measure is

$$\mu(\sigma) \propto \sigma^{-1} \ . \tag{3.4}$$

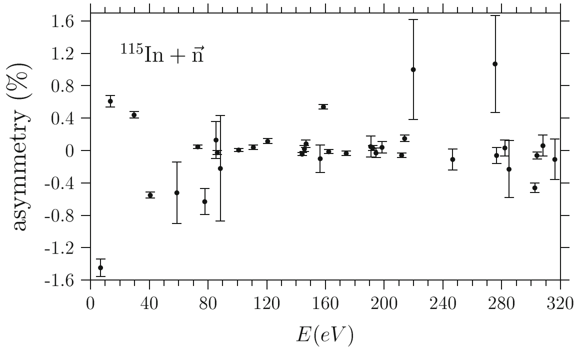


Fig. 3.2 Parity violating matrix elements from ^{115}In. The data on this figure are from [5, 6]. The asymmetry of the absorption cross-section of longitudinally polarised neutrons is shown. Its root mean square deviation from zero is the quantity of interest. The discussion in the present section disregards the systematic errors indicated by the *error bars*

This yields the posterior distribution

$$P(\sigma|x) = \left(\frac{2}{\pi}\right)^{1/2} x\sigma^{-2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x > 0, \tag{3.5}$$

for a single event of the experiment in Fig. 3.2. It is conditioned by the absolute value of the observed x. The interested reader should show that this distribution is normalised to unity.

Before we discuss the Bayesian interval of σ, let us justify the measure (3.4). We use the argument of Sect. 2.3 about distributions of the form (2.15). One can bring (3.3) into this form, invariant under translations, by help of the transformation

$$\eta = \ln \sigma^2 \tag{3.6}$$

which will yield the uniform measure if (3.4) is correct. At the same time we substitute

$$y = \ln(x^2/2) \tag{3.7}$$

and obtain

$$p_T(y|\eta) = \left(\frac{1}{\pi}\right)^{1/2} \exp\left(\frac{1}{2}[y - \eta] - e^{y-\eta}\right). \tag{3.8}$$

This result[1] depends on the difference $y - \eta$ only. Therefore the measure is uniform with respect to the parameter η, and the ansatz (3.4) is correct. We note that

[1] The corresponding expression (3.8) in the first edition of the present book contains an error. Together with the figures showing this distribution, the error is corrected in the present Eq. (3.8).

(3.8) is a version of the chi-squared distribution with 1 degree of freedom. See the discussion in Sect. 4.1.3.

These arguments can be generalised to find the measure $\mu(\sigma)$ of any model with the structure

$$q(x|\sigma) = \sigma^{-1} w\left(\frac{x}{\sigma}\right) , \tag{3.9}$$

where $w(y)$ is a normalised distribution of y. The interested reader is asked to work out the prior (3.4) from the above arguments. The general way to find the prior is given in Chap. 9. As an application, the reader should answer the following question. Is it possible to infer the mean lifetime of a radioactive substance from the observation of a single event? The model of radioactive decay is given by Eq. (4.40) in Sect. 4.2. The question has been debated in the published literature [7].

Let us return to the problem of inferring the standard deviation σ. In the parameterisation of Eqs. (3.6), (3.7), the model and the posterior are given by the same expression,

$$P_T(\eta|y) = p_T(y|\eta) , \tag{3.10}$$

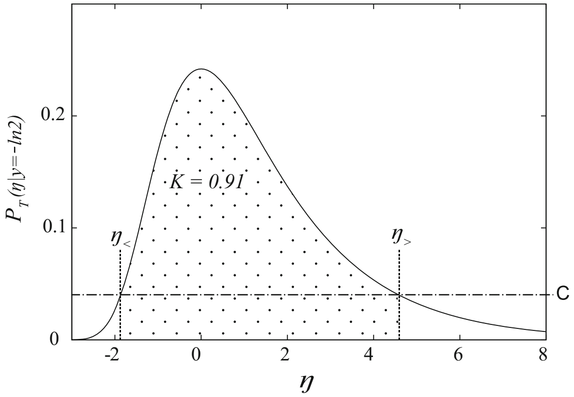
whence the posterior distribution is proper. It is displayed in Fig. 3.3 for the event $x = 1$; that is, $y = 0$. The Bayesian interval obviously exists. The maximum of P_T is at $\eta = y$ or $\sigma = x$. Unlike the Gaussian distribution, P_T is not symmetric with respect to this point. It falls off faster to the left than to the right-hand side. For $K = 0.90$, the limits of $\mathcal{B}$ are indicated in the figure.

The transformation from (3.3) via (3.6), (3.7) to (3.8) has led us to the structure of the model (2.15) in Sect. 2.3. Thus the model (3.3) provides another example of form invariance.

In the case of a multidimensional $x = (x_1, \dots, x_N)$ and σ to be inferred, the Gaussian model (2.23) turns into

$$p(x_1, ..., x_N \mid \sigma) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{N}{2\sigma^2}\langle x^2\rangle\right) , \quad x_k \text{ real} , \tag{3.11}$$

Fig. 3.3 The construction of a Bayesian interval. The distribution $P_T(\eta|y)$ of Eqs. (3.8), (3.10) is shown. The event is $y = -\ln 2$ or equivalently $x = 1$; that is, y is such that the maximum of the curve occurs at $\eta^{\mathrm{ML}} = 0$. The shaded area amounts to K = 0.90. The Bayesian interval is $[\eta_<, \eta_>]$. The borders are found by the intersection of a suitable level C with the curve



if we assume that $\xi = 0$ is known. The prior distribution of σ is given in (3.4), and the posterior is

$$P(\sigma|x_1, \dots, x_N) = \frac{2}{\Gamma\left(\frac{N}{2}\right)} \left[\frac{N}{2}\langle x^2\rangle\right]^{N/2} \sigma^{-N-1} \exp\left(-\frac{N}{2}\frac{\langle x^2\rangle}{\sigma^2}\right) . \tag{3.12}$$

This distribution is normalised to unity as one recognises via the substitution

$$\sigma \to \tau = \frac{N}{2}\frac{\langle x^2\rangle}{\sigma^2}$$

and the integral representation (B.23) of the Γ function in Chap. B.4. The interested reader may verify Eq. (3.12).

We introduce the "likelihood function"

$$\begin{aligned} L(\sigma|x) &= \frac{P(\sigma|x_1, \dots, x_N)}{\mu(\sigma)} \\ &= \frac{p(x_1, \dots, x_N|\sigma)}{m(x_1, \dots, x_N)} . \end{aligned} \tag{3.13}$$

It transforms like a function with respect to both variables, the event x and the parameter ξ. In contrast, the model $p(x|\xi)$ transforms like a distribution with respect to x and like a function with respect to ξ, whereas the posterior $P(\xi|x)$ transforms like a distribution with respect to ξ and like a function with respect to x. The quantity $m(x)$ transforms like the distribution p; it is defined by Eq. (2.5). For the definition of the Bayesian interval or area in Sect. 3.3 the numerical value of m is important. For other uses of L such as the determination of the ML estimator (see below) and the definition of the geometric measure (see Chap. 9) only logarithmic derivatives of L with respect to parameters (not to x) are needed; then m is immaterial and we restrict ourselves to define L in the form of $L \propto p(x|\sigma)$.

The likelihood function possesses a maximum. By requiring the derivative of $\ln L$ to vanish we find the maximum of (3.13) at

$$\sigma^{\mathrm{ML}}(x_1, \dots, x_N) = \left(\langle x^2\rangle\right)^{1/2} . \tag{3.14}$$

For this purpose L can be defined as

$$L(\sigma|x) \propto p(x_1, \dots x_N|\sigma) . \tag{3.15}$$

The quantity σ^{ML} is called the "maximum likelihood estimator" or briefly the "ML estimator" for the parameter σ. The reader is asked to verify Eq. (3.14).

Substituting σ according to Eq. (3.6) and setting

$$y = \ln\left(\frac{N}{2}\langle x^2\rangle\right) , \tag{3.16}$$

the posterior (3.12) takes the form

$$P_T(\eta|y) = \frac{1}{\Gamma(N/2)} \exp\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right) . \qquad (3.17)$$

This is a generalisation of the distribution (3.8) depending on the difference $y - \eta$ between event and parameter variables. It is form invariant with translational symmetry.

The event variable y of the posterior (3.17) is a function of $\langle x^2 \rangle$ according to Eq. (3.16). Thus, for the multidimensional x, the transformed posterior P_T is conditioned by a function of $\langle x^2 \rangle$ and only by this function of the N events. Indeed, the ML estimator of η turns out to be

$$\eta^{\mathrm{ML}} = \ln\langle x^2 \rangle . \qquad (3.18)$$

The interested reader is asked to prove it. Form invariance guarantees that the ML estimator supplies the sufficient statistic; see Sect. 6.7. Generally, however, it is not possible to write it down as a function of x in a closed and simple form, as was done in the present example.

We note that (3.12) and (3.17) are equivalent versions of the posterior of the chi-squared distribution with N degrees of freedom. The chi-squared distribution is defined in Sect. 4.1.3.

Note that in Fig. 3.3 the Bayesian interval is limited by the intersections of the function P_T with a horizontal line. This should be so according to a rule given in the next section.

3.3 Contour Lines

Consider the posterior distribution (3.8). Because it depends exclusively on the difference between event and parameter, the measure with respect to η is uniform. Then there is a positive number $C = C(K)$ such that the Bayesian interval $\mathcal{B}(K)$ consists of the points η that have the property

$$P_T(\eta|y) > C(K) . \qquad (3.19)$$

In Fig. 3.3, this interval is indicated. The borders are $\eta_<$ and $\eta_>$.

With the help of Fig. 3.4, we can show that $\mathcal{B}(K)$ has the minimum length out of all intervals in which η occurs with probability K. Replace $\mathcal{B}(K)$ by the interval $[a, b]$. Because

$$\int_a^b d\eta \, P_T(\eta|y) = K , \qquad (3.20)$$

the integrals over the intervals A and B are equal. In A, the value of P_T is everywhere larger than in B. Therefore the interval $[\eta_<, a]$, which has been cut away from $\mathcal{B}$, is

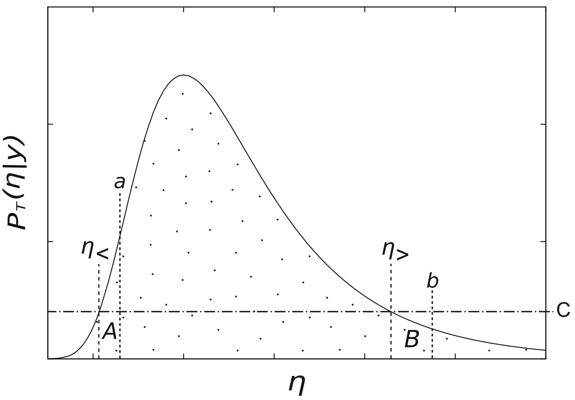


Fig. 3.4 Proof that the Bayesian interval has the minimum length. The curve is the same as in Fig. 3.3. The integrals over the intervals A and B are equal. The length of A is smaller than the length of B

shorter than the interval $[\eta_>, b]$, which has been added to $\mathcal{B}$, and therefore $[a, b]$ is longer than $\mathcal{B}$.

If one uses the parameterisation (3.5) - or any other one - then the likelihood function of Eq. (3.13) becomes

$$\begin{aligned} L(\sigma|x) &= \frac{P(\xi|x)}{\mu(\xi)} \\ &= \frac{P_T(\eta|x)}{\mu_T(\eta)} \,. \end{aligned} \tag{3.21}$$

It transforms as a function, not as a distribution, with respect to both its variables. Therefore the Bayesian interval $\mathcal{B}(K)$ consists of all points ξ with the property

$$\frac{P(\xi|x)}{\mu(\xi)} > C(K) \,, \tag{3.22}$$

and neither the ML estimator nor the Bayesian interval change their places under the transformation.

Let us adapt these arguments to a case where the parameter $\xi = (\xi_1, \xi_2)$ is two-dimensional. We speak of the “Bayesian area” because it is a surface in this case.

Consider the normalised distribution $Q(\xi)$ and suppose that the measure $\mu(\xi)$ is known. In more than one dimension, one cannot be sure of finding a reparameterisation such that the measure becomes uniform. Yet with the help of Fig. 3.5, we show that there is a positive number $C(K)$ such that the Bayesian area $\mathcal{B}(K)$ consists of the points ξ with the property

$$\frac{Q(\xi)}{\mu(\xi)} > C(K) \ , \tag{3.23}$$

that is, the points where the likelihood function is larger than C.

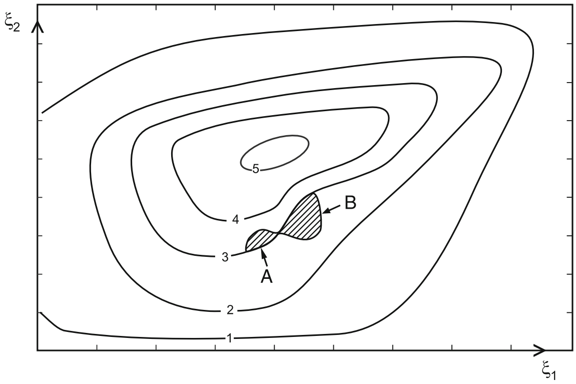


Fig. 3.5 Proof that the Bayesian area has the minimum size. The contour plot of the likelihood function (3.23) is given. The integrals over the shaded areas A and B are equal. The area of A is smaller than the area of B

Figure 3.5 is a contour plot of the likelihood function (3.23). The contour lines are the places where likelihood Q/μ assumes a constant value. This definition ensures that contour lines are invariant under reparameterisations. The interested reader should show this. Consider the contour line labelled 3 which encloses a domain $\mathcal{B}$ such that ξ is in $\mathcal{B}$ with probability K. We show that $\mathcal{B}$ is the domain with minimum area. For this we modify $\mathcal{B}$ by taking away the area A and adding B. The modified domain is again required to contain ξ with the probability K, whence

$$\int_A d\xi\, Q(\xi) = \int_B d\xi\, Q(\xi) \ . \tag{3.24}$$

The area of A is

$$\begin{aligned}\int_A d\xi\, \mu(\xi) &= \int_A d\xi\, \frac{\mu(\xi)}{Q(\xi)} Q(\xi) \\ &= \frac{\mu(\xi_a)}{Q(\xi_a)} \int_A d\xi\, Q(\xi) \,,\end{aligned} \tag{3.25}$$

where ξ_a is a suitable point inside A. Similarly, we have

$$\int_B d\xi\, \mu(\xi) = \frac{\mu(\xi_b)}{Q(\xi_b)} \int_B d\xi\, Q(\xi) \ , \tag{3.26}$$

where ξ_b is inside B. Inasmuch as the likelihood function in A is larger than in B, the area of B is larger than the area of A, whence the area of the modified domain is larger than that of $\mathcal{B}$.

One can summarise these results by saying: the limit of a Bayesian area $\mathcal{B}(K)$ is a contour line.

We require that there is a point $\xi = \xi^{\mathrm{ML}}$, where L assumes an absolute maximum. This point, the ML estimator ξ^{ML}, belongs to any Bayesian interval of Q. The reader should show that ξ^{ML} does not change under reparameterisations.

3.4 On the Existence of the Bayesian Area

In Sect. 2.4, we have discussed the model (2.28) which did not allow constructing the Bayesian interval. This happened because the posterior $P(\xi|x)$ equalled the measure $\mu(\xi)$, wherever P differed from zero. Thus the likelihood function $L(\xi|x)$ was constant; it did not possess a maximum.

The likelihood function must nowhere be independent of ξ because the model p shall allow us to infer ξ from the observed x. Form invariance guarantees that a maximum of the likelihood function exists. We show this for one-dimensional ξ.

Let ξ be one-dimensional. Then we can find a transformation such that the measure is $\mu \equiv$ const. We suppose that this is given. If there were an interval of ξ where the likelihood function is constant then the model $p(x|\xi)$ would be independent of ξ there. In that interval of its domain of definition the parameter ξ could not be inferred.

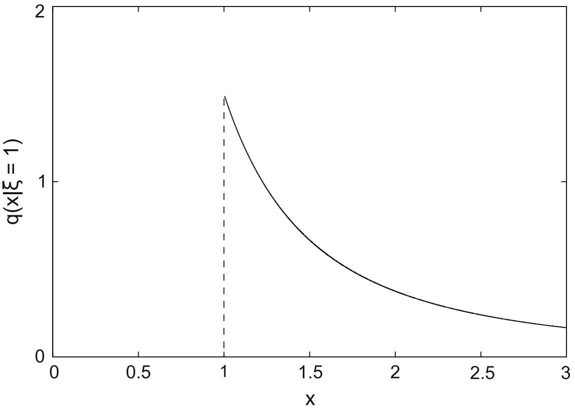
When the measure is uniform and the model is form invariant, the posterior P will depend on the difference $x - \xi$ and only on it. The event x amounts to a translation of the parameter ξ. All translations are allowed. Because P is normalised, it must possess a maximum. We assume that there is a unique absolute maximum. The general definition of form invariance is given in Chap. 6.

Although the point of maximum likelihood lies within every Bayesian interval, it would not be a good estimator if it were found at the edge of the Bayesian intervals. This cannot happen because we require the likelihood function to be regular. A counterexample is given by the Pareto model [8]

$$q(x|\xi) = \begin{cases} \frac{\alpha}{\xi}\left(\frac{\xi}{x}\right)^{1+\alpha} & \text{for} \quad x > \xi \\ 0 & \text{for} \quad x < \xi \end{cases}, \tag{3.27}$$

where $\xi > 0$ and $\alpha > 0$. In Fig. 3.6, we show q for $\alpha = 3/2$ and $\xi = 1$.

Fig. 3.6 The Pareto distribution for $\alpha = 3/2$ and $\xi = 1$



This is a model of the type (3.9). The prior is (3.4). The posterior

$$Q(\xi|x) = \begin{cases} \frac{3}{2}\xi^{1/2} \text{ for } & \xi < 1 \\ 0 \quad \text{for } & \xi > 1 \end{cases} \tag{3.28}$$

is given in Fig. 3.6 for $\alpha = 3/2$ and $x = 1$. The likelihood function is

$$L(\xi|x) = \begin{cases} \xi^{3/2} \text{ for } \xi < 1 \\ 0 \quad \text{for } \xi > 1 \end{cases} . \tag{3.29}$$

The point of maximum likelihood is at $\xi^{\mathrm{ML}} = 1$. This is the upper border of every Bayesian interval. Thus the estimator is intuitively not satisfactory. It should be a point where the derivative of Q vanishes. We ascribe this failure to the fact that there is a discontinuity in the dependence of the Pareto model on its parameter. The likelihood L should be a regular function of ξ. Then there will be a point ξ^{ML}, where L is maximal. This point will not be at the border of the domain of definition of ξ. Because the prior μ is monotonic, the posterior Q will be maximal at ξ^{ML}, too (Fig. 3.7).

The Pareto model has played a role in econometric investigations [9]. Distributions of wealth tend to be quite different from Gaussian. They have no reflection symmetry and decrease slowly with increasing wealth. An unusual example is Abul-Magd's study of wealth in ancient Egypt [10]. From excavations at Tell el-Amarna in Middle Egypt he obtained a distribution of the area A of houses in the fourteenth century B.C. The excavations concerned a city called Akhetaten at its time. For religious reasons this city existed for only about 18 years starting from 1372 B.C. So the distribution of A served as the distribution of wealth at one moment in ancient Egypt. The histogram of the data is shown in Fig. 3.8. Abul-Magd tries to fit the data with the Pareto model

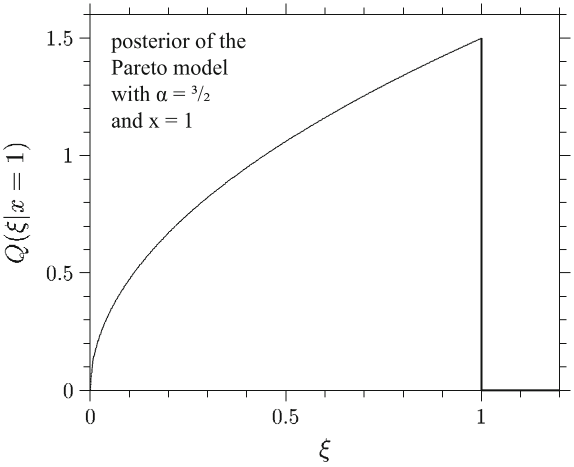
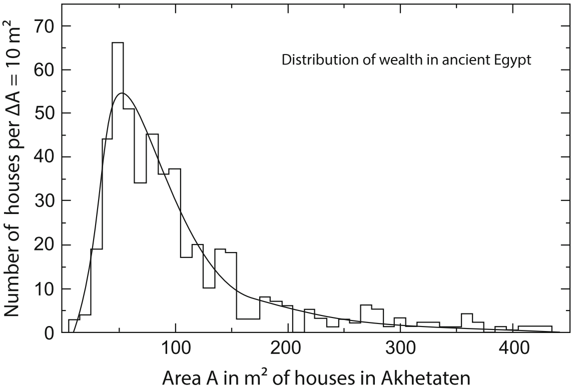


Fig. 3.7 The posterior (3.29) of a Pareto distribution

Fig. 3.8 The distribution of wealth in ancient Egypt according to Ref. [10]. The fit to the histogram is shown by the curve. This curve is a version of the chi-squared distribution with 3.8 degrees of freedom; see Sect. 4.1.3



but he does not find the discontinuity in the data and therefore replaces the model (3.27) by

$$w(A|\alpha) \propto A^{-1-\alpha} \exp\left(-\frac{1}{\tau A}\right), \tag{3.30}$$

where τ is a positive number. For large A this distribution behaves as the Pareto distribution, but there is no discontinuity for $A > 0$. The parameter α is found to be $\alpha = 3.76 \pm 0.19$. The statistical error has not been obtained via Bayesian statistics. Our interest in Ref. [10] is due to the fact that the author had to look for a model that depends regularly on its parameter(s).

Let us note that the distribution w becomes a chi-squared distribution with $N = \alpha$ degrees of freedom by transforming A to A^{-1}. The chi-squared model is treated in Sect. 4.1.3.

References

1. E.T. Jaynes, Confidence intervals vs. Bayesian intervals, in *Papers on Probability, Statistics and Statistical Physics*, vol. 158, Synthese Library, ed. by R.D. Rosenkrantz (Reidel, Dordrecht, 1983), pp. 149–209. The original article is in [11]
2. V.P. Alfimenkov, S.B. Borzakov, V. Van Tkhuan, Y.D. Mareev, L.B. Pikel'ner, A.S. Khrykin, E.I. Sharapov, Parity violation at the 0.75-ev neutron resonance of ^{139}La. JETP Lett. **35**(1), 51–54 (1982)
3. V.P. Alfimenkov, S.B. Borzakov, V. Van Tkhuan, Y.D. Mareev, L.B. Pikel'ner, A.S. Khrykin, E.I. Sharapov, Parity nonconservation in neutron resonances. Nucl. Phys. A **398**, 93–106 (1983)
4. G.E. Mitchell, J.D. Bowman, H.A. Weidenmüller, Parity violation in the compound nucleus. Rev. Mod. Phys. **71**, 445–457 (1999)
5. G.E. Mitchell, J.D. Bowman, S.I. Penttilä, E.I. Sharapov, Parity violation in compound nuclei: experimental methods and recent results. Phys. Rep. Rev. Sec. Phys. Lett. **354**, 157–241 (2001). The original publication on parity violation in ^{115}In is [6]
6. S.L. Stephenson, J.D. Bowman, F. Corvi, B.E. Crawford, P.P.J. Delheij, C.M. Frankle, M. Iinuma, J.N. Knudson, L.Y. Lowie, A. Masaike, Y. Masuda, Y. Matsuda, G.E. Mitchell, S.I.

Pentilä, H. Postma, N.R. Roberson, S.J. Seestrom, E.I. Sharapov, H.M. Shimizu, Y.-F. Yen, V.W. Yuan, L. Zanini, Parity violation in neutron resonances in ^{115}in. Phys. Rev. C **61**, 045501, 1–11 (2000)

7. H.W. Lewis, What is an experiment? Am. J. Phys. **50**, 1164–1165 (1982)
8. V. Pareto, *Cours d'Économie Politique* (Macmillan, Paris, 1896)
9. H. Aoyama, Y. Nagahara, M.P. Okazaki, W. Sauma, H. Takayasu, M. Takyasu, Pareto's law for income of individuals and debt of bankrupt companies. Fractals **8**, 293–300 (2000)
10. A.Y. Abul-Magd, Wealth distribution in an ancient Egyptian society (2002)
11. W.L. Harper, C.A. Hooker (eds.), *Foundations of ProbabilityTheory, Statistical Inference, and Statistical Theories of Science* (Reidel, Dordrecht, 1976)

Chapter 4
Description of Distributions I: Real *x*

In the present chapter, some important distributions are defined and described. The event variable x is real, as opposed to discrete; that is, the distributions are probability densities. In Sect. 4.1, we describe Gaussian models. The exponential distribution is described in Sect. 4.2. The Cauchy and Student's t-distribution are defined in Sect. 4.3. Section A.4 gives the solutions of the problems suggested to the reader.

4.1 Gaussian Distributions

The Gaussian[1] distribution which seems to have been discovered by A. de Moivre,[2] is the most frequently used statistical model. Its simplest version is treated in Sect. 4.1.1; the multidimensional Gaussian is introduced in Sect. 4.1.2 and the family of chi-squared distributions in Sect. 4.1.3.

4.1.1 The Simple Gaussian

The Gaussian distribution of a single event variable x has two parameters, the central value ξ and the variance σ. It is given by

$$q(x|\xi,\sigma) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right)\,, \tag{4.1}$$

[1] Carl Friedrich Gauß, 1777–1855, German mathematician, astronomer, and physicist. He contributed to number theory, celestial and general mechanics, geodesy, differential geometry, magnetism, optics, the theory of complex functions, and statistics.

[2] Abraham de Moivre, 1667–1754, French mathematician, emigrated to England after the revocation (1685) of the tolerance edict of Nantes.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_4

and is represented in Fig. 2.1. The event x and the parameter ξ are defined on the whole real axis. The normalising factor is obtained from the integral

$$\begin{aligned} Z(a) &= \int_{-\infty}^{\infty} dx \exp\left(-a(x-\xi)^2\right) \\ &= \sqrt{\pi/a}\,; \end{aligned} \tag{4.2}$$

compare Problem A.3.3. The interested reader should convince himself or herself that the mean value $\overline{x}$ is equal to ξ.

The *variance* of a random variable x, is the mean square deviation from $\overline{x}$; that is,

$$\text{var}(x) = \overline{(x-\overline{x})^2}\,. \tag{4.3}$$

Here, the overlines denote expectation values with respect to the distribution of x. The square root of the variance is called the *standard deviation* or root mean square deviation. It quantifies the fluctuations of x about its mean value. A transformation $y \to Tx$, however, changes the value of the standard deviation. The interested reader should show that the variance can also be expressed as

$$\text{var}(x) = \overline{x^2} - \overline{x}^2\,. \tag{4.4}$$

The reader is asked to prove that, for the Gaussian distribution (4.1), one has

$$\begin{aligned} \text{var}(x) &= \sigma^2\,, \\ \overline{(x-\xi)^4} &= 3\sigma^2\,. \end{aligned} \tag{4.5}$$

To calculate the moments $\overline{(x-\xi)^{2n}}$, it is helpful to consider the derivatives of the function $Z(a)$ of (4.2).

The prime importance of the Gaussian distribution is a consequence of the central limit theorem. This theorem can be stated as follows. Let there be N random variables $x_1, \ldots, x_N$ that follow the distribution $w(x_k)$. Then the distribution $W(z)$ of their average $z = \langle x \rangle$ tends to a Gaussian for $N \to \infty$, if the mean value $\overline{x_k}$ and the variance $\overline{x_k^2} - \overline{x_k}^2$ exist. The notation $\langle \ldots \rangle$ is defined in Eq. (2.21).

Thus for large N, one obtains the approximation

$$W(z) \propto \exp\left(-\frac{(z-A)^2}{2B^2}\right)\,, \tag{4.6}$$

where the central value A and the variance B^2 are approximately

$$A \approx \langle x \rangle \tag{4.7}$$

and

$$B^2 \approx \langle x^2 \rangle - \langle x \rangle^2 \,, \tag{4.8}$$

compare Sect. 2.5.

We do not prove the central limit theorem here, but we illustrate it below. Based on this theorem, the Gaussian model is invoked whenever the observable x can be considered as the sum of several contributions that all follow the same or similar distributions. Examples are given by the logarithm of the joint probability of several independent events, the noise in an instrument of measurement, the amplitude of waves in the oceans. To assume the Gaussian model can be misleading because the central limit theorem does not specify how large N must be in a given context. A strict application of the Gaussian distribution to the amplitudes of waves would practically exclude the occurrence of monster waves in the oceans: a wrong result.

The central limit theorem is illustrated in Fig. 4.1. In each of the four parts of the figure, a histogram with 100 bins is shown. The binning corresponds to an equidistant partition of the interval [0, 1]. By use of a random number generator, the x_k have been drawn from the distribution w which is uniform on the interval [0, 1]. This gives

$$\begin{aligned} A &= 0.5 \,, \\ B &= (12N)^{-1/2} \end{aligned} \tag{4.9}$$

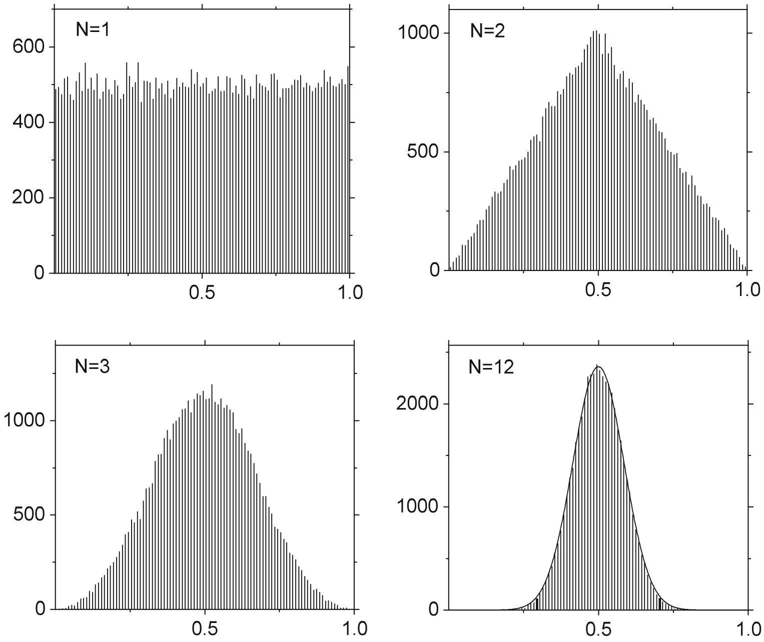


Fig. 4.1 Illustration of the central limit theorem. Shown is the distribution of the average z of N random variables x_k. The x_k are drawn from a uniform distribution w

The height of each bar is the number of cases in which the variable z has fallen into the corresponding bin. For each part of the figure, 5×10^4 values of z have been drawn. The part with $N = 1$ shows the uniform distribution w. The part with $N = 2$ displays a triangular distribution. For $N = 3$, the distribution of z starts to resemble a Gaussian. For $N = 12$, the Gaussian is given in the figure; it closely follows the histogram. This figure was inspired by Fig. 4.1 of [1]. Note, however, that the Gaussian approximation may require N to be much larger than $N = 12$ when w decreases too slowly towards the ends of its domain of definition: in other words, when it conveys little information.

Very often, the statistical fluctuations of an observable are the sum of many contributions. Because of the central limit theorem, the Gaussian model can be used to describe the fluctuations. In the present book, the Gaussian approximation is applied to the logarithm of a likelihood function $L(\xi|x_1, \dots, x_N)$ which results from a large number N of events given by one and the same model $p(x|\xi)$. Then ln L will display a single strong peak in the domain, where it is "essentially different from zero," and it will be approximated by a Gaussian.

The quantity A of Eq. (4.7) is given by the ML estimator ξ^{ML}; compare Sect. 3.2. The Gaussian approximation is justified by the observation that the likelihood function is proportional to the product

$$\frac{P(\xi|x)}{\mu(\xi)} \propto \prod_{k=1}^{N} q(x_k|\xi)\,, \tag{4.10}$$

of the probabilities to obtain x_k, whence it is a function of the sum over the logarithms ln $p(x_k|\xi)$ of the probabilities. Thus the ML estimator is a function of their average $\langle \ln q \rangle$ which is assumed to have a Gaussian distribution. The information about ξ will increase with increasing N in that the peak in L will become narrower and narrower. Eventually, the measure μ — even if it depends on ξ — can be considered constant within the domain, where L is essentially different from zero. Then L becomes a Gaussian function of ξ and P a Gaussian distribution. Therefore the Bayesian interval for the parameter ξ derived from a sufficiently high number N of observations is usually given as the Gaussian "error interval", that is, the root mean square value B of the assumed Gaussian distribution. The inverse of B^2 is given by the second derivative of the logarithm of the likelihood function (3.13),

$$\begin{aligned} B^{-2} &= \lim_{N \text{ large}} -\frac{\partial^2}{\partial \xi^2} \ln \frac{P(\xi|x_1, \dots, x_N)}{\mu(\xi)}\bigg|_{\xi=\xi^{\mathrm{ML}}} \\ &= \lim_{N \text{ large}} -\sum_{k=1}^{N} \frac{\partial^2}{\partial \xi^2} \ln q(x_k|\xi)\bigg|_{\xi=\xi^{\mathrm{ML}}} . \end{aligned} \tag{4.11}$$

For large N the sum in the second line of this equation can be replaced by an integration over the distribution of the x_k and we obtain

$$B^{-2} = NF(\xi^{\mathrm{ML}})\,, \tag{4.12}$$

where

$$F(\xi) = -\int \mathrm{d}z\, q(z|\xi) \frac{\partial^2}{\partial \xi^2} \ln q(z|\xi) \tag{4.13}$$

is called the Fisher[3] [2] function of the model $q(z|\xi)$. It is also called the "Fisher information" on the parameter ξ. Note, however, that the value of F changes when the parameter ξ is transformed. Thus F is not a property of the distribution q; it is a property of the specific parameterisation of q.

4.1.2 The Multidimensional Gaussian

The Gaussian model (4.1) can be generalised to the n-dimensional event

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{4.14}$$

We similarly introduce the vector

$$\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} . \tag{4.15}$$

The multidimensional Gaussian reads

$$p(\boldsymbol{x}|\boldsymbol{\xi}) = Z^{-1} \exp\left(-(\boldsymbol{x}-\boldsymbol{\xi})^\dagger (2C)^{-1} (\boldsymbol{x}-\boldsymbol{\xi})\right) . \tag{4.16}$$

Here, $\boldsymbol{x}$ and $\boldsymbol{\xi}$ are vectors in a Cartesian[4] space; that is, their components are real numbers and the metric applies which says that $\sqrt{\boldsymbol{x}^\dagger \boldsymbol{x}}$ is the length of the vector $\mathbf{x}$. The symbol $\boldsymbol{x}^\dagger$ stands for the transpose of $\boldsymbol{x}$. The quantity C in (4.16) is a symmetric n-dimensional matrix with positive eigenvalues, called the correlation matrix. Because all eigenvalues are positive, the expression $\boldsymbol{x}^\dagger C^{-1} \boldsymbol{x}$ is positive for every vector $\boldsymbol{x} \neq 0$. The normalising factor in Eq. (4.16) is

[3] Sir Ronald Aylmer Fisher, 1890–1962, British statistician and geneticist. He introduced statistical arguments into the design of scientific experiments. Fisher's work prepared the rediscovery of Bayesian inference in the twentieth century.

[4] René Descartes, 1596–1650, French philosopher, mathematician, and physicist. His "Discours de la méthode pour bien conduire sa raison et chercher la vérité" is a foundation of scientific thinking. His work "La géométrie" founded the formalism of analytical geometry.

$$\begin{aligned} Z(C) &= \int \mathrm{d}x_1 \dots \mathrm{d}x_n \exp\left(-\boldsymbol{x}^\dagger (2C)^{-1} \boldsymbol{x}\right) \\ &= \left((2\pi)^n \det C\right)^{1/2} \end{aligned} \tag{4.17}$$

as the interested reader can verify with the help of a transformation that diagonalises C. In (4.17), the translation by $-\xi$ that appears in (4.16) has been omitted because the integral extends over the whole real axis in each variable and is therefore independent of the shift. Very much as in the one-dimensional case of Sect. 4.1.1, the parameter ξ_ν is the expectation value of x_ν.

The correlation matrix is the generalisation of the variance σ^2 that appears in the one-dimensional Gaussian. Indeed, the elements of C are

$$\begin{aligned} C_{\nu\nu'} &= \overline{(x_\nu - \xi_\nu)(x_{\nu'} - \xi_{\nu'})} \\ &= \int \mathrm{d}x_1 \dots \mathrm{d}x_n \, (x_\nu - \xi_\nu)(x_{\nu'} - \xi_{\nu'}) \, p(\boldsymbol{x}|\boldsymbol{\xi}) \end{aligned} \tag{4.18}$$

as we prove in Sect. B.1. The expectation value of $(x_\nu - \xi_\nu)(x_{\nu'} - \xi_{\nu'})$ is also called the correlation between $(x_\nu - \xi_\nu)$ and $(x_{\nu'} - \xi_{\nu'})$, whence C is termed the "correlation matrix".

The multidimensional Gaussian has a very convenient property: the integration over any number of the event variables x_ν yields a result that is known by a simple rule. Let us integrate Eq. (4.16) over x_n; the result

$$\begin{aligned} \int \mathrm{d}x_n \, p(\boldsymbol{x}|\boldsymbol{\xi}) &= ((2\pi)^{n-1} \det K)^{-1/2} \\ &\quad \times \exp\left(- \sum_{\nu,\nu'=1}^{n-1} (x_\nu - \xi_\nu)(2K)^{-1}_{\nu,\nu'}(x_{\nu'} - \xi_{\nu'})\right) \end{aligned} \tag{4.19}$$

is again a multidimensional Gaussian and its correlation matrix K has the elements

$$K_{\nu,\nu'} = C_{\nu,\nu'} \quad \text{for } \nu, \nu' = 1, \dots, n-1 \,, \tag{4.20}$$

that is, the $(n-1)$-dimensional matrix K is obtained from C by simply omitting the last row and the last column of C. This procedure can be repeated until only x_1 and ξ_1 are left so that

$$p^\downarrow(x_1|\xi_1) = (2\pi C_{1,1})^{-1/2} \exp\left(-\frac{(x_1 - \xi_1)^2}{2C_{1,1}}\right) . \tag{4.21}$$

The rule formulated by Eqs. (4.19) and (4.20) is proven in Sect. B.2.

4.1.3 The Chi-Squared Model

Let the quantity x_k , $k = 1, \dots, N$ be distributed according to the Gaussian

$$w(x_k|\sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{x_k^2}{2\sigma^2}\right) . \tag{4.22}$$

We ask for the distribution $\chi^{\rm sq}$ of the sum

$$T = \sum_{k=1}^{N} x_k^2 . \tag{4.23}$$

This quantity (or a proportional one) was called χ^2 - in words, "chi-squared" - in earlier publications on statistical inference, whence the title of the present section and the symbol $\chi^{\rm sq}$ used below for the desired distribution.

As a first step, we substitute the variables x_k^2 by the positive definite variables r_k. The normalised distribution of r_k is

$$\tilde{w}(r_k|\sigma) = (2\pi\sigma^2)^{-1/2}\, r_k^{-1/2} \exp\left(-\frac{r_k}{2\sigma^2}\right) , \quad r_k > 0 . \tag{4.24}$$

This amounts to the model treated in Sect. 3.2.2. In physics, this distribution is called the Porter-Thomas distribution. More steps are necessary in order to generalise w to the case where $N > 1$.

We introduce the variables T and t_k via

$$r_k = T t_k, \quad k = 1, \dots N . \tag{4.25}$$

Equation (4.23) says that

$$\sum_{k=1}^{N} t_k = 1 , \tag{4.26}$$

which means that the t_k are not independent of each other. One can, for example, express t_N by $t_1, \dots, t_{N-1}$. Yet it is possible to substitute the variables $r_1, \dots, r_N$ by the variables $T, t_1, \dots, t_{N-1}$. This transformation recalls the introduction of polar coordinates in N dimensions, where the Cartesian vector x is replaced by its radius R and trigonometric functions of angles. The present variable T equals R^2. The present variables t_k are restricted to nonnegative values $0 \le t_k \le 1$. The Jacobian determinant of the transformation is

$$\left| \frac{\partial(r_1, \dots, r_N)}{\partial(T, t_1, \dots, t_{N-1})} \right| = T^{N-1} \tag{4.27}$$

as we show in Sect. B.3. Thus the joint distribution

$$W(r_1, \dots, r_N|\sigma) = \prod_{k=1}^{N} \tilde{w}(r_k|\sigma) \tag{4.28}$$

of the r_k takes the form

$$\begin{aligned}\tilde{W}(T, t_1, \dots t_{N-1}|\sigma) &\propto T^{N-1} \prod_{k=1}^{N} \tilde{w}(r_k|\sigma) \\ &\propto T^{N-1} \left(\prod_{k=1}^{N} (T t_k)^{-1/2} \right) \exp\left(-\frac{T}{2\sigma^2} \right) \\ &\propto T^{N/2-1} \left(\prod_{k=1}^{N} t_k^{-1/2} \right) \exp\left(-\frac{T}{2\sigma^2} \right) . \end{aligned} \tag{4.29}$$

This result says that $\tilde{W}$ factorises into a distribution of T and a distribution of $t_1, \dots, t_{N-1}$. The latter does not depend on σ. Therefore the variable T is statistically independent of the variables $t_1, \dots, t_{N-1}$. The distribution of T is

$$\chi_N^{\text{sq}}(T|\sigma) \propto T^{N/2-1} \exp\left(-\frac{T}{2\sigma^2} \right) . \tag{4.30}$$

We write

$$\tau = 2\sigma^2 \tag{4.31}$$

and obtain

$$\chi_N^{\text{sq}}(T|\tau) = Z^{-1} T^{N/2-1} \exp\left(-\frac{T}{\tau} \right) . \tag{4.32}$$

The factor

$$\begin{aligned} Z &= \int_0^\infty \mathrm{d}T \, T^{N/2-1} \exp\left(-\frac{T}{\tau} \right) \\ &= \tau^{N/2} \, \Gamma(N/2) \end{aligned} \tag{4.33}$$

which normalises χ_N^{sq} to unity, can be found with the help of the integral representation of the Γ function given in Sect. B.4. This yields the chi-squared model[5] with N degrees of freedom

$$\chi_N^{\text{sq}}(T|\tau) = \frac{1}{\Gamma(N/2)} \tau^{-1} \left(\frac{T}{\tau} \right)^{N/2-1} \exp\left(-\frac{T}{\tau} \right) . \tag{4.34}$$

[5] The present definition (4.34) of the chi-squared distribution differs from the corresponding Eq. (4.38) in the first edition of this book: the Fourier transform of (4.34) has a convenient property.

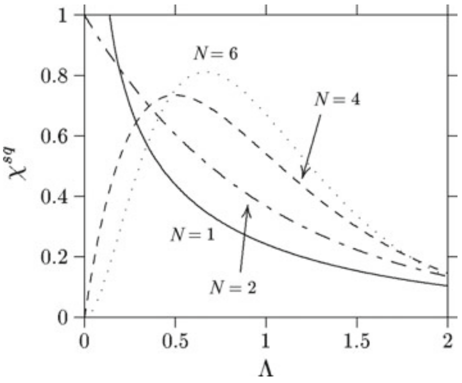


Fig. 4.2 The chi-squared distribution (4.34) for various degrees of freedom N. We have set $\tau = 2/N$. Then the expectation value $\overline{T}$ equals unity for every N. When N becomes large, the chi-squared distribution tends to the Gaussian with the parameters given by (4.35) and (4.36)

For some values of N, it is displayed in Fig. 4.2. Note that N cannot be inferred; it must be known. For this reason we do not list it among the parameters of the model but rather write it as an index to the letter χ. The scaling factor τ can be inferred via Bayesian inference. In the present case, the number of degrees of freedom equals the number of terms in the sum (4.23) because every term x_k^2 has a chi-squared distribution with one degree of freedom which is given by Eq. (4.24). A sum over N quantities, each of which obeys a chi-squared distribution with f degrees of freedom, possesses a chi-squared distribution with Nf degrees of freedom; see Sect. H.2. Here, f can be any positive real number.

The chi-squared distribution (4.34) of T has the mean value

$$\overline{T} = \frac{N}{2}\tau \tag{4.35}$$

and the variance

$$\overline{(T-\overline{T})^2} = \frac{N}{2}\tau^2 \,. \tag{4.36}$$

The interested reader should verify (4.35) and (4.36). One can do so with the help of the normalising factor Z in (4.33). By differentiating Z with respect to $1/\tau$ one obtains the moments of the chi-squared distribution.

For every number of degrees of freedom, the chi-squared model has the structure of Eq. (3.9) discussed in Sect. 3.2.2. Thus the prior distribution is

$$\mu(\tau) \;=\; \tau^{-1} \,, \tag{4.37}$$

and we expect that there is a transformation that brings the model (4.34) into a form corresponding to (3.8), where it depends on the difference between event and parameter. Indeed, the transformations

$$\begin{aligned} y &= \ln T\,, \\ \eta &= \ln \tau \end{aligned} \tag{4.38}$$

yield the somewhat unusual form

$$\tilde{\chi}_N^{\text{sq}}(y|\eta) = \frac{1}{\Gamma(N/2)} \exp\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right) \tag{4.39}$$

of the chi-squared model with N degrees of freedom. It is used in Chap. 13 to obtain a Gaussian approximation to the chi-squared model. This form shows that the chi-squared distribution is proper for every positive N. (The number of degrees of freedom need not be an integer). When N decreases, the convergence for $y \to -\infty$ occurs ever more slowly. Then the distribution becomes less and less concentrated. This concentration is quantified by the Shannon information introduced in Sect. 2.6. Thus with decreasing N the information conveyed by χ_N^{sq} decreases. Indeed, the χ_N^{sq} are a family of distributions with adjustable Shannon information. Their analytical properties are well known, and they can be used to approximate distributions that are difficult to handle. This is worked out in Chap. 13.

4.2 The Exponential Model

The exponential model

$$p(x|\xi) = \xi^{-1} \exp(-x/\xi)\,, \quad x, \xi > 0\,, \tag{4.40}$$

is the law of radioactive decay. Less obvious is the fact that it describes the distribution of the distance between independent random events that occur at the constant mean distance ξ. This is realised, for example, by a radioactive source that emits quanta at a (approximately) time-independent rate. The difference x of the times of arrival of two successive particles in a detector is distributed according to (4.40). The event x and the parameter ξ are defined on the positive real axis. Note that the exponential distribution equals the chi-squared distribution with $N = 2$ degrees of freedom.

The event has the expectation value

$$\overline{x} = \xi \tag{4.41}$$

and the variance

$$\text{var}(x) = \xi^2\,. \tag{4.42}$$

The reader is asked to use the properties of the Γ function given in Sect. B.4 to prove this.

4.3 Student's *t*-Distribution

The model

$$p(t|\gamma) = B^{-1}\gamma^{-1}\left(1 + \frac{t^2}{\gamma^2}\right)^{-\nu}, \quad t \text{ real}, \ \gamma > 0\,, \nu > 1/2\,, \tag{4.43}$$

is called the Student's t-distribution; compare Sect. 3.3 of [3]. By number 3 of Sect. 3.194 of [4], the normalising factor is the beta function

$$B = B(1/2, \nu - 1/2) \tag{4.44}$$

defined in Sect. B.5. We have taken t to be defined on the whole real axis.

For the nth moment of the t-distribution to exist, the value of ν must be larger than $(n + 1)/2$. Hence, for $\nu \le 3/2$, the second moment does not exist, and random numbers drawn from (4.43) do not conform to the central limit theorem. This means that the t-distribution models random numbers with large fluctuations. They occur, for example, in financial time series [5–8].

For $\nu \le 1$, not even the mean $\overline{t}$ exists. This is true especially for the Cauchy distribution

$$p(t|\gamma) = (\pi\gamma)^{-1}\left(1 + \frac{t^2}{\gamma^2}\right)^{-1} \tag{4.45}$$

obtained from (4.43) with $\nu = 1$. In both classical and quantum mechanics this function turns out to be the shape of a resonance line. To physicists, it is known under the name of the Lorentzian or Breit–Wigner distribution. It violates the central limit theorem: if the variables t_k for $k = 1, \dots, N$ follow one and the same Cauchy distribution $p(t|\gamma)$, their average

$$z = \frac{1}{N}\sum_{k=1}^{N} t_k \tag{4.46}$$

again follows a Cauchy distribution, albeit with a smaller value of γ; that is, the distribution becomes narrower with increasing N. Still the variable z never follows a Gaussian distribution.

In Fig. 4.3, the Cauchy distribution (4.45) is compared to the Gaussian

$$p_{\text{Gauss}}(t|\sigma) = (2\pi)^{-1/2}\sigma^{-1}\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{4.47}$$

For the comparison, the values of γ and σ must be chosen such that the unit of length along the t-scale is the same for both models. This becomes clear from Sects. 6.4 and 9.2, where the geometric measure in the space of a parameter is discussed. With

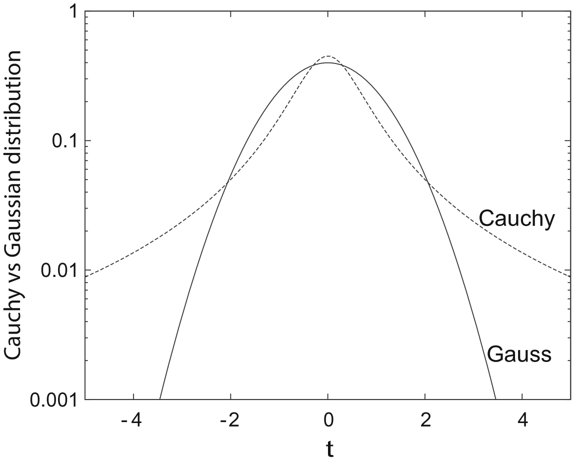


Fig. 4.3 The Cauchy distribution (4.45) with $\gamma = 2^{-1/2}$ compared to the Gaussian (4.47) with $\sigma = 1$

$\gamma = 2^{-1/2} = 0.7071$ and $\sigma = 1$ both models yield the same unit of length in the space of t. The slow decay of the Cauchy distribution is conspicuous.

The variable z of Eq. (4.46), in connection with the Cauchy distribution, remains a quantity with large fluctuations for any N. However, the ML estimator for the parameter γ from N events is different from z and the posterior distribution allows for fewer and fewer fluctuations with growing N. We give an argument for this. The model (4.45) is of the type (3.9), whence the prior distribution is $\mu(\gamma) \propto \gamma^{-1}$, and the posterior distribution

$$P(\gamma \mid t_1, \ldots, t_N) \propto \gamma^{-N-1} \prod_{k=1}^{N} \left(1 + \frac{t_k^2}{\gamma^2}\right)^{-1} . \tag{4.48}$$

vanishes as γ^{-N-1} for $|\gamma| \to \infty$.

A generalisation of the Student's t-distribution is

$$p(t|\gamma) = B^{-1}\gamma^{-1} \frac{\left(\frac{t^2}{\gamma^2}\right)^{\mu-1/2}}{\left(1 + \frac{t^2}{\gamma^2}\right)^{\nu}} , \quad t \text{ real} , \gamma > 0 , \nu > \mu . \tag{4.49}$$

As with (4.43) this distribution decays algebraically according to $t^{2\mu-2\nu-1}$ for large t. It is again a distribution with large fluctuations. In Eq. (4.49), the normalising factor is the beta function

$$B = B(\mu, \nu - \mu) , \tag{4.50}$$

and t is defined on the whole real axis. The integral representation of the beta function is found in Sect. B.5.

References

1. R. Barlow, *Statistics* (Wiley, Chichester, 1989)
2. R.A. Fisher, Theory of statistical information. Proc. Cambridge Philos. Soc. **22**, 700–725 (1925)
3. D.S. Sivia, *Data Analysis* (A Bayesian Tutorial. Clarendon, Oxford, 1998)
4. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products* (Academic Press, New York, 2015)
5. Benoît Mandelbrot, The variation of certain speculative prices. J. Bus. **36**, 394–419 (1963)
6. H.E. Stanley, L.A.N. Amaral, D. Canning, P. Gropokrishnan, Y. Lee, Y. Liu, Econophysics: Can physicists contribute to the science of economics? Physica A **269**, 156–169 (1999)
7. R.N. Mantegna, H.E. Stanley, Scaling behaviour in the dynamics of an economic index. Nature **376**, 46–49 (1996)
8. V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Econophysics: Financial time series from a statistical physics point of view. Physica A **279**, 443–456 (2000)

Chapter 5
Description of Distributions II: Natural *x*

In the present chapter, distributions are described in which the event x is a number of hits or a count rate; that is, x is a natural number. Therefore the distribution $p(x|\xi)$ is not a probability density but, rather, a probability. The parameter ξ, however, is real. In Sect. 5.1, the binomial distribution is presented. In Sect. 5.2, the multinomial model follows. The Poisson distribution is introduced in Sect. 5.3. Section A.5 gives the solutions to the problems suggested to the reader.

5.1 The Binomial Distribution

Historically, the binomial distribution is the root of all probability distributions [1]. It served to define precisely the frequentist notion of probability. It was the first example used to discuss statistical inference [2, 3].

The binomial model describes the distribution of the results obtained from a simple alternative. Let η be the probability of winning in drawing lots, and $1-\eta$ the probability of drawing a blank. This describes the odds when one lot is drawn. The probability of winning x times when N lots are drawn is

$$p(x|\eta) = \binom{N}{x} \eta^x (1-\eta)^{N-x} ; \quad x = 0, 1, 2, \ldots, N;\; 0 \le \eta \le 1 . \tag{5.1}$$

This expression is normalised; that is, one has

$$\sum_{x=0}^{N} p(x|\eta) \;=\; 1 \tag{5.2}$$

by virtue of the binomial theorem. This theorem states

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_5

$$(\eta_1 + \eta_2)^N = \sum_{x=0}^{N} \binom{N}{x} \eta_1^x \eta_2^{N-x} . \tag{5.3}$$

The binomial distribution is represented in Fig. 5.1 for several values of the parameter η.

We want to evaluate the moments $\overline{x}$ and $\overline{x^2}$ of the binomial distribution. In order to obtain $\overline{x}$, the identity

$$x = \frac{\partial}{\partial\zeta}\zeta^x \bigg|_{\zeta=1} \tag{5.4}$$

is useful. It shows that

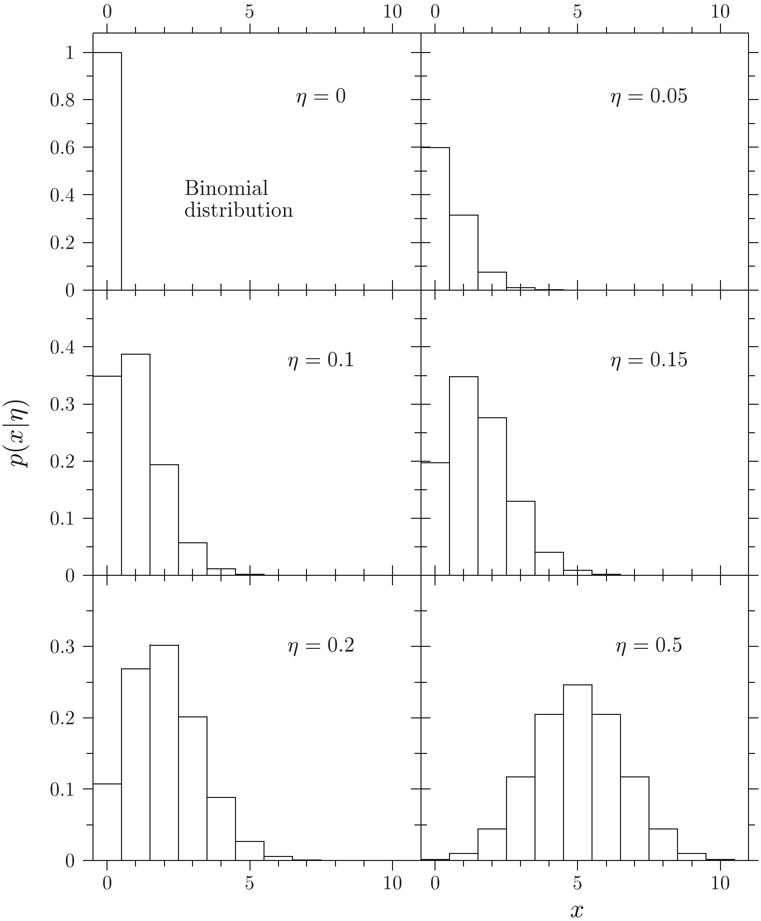


Fig. 5.1 The binomial distribution (5.1) for several mean values η

$$\begin{aligned}\overline{x} &= \sum_{x=0}^{N} x\, p(x|\eta) \\ &= \frac{\partial}{\partial\zeta} \sum_x \binom{N}{x} \zeta^x \eta^x (1-\eta)^{N-x} \bigg|_{\zeta=1} \\ &= \frac{\partial}{\partial\zeta} (\zeta\eta + 1 - \eta)^N \bigg|_{\zeta=1} \\ &= N\eta\,. \end{aligned} \tag{5.5}$$

The third line of this equation is obtained with the help of the binomial theorem. The derivative with respect to ζ must be taken before η is set equal to 0 or 1. The result of the last line, however, is meaningful for $\eta = 0, 1$ too.

The last line of Eq. (5.5) illustrates the frequency interpretation of probability: from N drawings one expects to obtain a number of hits equal to N multiplied with the probability to win in one drawing. For this reason, the parameter η has been called a probability.

The frequency interpretation can be verified by drawing a large number of lots. The fluctuations of the result should then become small as compared to the expectation value (5.5). We can estimate the fluctuations by the variance of x. In order to obtain the variance, we use the identity

$$x(x-1) = \frac{\partial^2}{\partial\zeta^2} \zeta^x \bigg|_{\zeta=1} \tag{5.6}$$

and proceed as in (5.5). This yields

$$\overline{x^2} = N(N-1)\eta^2 + N\eta \tag{5.7}$$

and

$$\begin{aligned}\text{var}(x) &= \overline{x^2} - \overline{x}^2 \\ &= N\eta(1-\eta)\,. \end{aligned} \tag{5.8}$$

The normalised variance

$$\frac{\text{var}(x)}{\overline{x}^2} = \frac{1-\eta}{N\eta}\,; \qquad \eta > 0 \tag{5.9}$$

is of the order of N^{-1}. This was called the law of large numbers by de Moivre and Laplace [4].

In the next section, the simple alternative is generalised to the M-fold alternative.

5.2 The Multinomial Distribution

The multinomial distribution generalises the binomial distribution to an alternative that offers M possibilities instead of two. We call this an M-fold alternative. The possible outcomes of one drawing are called the bins $k = 1 \dots M$. The probability that the kth bin is obtained in one drawing is represented by η_k An "alternative" means that one of the bins must be chosen; therefore

$$\sum_{k=1}^{M} \eta_k = 1 \,. \tag{5.10}$$

The probability of hitting the first bin x_1 times and the second bin x_2 times, and so on, in N drawings, is

$$p(x_1, \dots, x_{M-1} | \eta_1, \dots, \eta_{M-1}) \;=\; N! \prod_{k=1}^{M} \frac{\eta_k^{x_k}}{x_k!} \,; \quad 0 \le \eta_k \le 1 \,. \tag{5.11}$$

This is the multinomial model. Every x_k may take the values $x_k = 0, 1, \dots, N$ such that the set $x = (x_1, \dots, x_M)$ has the property

$$\sum_{k=1}^{M} x_k = N \,. \tag{5.12}$$

This means that we can express x_M in terms of the variables $x = (x_1, \dots, x_{M-1})$. An analogous situation applies for the hypothesis parameters $\eta = (\eta_1, \dots, \eta_{M-1})$. The quantity η_M is defined by (5.10). The multinomial model (5.11) is normalised according to

$$\sum_{x} p(x|\eta) = 1 \,. \tag{5.13}$$

The normalisation is a consequence of the multinomial theorem which states that

$$\left(\sum_{k=1}^{M} \eta_k \right)^{N} = N! \sum_{x} \prod_{k=1}^{M} \frac{\eta_k^{x_k}}{x_k!} . \tag{5.14}$$

Note that $\eta_k^0 = 1$ and that $0! = 1$ according to Sect. B.4.

We want to calculate the moments $\overline{x_k}$ and $\overline{x_k x_{k'}}$ of the multinomial distribution. Similarly to what we did in Sect. 5.1, the identity (5.4) yields

$$\begin{aligned}\overline{x_k} &= \frac{\partial}{\partial\zeta}\sum_x \zeta^{x_k}\, p(x|\eta)\bigg|_{\zeta=1} \\ &= \frac{\partial}{\partial\zeta}\,(\zeta\eta_k - \eta_k + 1)^N\bigg|_{\zeta=1} \\ &= N\eta_k\,. \end{aligned} \tag{5.15}$$

The second line of the equation is obtained with the help of the multinomial theorem. Again the last version corresponds to the frequency interpretation of probability.

To obtain the second moments, the identities (5.4) and (5.6) are used once more. We find

$$\overline{x_k x_{k'}} = N(N-1)\,\eta_k\eta_{k'} + \delta_{kk'}N\eta_k \tag{5.16}$$

which is the generalisation of (5.7). The interested reader should work out the details. The quantity

$$\overline{x_k x_{k'}} - \overline{x_k}\,\overline{x_{k'}} = N\eta_k(\delta_{kk'} - \eta_{k'}) \tag{5.17}$$

is called the correlation between x_k and $x_{k'}$ (see Sect. 4.1.2). The x_k are correlated with each other because they respect the sum rule (5.12). The ratio

$$\frac{\overline{x_k x_{k'}} - \overline{x_k x_{k'}}}{\overline{x_k}\,\overline{x_{k'}}} = \frac{\delta_{kk'} - \eta_{k'}}{N\eta_{k'}}\,; \qquad \eta_k, \eta_{k'} \neq 0\,, \tag{5.18}$$

is called a correlation coefficient. It is a generalisation of (5.9) and is of the order of N^{-1}.

5.3 The Poisson Distribution

The Poisson[1] distribution gives the probability of counting x events when nothing but the mean value λ is specified. It can be obtained from the binomial model if the number N of trials is taken to infinity and the probability of success in one trial is $\eta = \lambda/N$. In this limit, the binomial distribution tends towards

$$p(x|\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda)\,; \qquad x = 0, 1, 2, \ldots, \; 0 \le \lambda\,; \tag{5.19}$$

which is the Poisson distribution. The proof is left to the interested reader. From the Taylor expansion of the exponential function, one can recognise that p is normalised according to

[1] Siméon Denis Poisson, 1781–1840, French mathematician and physicist. He contributed to the theory of electrostatic and magnetic potentials, the theories of Fourier series and differential equations, and to statistics.

$$\sum_{x=0}^{\infty} p(x|\lambda) = 1 . \tag{5.20}$$

With the help of (5.4), one can verify the mean value

$$\overline{x} = \lambda . \tag{5.21}$$

With the help of (5.6), one finds

$$\overline{x^2} = \lambda(\lambda + 1) . \tag{5.22}$$

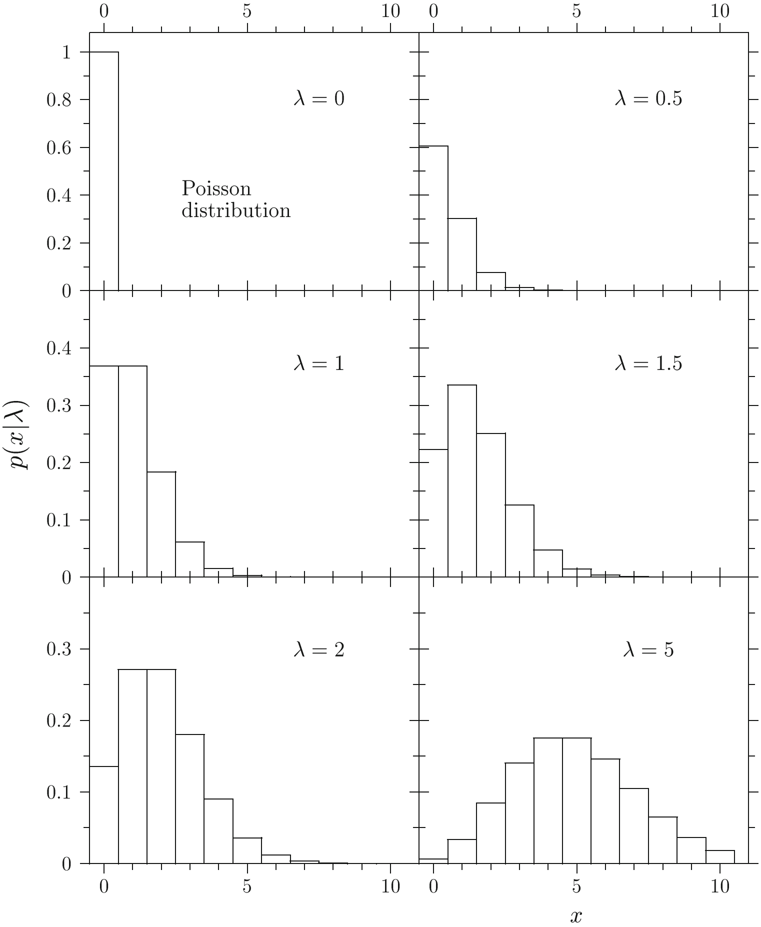


Fig. 5.2 The Poisson distribution (5.19) for several values of the mean λ

Hence, both the mean value and the variance of x are equal to λ. The Poisson distribution is given in Fig. 5.2 for several values of λ.

We note that the model of the histogram is given by the count rates $x_1, \dots, x_M$ in M bins when x_k follows the Poisson distribution with the mean λ_k for $k = 1, \dots, M$. This means that the count rates x_k are independent of each other and that their joint distribution is the product of the Poisson distributions in every bin. The histogram together with its posterior distribution is considered in Sect. 13.3.

References

1. J. Bernoulli. *Wahrscheinlichkeitsrechnung*. Harri Deutsch, Thun, 1999. Reprint of vols. 107 and 108 of the series *Ostwalds Klassiker der exakten Wissenschaften.* The Latin original appeared in 1713 under the title of *Ars conjectandi*
2. T. Bayes, An essay towards solving a problem in the doctrine of chances. Phil. Trans. Roy. Soc., **53**, 330–418 (1763). Reprinted in Biometrika 45, 293–315 (1958) and in ‘Studies in the History of Statistics and Probability’, ed. by E.S. Pearson, M.G. Kendall, C. Griffin and Co. Ltd., London (1970) and in ’Two papers by Bayes with commentaries’, ed. by W.E. Deming, Hafner Publishing Co., New York (1963)
3. P.S. De Laplace. Mémoire sur la probabilité des causes par les événements. *Mém. de math. et phys. présentés à l’Acad. roy. des sci.*, **6**, 621–656 (1774). Reprinted in [5], vol. 8, pp 27–65. An English translation can be found in [6]
4. P.S. De Laplace. *A Philosophical Essay on Probabilities* (Dover, New York, 1951). Original title *Essay philosophique sur les probabilités*
5. P.S. De Laplace, *Œuvres complètes de Laplace* (Gauthier-Villars, Paris, 1886–1912). 14 volumes
6. S.M. Stigler, Laplace’s 1774 memoir on inverse probability. Stat. Sci. **1**, 359–378 (1986)

Chapter 6
Form Invariance I

Ignorance about the hypothesis ξ cannot in general be expressed by the uniform prior. This is a consequence of the transformation law of a probability density discussed in Sect. 2.2. Under a reparameterisation of the hypothesis, the uniform density generally changes into another one that is no longer uniform. If there were a distribution invariant under all transformations, it would be the universal ignorance prior. Such a distribution does not exist. However, there are distributions that remain invariant under a group of transformations. If the group "describes" a symmetry of the model p, we consider the invariant distribution to be the prior. In more technical language, we can say that if the group of transformations is the symmetry group of the model, the prior is required to be invariant under the group. Symmetries and, in particular, the symmetries of form-invariant models are discussed below.

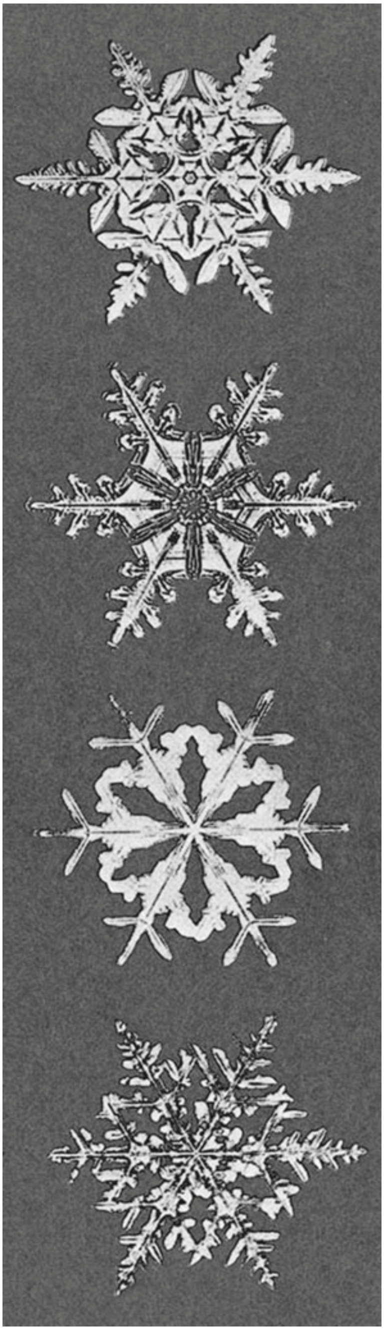
The present chapter owes much to the work of Hartigan [1, 2], Stein [3], and Jaynes [4]; their work was extended by by Villegas [5–8]. For context, consult the review article by [11].

What is symmetry? The snowflake crystals [12, 13] in Fig. 6.1 are all different from each other but every one has the property: if one rotates the crystal by 60^0 or $\pi/3$ rad, its appearance is not changed. This is not the only rotation leaving the crystal invariant. A multiple $n\pi/3$ of the elementary rotation again leads to the same appearance. Here, n is an integer number, and rotations in the positive as well as the negative sense are admitted. Also admitted is $n = 0$, that is, the identity. Actually, one need only consider the rotations given by n modulo 6, because the rotation by 2π is equivalent to the identity. Hence, there are 6 transformations with $n = 0 \ldots 5$ of the snowflake which leave its appearance unchanged. This "group" of transformations is the mathematical essence of the symmetry that we perceive in Fig. 6.1. Many more beautiful examples can be found in the booklet [14] by Hermann Weyl.

The notion of a mathematical group is defined in Sect. 6.1. The symmetry of form invariance is introduced in Sect. 6.2. The invariant measure (i.e. the prior distribution) is defined in Sect. 6.3. In Sect. 6.4, we compare the invariant measure with the measure of differential geometry. Finally, in Sect. 6.5, it is shown that form invariance of the

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_6

Fig. 6.1 Snowflakes are invariant under a symmetry group that contains six different rotations. This figure is due to the collection of [12]; see also p. 125 of [13]



model $p(x|\xi)$ entails form invariance of the posterior $P(\xi|x)$. In Sect. 6.6, we define the maximum likelihood estimator of the parameter ξ of a model $p(x|\xi)$, and we show in Sect. 6.7 that this estimator is the sufficient statistic for the parameter. Section A.6 gives the solutions to the problems suggested to the reader.

6.1 Groups

The mathematical groups $\mathcal{G}$ considered in the present context are sets of transformations G_ξ of the events x. A group has the following four properties:

1. If G_ξ and $G_{\xi'}$ are in $\mathcal{G}$, then the product $G_\xi G_{\xi'}$ is defined and is contained in $\mathcal{G}$. This product is the transformation obtained by first applying $G_{\xi'}$ and then G_ξ.
2. The product is associative; that is, one has $G_\xi(G_{\xi'}G_{\xi''}) = (G_\xi G_{\xi'})G_{\xi''}$.
3. The identity $\mathbf{1}$ is in $\mathcal{G}$. It is also called the unit element of $\mathcal{G}$ and carries the index ϵ.
4. For any element G_ξ, the group contains an inverse G_ξ^{-1}. The inverse has the property $G_\xi^{-1}G_\xi = \mathbf{1}$.

Note that these axioms do not require the commutativity,

$$G_\xi G_{\xi'} = G_{\xi'} G_\xi \ . \tag{6.1}$$

Compare Chap. 1 of [15]. Below in the present section, an example is given of a group with elements that do not commute. If (6.1) holds for any pair of transformations in $\mathcal{G}$, the group is called Abelian.[1]

The properties of a group entail that every element of the group can be considered the "origin" of the group: let ξ run over all values of the group parameter and G_τ be an arbitrary but fixed element of the group. Then

$$G_\rho = G_\xi G_\tau \tag{6.2}$$

runs over all elements of the group exactly once; that is, the multiplication by G_τ is a one-to-one mapping of the group onto itself. The proof is left to the reader. In a generalised sense, this mapping can be called a "shift" of the index ξ.

The symmetries of conditional probabilities considered below are not described by finite groups nor by groups with a countable number of elements but rather by groups with a manifold of elements. That is to say, the index ξ that labels the transformations G_ξ in the group $\mathcal{G}$, is a real number or even a vector of real numbers. We call this a Lie group.[2] Inasmuch as the symmetries we consider are given by Lie groups,

[1] Niels Henrik Abel (1802–1829), Norwegian mathematician. He investigated the question of which algebraic equations are solvable. He founded the general theory of integrals of algebraic functions.

[2] Marius Sophus Lie (1842–1899), Norwegian mathematician. He developed the theory of continuous transformation groups which nowadays carry his name.

they unfortunately cannot be visualised as nicely as the symmetry of snowflakes in Fig. 6.1.

A simple example of a Lie group is given by the set of transformations

$$G_\phi = \begin{pmatrix} \cos\phi\,, & -\sin\phi \\ \sin\phi\,, & \cos\phi \end{pmatrix}\,,$$
$$0 \le \phi < 2\pi\,, \tag{6.3}$$

which rotate the plane by the angle of ϕ about the origin. The symmetry of the circle is described by this group or - in other words - this is the symmetry group of the circle in the sense that the circle does not change its appearance when it is rotated about its center. See Fig. 6.2. As is well known,

$$a(\phi) = G_\phi \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{6.4}$$

is a parametric representation of the circle. Here, a is the two-dimensional vector

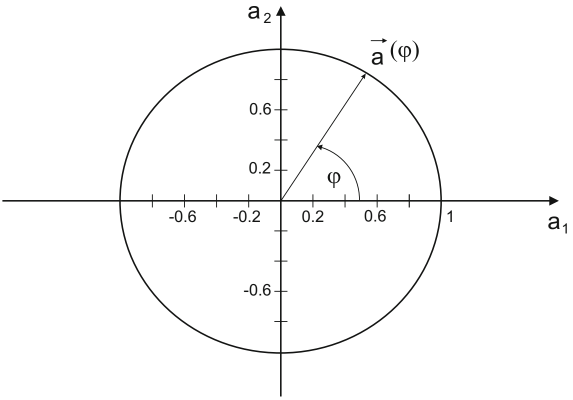
$$a = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}\,. \tag{6.5}$$

Why do the rotations (6.3) with $0 \le \phi \le \pi$ no longer form a group? The domain of definition of the group parameter is important.

Another example of a Lie group is given by the hyperbolic transformations

$$G_\phi = \begin{pmatrix} \cosh\phi\,, & \sinh\phi \\ \sinh\phi\,, & \cosh\phi \end{pmatrix}\,,$$
$$-\infty < \phi < \infty\,. \tag{6.6}$$

Fig. 6.2 The parametric representation (6.4) of the circle. The transformations (6.3) form the symmetry group of the circle



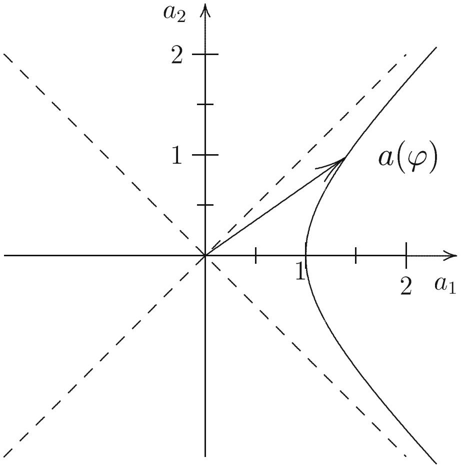


Fig. 6.3 The parametric representation (6.4) of the hyperbola. The transformations (6.6) form the symmetry group of the hyperbola

The interested reader may show that (6.4) is a parametric representation of the hyperbola of Fig. 6.3, if G_ϕ is taken from (6.6). The group of transformations (6.6) is the symmetry group of the hyperbola.

However, the hyperbolic symmetry (6.6) will not occur among the form- invariant statistical models defined below in Sect. 6.2 because we consider transformations that conserve the normalisation (2.3) and (2.4) of a given statistical model. We show in Sect. 8.4 that this requires conserving the absolute value $a^\dagger a$ of the vector a in Eq. (6.5). Here, the dagger † denotes the conjugate or transpose of a. In other words we consider orthogonal transformations. They are defined by the property

$$G_\phi G_\phi^\dagger = \mathbf{1} \tag{6.7}$$

The rotations (6.3) are orthogonal; the hyperbolic transformations (6.6) are not.

The abstract structure of a group is contained in the multiplication function $\Phi = \Phi(\xi'; \xi)$ which labels the product

$$G_\Phi = G_\xi G_{\xi'} \, . \tag{6.8}$$

According to general usage in group theory, the order of the arguments of Φ is the reverse of the order of the corresponding operators in (6.8) (see e.g. Chap. 8 of [15] or Chap. 4 of [16]). Thus, in terms of the multiplication function, axiom 2 above, stipulating that the product of the group elements is associative, reads

$$\Phi(\Phi(\xi''; \xi'); \xi) = \Phi(\xi''; \Phi(\xi'; \xi)) \, . \tag{6.9}$$

The reader may show that for both of the groups (6.3) and (6.6), the multiplication function is

$$\Phi(\phi', \phi) = \phi + \phi' \, . \tag{6.10}$$

Hence, the groups are Abelian. In the case of the circle, the domain of definition of ϕ', ϕ, Φ, given in Eq. (6.3), must be understood modulo 2π.

The translations

$$G_\xi \, x = x + \xi \ , \tag{6.11}$$

where x and ξ are real, form an Abelian group. The multiplication function corresponds to (6.10). The interested reader should show that the dilations

$$G_\sigma \, x = \sigma x, \qquad 0 < \xi < \infty \, , \tag{6.12}$$

form an Abelian group, and that

$$\Phi(\sigma'; \sigma) = \sigma \, \sigma' \tag{6.13}$$

is its multiplication function. Note that groups with one parameter are always Abelian. This is explained in Sect. 8.4.

The combination

$$G_{\xi,\sigma} \, x = \xi + \sigma x \tag{6.14}$$

of translation and dilation, where

$$-\infty < \xi < \infty \quad \text{and} \quad 0 < \sigma < \infty \ , \tag{6.15}$$

presents an example of a non-Abelian group. It contains the subgroup of the translations $G_{\xi,1}$ as well as the subgroup of the dilations $G_{0,\sigma}$. The notation $G_{\xi,\sigma}$ means

$$G_{\xi,\sigma} = G_{0,\sigma} G_{\xi,1} \ , \tag{6.16}$$

that is, the translation by the amount ξ is performed first and yields the mapping

$$x \longrightarrow G_{\xi,1} x = \xi + x \ . \tag{6.17}$$

The subsequent dilation of x leads to

$$x \longrightarrow G_{0,\sigma} G_{\xi,1} x = \xi + \sigma x \ . \tag{6.18}$$

The reader may work out the multiplication function[3]

$$\Phi(\xi', \sigma'; \xi, \sigma) = (\xi' + \xi\sigma'; \sigma\sigma') \ . \tag{6.19}$$

[3] In the first edition of the present book, a confusion has occurred about the order of the operations of translation and dilation. The multiplication function given in Eqs. 6.15 and (A.81) of that edition is incorrect, whereas Eq. 7.16 is correct. The present Eqs. (6.16)–(6.18) make clear that every operation acts on the event variable x, not on the result of the foregoing operation. Thus $G_{0,\sigma}$, acting on $\xi + x$, generates $\xi + \sigma x$ not $\sigma(\xi + x)$.

of the transformations defined by (6.14). The reader may convince herself that reversing the order of translation and dilation changes both the definition (6.14) and the multiplication function (6.19). For later use we note that the inversion $G^{-1}_{\xi,\sigma}$ interchanges the order of the operations and replaces ξ, σ by the inverse translation $\overline{\xi} = -\xi$ and the inverse dilation $\overline{\sigma} = \sigma^{-1}$. We obtain

$$\begin{aligned} G^{-1}_{\xi,\sigma}\, x &= G^{-1}_{\xi,1} G^{-1}_{0,\sigma}\, x \\ &= G_{\overline{\xi},1} G_{0,\overline{\sigma}} \\ &= \frac{x-\xi}{\sigma} \, . \end{aligned} \tag{6.20}$$

Any group of transformations G_ξ acting on x, induces a group of transformations $\tilde{G}_\xi$ acting on the domain of definition of the index (or the set of indices) ξ. Let ρ be one such index; we define the transformation $\tilde{G}_\xi$ by the mapping

$$\tilde{G}_\xi\, \rho = \Phi(\rho; \xi) \, . \tag{6.21}$$

This is a transformation of the domain of definition of the indices. The interested reader should convince himself of this. The set of transformations obtained by letting ξ run over all values of the group parameter is a group $\tilde{\mathcal{G}}$ that is "isomorphic" to $\mathcal{G}$; that is, it has the same multiplication function as the group $\mathcal{G}$. The proof is left to the reader. Therefore, and because it will always be clear from the context whether we mean the transformation G_ξ of x or the transformation $\tilde{G}_\xi$ of the group indices, the tilde will henceforth be omitted.

As a consequence of this isomorphism, there is exactly one transformation G_ξ which takes a given τ to a given ρ, so that

$$\rho = G_\xi \tau \, . \tag{6.22}$$

Letting ξ run over all indices, one obtains a reparameterisation of the group $\mathcal{G}$. The mapping $\tau \to \rho$ can be called a "shift" of the index ξ.

For more examples of Lie groups and a deeper insight into the fascinating realm of group theory, consult textbooks such as [15–22].

In the next section, the symmetry of form invariance of a conditional probability is defined. It is not tied to a definite group. The above Lie groups and other ones may occur as the symmetry group.

6.2 The Symmetry of Form Invariance

We consider a model $p(x|\xi)$ where both the event x and the parameter ξ, are real numbers or sets of real numbers. The model is called form invariant if the domain of definition of the parameter ξ labels a group $\mathcal{G}$ of transformations of x such that

p remains invariant under simultaneous application of a group element G_ρ to x and to ξ,

$$p(x|\xi) \;=\; p(G_\rho x|G_\rho \xi) \left| \frac{\partial G_\rho \, x}{\partial x} \right| \,. \tag{6.23}$$

Because p is a probability density the transformation of x entails the multiplication with the determinant of the Jacobian matrix of the transformation.

This definition of form invariance is equivalent to the statement that for every ξ the distribution p emerges from one common "form" $w(x)$ such that

$$p(x|\xi) = w(G_\xi^{-1} \, x) \left| \frac{\partial G_\xi^{-1} \, x}{\partial x} \right| \,, \tag{6.24}$$

and the set of transformations G_ξ is the group $\mathcal{G}$. The interested reader should show that w is normalised and can be stated as

$$w(x) = p(x|\xi = \epsilon) \,. \tag{6.25}$$

Remember that G_ϵ is the unit element of the group $\mathcal{G}$.

In Sect. 2.3, the simplest examples of form invariance were introduced, namely the models with the structure

$$p(x|\xi) = w(x - \xi) \,, \qquad -\infty < x, \xi < \infty \,, \tag{6.26}$$

depending on the difference of x and ξ (and only on this difference). One also says that x and ξ are defined "on a common scale". The Gaussian of (2.16) is such a model. The group (6.11) of translations is its symmetry group. The measure μ on the common scale must be uniform (i.e. constant) so that the difference $x - \xi$ refers to a distance which is independent of the values of x and ξ. The "invariant measure" μ, introduced in the next section, is indeed uniform for the group of translations. In Eq. (6.26) the negative sign of ξ recalls the inverse G_ξ^{-1} in (6.24). Examples of other symmetry groups are given in Chap. 7.

When N events $x_1, \ldots, x_N$ from the same form-invariant distribution $q(x_k|\xi)$ are observed then the joint distribution

$$p(x_1, \ldots, x_N|\xi) = \prod_{k=1}^{N} q(x_k|\xi) \tag{6.27}$$

is form invariant too, and the symmetry group is independent of N . This is obvious because the symmetry property

$$p(G_\rho x_1, \ldots, G_\rho x_N|G_\rho \xi) = \prod_{k=1}^{N} \left(q(G_\rho x_k|G_\rho \xi) \left| \frac{\partial G_\rho \, x_k}{\partial x_k} \right| \right) \tag{6.28}$$

is an N-fold repetition of the elementary symmetry (6.23). The transformations

$$\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \longrightarrow \begin{pmatrix} G_\xi^{-1} x_1 \\ \vdots \\ G_\xi^{-1} x_N \end{pmatrix} \tag{6.29}$$

of the vector x form a group which is isomorphic to $\mathcal{G}$. Up to this isomorphism, the symmetry group does not depend on N.

A Lie group defines an "invariant measure" in the space of the group parameter. This is described in the next section.

6.3 The Invariant Measure

There is a measure $\mu(\xi)$ which is left unchanged under all transformations $G_\rho \in \mathcal{G}$ that is,

$$\mu(\xi) = \mu(G_\rho \xi) \left| \frac{\partial G_\rho \xi}{\partial \xi} \right| . \tag{6.30}$$

With this measure, the volume

$$V = \int_{\mathcal{A}} \mu(\xi)\, d\xi \tag{6.31}$$

of an area $\mathcal{A}$ in the space of ξ does not change when $\mathcal{A}$ is "shifted" to $G_\rho \mathcal{A}$ by any G_ρ of the group. By the "shift" we mean that the image of every point ξ in $\mathcal{A}$ is obtained by applying G_ρ to it; in other words, $\xi \in \mathcal{A}$ is mapped onto $\xi' = \Phi(\xi; \rho)$. The mapping $\mathcal{A} \to G_\rho \mathcal{A}$ is then a generalisation of the translation of an interval in the space of real numbers. A reasonable definition of the length of the interval must be invariant under translations. Indeed, one has

$$\begin{aligned} V &= \int_{G_\rho \mathcal{A}} \mu(G_\rho \xi)\, dG_\rho \xi \\ &= \int_{G_\rho \mathcal{A}} \mu(\xi)\, d\xi \ , \end{aligned} \tag{6.32}$$

when μ satisfies the invariance property (6.30). The first line of this equation is obtained from (6.31) by a change of the integration variables. The second line results from the invariance property (6.30).

Because ϵ is the value of the group parameter that labels the unit element, the invariant measure can be written

$$\mu(\xi) = \mu(\epsilon) \left| \frac{\partial \Phi(\tau; \xi)}{\partial \tau} \right|^{-1}_{\tau=\epsilon} \tag{6.33}$$

as we show in Sect. C.1. In group theory, this expression is called the "left" invariant Haar measure [23]. It vanishes nowhere; in analogy to the transformation (6.21) of ξ, the mapping $\tau \to \Phi(\tau; \xi)$ is a transformation of τ. Therefore the Jacobian in (6.33) does not vanish. Hence, the invariant measure exists and is everywhere different from zero. The factor $\mu(\epsilon)$ is an arbitrary positive number.

There exists also a "right" invariant Haar measure. It satisfies the equation

$$\begin{aligned} V &= \int_{\mathcal{A}} \mu_r(\xi) \mathrm{d}\xi \\ &= \int_{\mathcal{A}'} \mu_r(\xi)\, d\xi \ , \end{aligned} \tag{6.34}$$

if $\mathcal{A}'$ is obtained by mapping $\xi \in \mathcal{A}$ onto $\xi' = G_\xi \rho = \Phi(\rho; \xi)$. The right invariant measure has been proposed [3, 7, 24–35] as a prior distribution because it seemed to allow a frequentist interpretation of the Bayesian area $\mathcal{B}(K)$. However, in the context of Bayesian statistics, a decision between the left and right invariant measures is not possible, because the label ξ of the group elements G_ξ can be redefined such that the right invariant measure (in terms of the original labels) is obtained from expression (6.33). This transition from the left to the right invariant measure occurs when a reparameterisation interchanges the order of the operations implied by the index ξ. Remember that we are assigning to ξ a dimension higher than 1, because a 1-dimensional ξ implies an Abelian group. The order of operations labelled by ξ is inverted, for example, by reparameterising according to $\xi \to \overline{\xi}$. Compare Reference [3]. Now, the definition of the symmetry group $\mathcal{G}$ can be taken either from Eq. (6.23) or from Eq. (6.24); that is, it can be given by the set of transformations G_ξ implying the multiplication function Φ or by the set of transformations G_ξ^{-1} implying the multiplication function $\overline{\Phi}$. The left invariant measure (6.33) will be interchanged with the (former) right invariant measure. We have no argument to prefer one over the other.

Here, we take the definition of the symmetry group from Eq. (6.23) and identify the left invariant measure (6.33) with the prior distribution postulated by Bayes. Henceforth, we simply speak of "the invariant measure". Examples of invariant measures are given in Chap. 7.

In the next section, an alternative definition of the measure is taken from differential geometry. In Chap. 9, it is shown that the invariant measure agrees with the geometric one up to the constant factor $\mu(\epsilon)$ in Eq. (6.33). This factor is left free in group theory; it is well defined in differential geometry.

6.4 The Geometric Measure

Consider a model $p(x|\xi)$, where x assumes only two possible values, $x = 0$ and $x = 1$. The normalisation $p(0|\xi) + p(1|\xi) = 1$ entails that we may write $p(0|\xi) = \sin^2 \xi$ and $p(1|\xi) = \cos^2 \xi$. Let us replace the probabilities by the probability "amplitudes"

$$a_x(\xi) = \sqrt{p(x|\xi)} \tag{6.35}$$

and define the transformation of the events x by the linear transformation (6.3) of the vector a so that Eq. (6.4) holds; that is,

$$a(\xi) = \begin{pmatrix} \cos \xi \\ \sin \xi \end{pmatrix} . \tag{6.36}$$

This is a parameter representation of a curve - a circle - in the plane. Thus a_0, a_1 are Cartesian coordinates in the plane. When ξ runs over the interval $\mathcal{A}$, then $a(\xi)$ runs over an arc with the length

$$V = \int_{\mathcal{A}} d\xi \left(\left(\frac{da_0}{d\xi} \right)^2 + \left(\frac{da_1}{d\xi} \right)^2 \right)^{1/2} . \tag{6.37}$$

This is a well-known result of simple differential geometry. It means that the geometric measure μ_g on the curve is given by

$$\begin{aligned} \left(\mu_g(\xi) \right)^2 &= \left(\frac{da_0}{d\xi} \right)^2 + \left(\frac{da_1}{d\xi} \right)^2 \\ &= \left(\frac{\partial}{\partial \xi} a(\xi) \right)^\dagger \left(\frac{\partial}{\partial \xi} a(\xi) \right) . \end{aligned} \tag{6.38}$$

Here, we have written the dagger † to denote the transpose of a vector. Taking $a(\xi)$ from (6.36) one finds

$$\mu_g(\xi) \equiv 1 . \tag{6.39}$$

This does not contradict the invariant measure (6.33) because, in analogy to (6.10), the multiplication function is $\Phi(\xi, \xi) = \xi + \xi$ and we find

$$\mu(\xi) \equiv \mu(\epsilon) . \tag{6.40}$$

However, the invariant measure (6.33) is defined up to a factor only while the geometric measure yields a well-defined value of μ. The invariant measure agrees with the geometric measure in that both are constant functions of ξ (in the present case). This holds for every linear transformation of ξ. A unique definition of μ requires, for example, the additional condition that ξ must be transformed such that $\mu(\xi) \equiv 1$.

Then Eq. (6.37) yields $V = 1$ for the length of the straight line between

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix} ,$$

that is, μ is in agreement with the notion of length in a two-dimensional Cartesian space.

Because of the identity

$$\frac{\partial}{\partial \xi} a_x = \frac{1}{2} \sqrt{p(x|\xi)} \frac{\partial}{\partial \xi} \ln p(x|\xi) , \tag{6.41}$$

the geometric measure can be written in terms of probabilities, without explicitly introducing probability amplitudes. Equation (6.41) allows us to express the geometric measure in the form of

$$\mu_g^2(\xi) = \frac{1}{4} \sum_x p(x|\xi) \left(\frac{\partial}{\partial \xi} \ln p(x|\xi) \right)^2 . \tag{6.42}$$

This can be rewritten as

$$\mu_g^2(\xi) = -\frac{1}{4} \sum_x p(x|\xi) \frac{\partial^2}{\partial \xi^2} \ln p(x|\xi) . \tag{6.43}$$

The last form is a consequence of the fact that $p(x|\xi)$ is normalised to unity for every value of ξ. The interested reader is asked to prove it.

In Chap. 9, the geometric measure (6.38) will allow us to define the prior distribution of models $p(x|\xi)$ that are not form invariant. The idea to use the geometric measure as the prior distribution was introduced by Kass [36, 37] and Amari [38].

The introduction of the probability amplitudes (6.35) has allowed us to ascribe the linear representation (6.4) to the symmetry of the binomial model introduced in the text at the beginning of this section. Chapter 8 extends the subject of linear representations to the case where x is a continuous variable.

Form invariance of a model p entails an analogous symmetry of the posterior P. This is described in the following section.

6.5 Form Invariance of the Posterior Distribution

As a consequence of the symmetries (6.23) and (6.30) of p and μ, the distribution $m(x)$ is invariant under the group $\mathcal{G}$, thus it is an invariant measure in the space of x. One sees this by rewriting the definition (2.5) of m in several steps:

$$\begin{aligned}
m(x) &= \int p(x|\xi)\mu(\xi)d\xi \\
&= \int p(x|\xi)\mu(G_\rho\,\xi)\,dG_\rho\,\xi \\
&= \left[\int p(G_\rho x|G_\rho\,\xi)\mu(G_\rho\xi)\,dG_\rho\,\xi\right]\left|\frac{\partial G_\rho x}{\partial x}\right| \\
&= \left[\int p(G_\rho x|\xi)\mu(\xi)d\,\xi\right]\left|\frac{\partial G_\rho\,x}{\partial x}\right| \\
&= m(G_\rho\,x)\left|\frac{\partial G_\rho\,x}{\partial x}\right| \,.
\end{aligned} \tag{6.44}$$

One proceeds from the first to the second line of this equation by the invariance property of $\mu(\xi)$. The invariance of $p(x|\xi)$ yields the third line, and a change of the integration variable the fourth line. By use of the definition of m, one obtains the last line. Hence, m is invariant under all transformations in $\mathcal{G}$; that is, m is an invariant measure.

However, m is not *the* invariant measure in the space of x because in general it is impossible to map x on ξ one to one. Not even the dimension of x has anything to do with the dimension of ξ. Indeed, for N events the dimension of x is at least N whereas the number n of hypothesis parameters is independent of N. The symmetries of p and μ and m entail

$$P(\xi|x) = P(G_\rho\,\xi|G_\rho\,x)\left|\frac{\partial G_\rho\,\xi}{\partial \xi}\right| \,. \tag{6.45}$$

The proof is left to the reader. Hence, P is form invariant.

We have derived the invariance of the measure m under the transformations in $\mathcal{G}$, but we have not yet shown that m exists in the sense that the integral (2.5) exists which defines m. This is discussed in Sect. 6.6 which deals with the existence of the ML estimator. Section 6.7 reveals an important property of the ML estimator: the function $\xi^{\mathrm{ML}}(x)$ is the sufficient statistic of the model $p(x|\xi)$.

6.6 The Maximum Likelihood Estimator

In Chap. 2, after Eq. (2.24), we noticed that a well-known model $p(x|\xi)$ possesses a maximum as a function of the parameter ξ. This value of ξ is called the estimator ξ^{ML}, more precisely, the maximum likelihood (ML) estimator, of ξ given the event x. In Eq. (3.13) of Chap. 3 the likelihood function was defined as

$$L(\xi|x) = \frac{P(\xi|x)}{\mu(\xi)} \,. \tag{6.46}$$

This expression transforms like a function with respect to ξ and x, and, by virtue of the symmetries of P and μ, it has the symmetry property

$$L(x|\xi) = L(G_\rho x|G_\rho \xi)\,. \tag{6.47}$$

We require that for any given x the likelihood function have a unique absolute maximum as a function of ξ. The value of ξ, where L is maximal, is denoted $\xi^{\mathrm{ML}} = \xi^{\mathrm{ML}}(x)$. In Sect. 3.4, we have argued that form invariance entails the existence of the ML estimator because $L(\xi|x)$ cannot be independent of ξ, not even piecewise. For $|\xi| \to \infty$ it must vanish because P is normalised. Therefore it must have a maximum. Any finite domain of definition of ξ can be transformed to an infinite domain of definition. See also the argument following Eq. (4.10). We are aware of the fact that these arguments at best are a sketch of a proof for the existence of ξ^{ML}. Therefore, to the best of our present knowledge, the existence of the ML estimator is a requirement in addition to the form invariance of $p(x|\xi)$.

If the ML estimator exists, then the value $L(\xi^{\mathrm{ML}}(x)|x)$ of L at $\xi = \xi^{\mathrm{ML}}$ is independent of x. This is a consequence of the symmetry property (6.47) of L as we show in the following discussion.

Given the existence of the ML estimator, one can introduce classes of events: we consider all values of x that lead to the same $\xi^{\mathrm{ML}}(x)$, as belonging to the same class of values. The class is defined and called by the value of ξ^{ML}. The classes are equivalent in the sense that they can be mapped onto each other. For this, we consider the epsilon-class $\{x|\xi^{\mathrm{ML}}(x) = \epsilon\}$. We map it onto the class ρ via the transformation $x \to x' = G_\rho x$. For two different elements x_1, x_2 from the epsilon class, the images x_1', x_2' are different from each other because G_ρ is a transformation; that is, it is reversible. Hence, the epsilon and rho classes are equivalent. Hence, every x can uniquely be expressed by its ML estimator $\xi^{\mathrm{ML}}(x)$ and its correspondent in the epsilon class which is

$$x^{(\epsilon)} = G^{-1}_{\xi^{\mathrm{ML}}(x)}\, x\,. \tag{6.48}$$

In other words, the mapping

$$Tx = (\xi^{\mathrm{ML}}(x),\; x^{(\epsilon)}(x))\,. \tag{6.49}$$

is a transformation of the space of events.

The coordinates Tx are chosen such that the application of any G_ρ from $\mathcal{G}$ touches the first part of Tx only,

$$\begin{aligned} G_\rho(\xi^{\mathrm{ML}},\; x^{(\epsilon)}(x)) &= (G_\rho \xi^{\mathrm{ML}},\; x^{\epsilon}(x)) \\ &= (\Phi(\xi^{\mathrm{ML}};\rho),\; x^{(\epsilon)}(x))\,. \end{aligned} \tag{6.50}$$

The coordinates $x^{(\epsilon)}(x)$ remain unaffected by G_ρ. The value of ξ^{ML} is a point in the n-dimensional parameter space. Thus the ξ space has become a subspace, possibly curved, in the x space. The coordinates $x^{(\epsilon)}(x)$ complement the coordinates ξ so that every point x is obtained exactly once. The transformation $x \to Tx$ lets the integration over ξ become an integration over a subspace of the x space. Inasmuch as $p(x|\xi)$ is assumed to be proper (i.e. its integral over all x exists), the integral m, which extends over ξ, exists too. An additional argument is needed to show that the measure $\mu(\xi)$ is compatible with this expectation. The argument is given in Sect. C.2.

6.7 The Sufficient Statistic

The transformation T of Eq. (6.49) entails that the ML estimator $\xi^{\mathrm{ML}}(x)$ is the sufficient statistic of the model $p(x|\xi)$. This says that the function ξ^{ML} summarises all that can be known from the multidimensional x on the parameter ξ, because $\xi^{\mathrm{ML}}(x)$ completely determines the posterior distribution $P(\xi|x)$.

Let x be an event that leads to the estimator $\xi^{\mathrm{ML}}(x) = \epsilon$. Its posterior distribution is

$$P(\xi|x)\mathrm{d}\xi = L(\xi|x)\mu(\xi)\mathrm{d}\xi\,. \tag{6.51}$$

For $\xi = \epsilon$, the maximum value $L(\epsilon|x)$ of the likelihood function is reached. This holds for every x belonging to the class ϵ defined in Sect. 6.6. Now consider an event from the class ρ. By definition there is an x belonging to the class ϵ such that the class ρ event is given by $G_\rho x$. The posterior distribution of the latter is given by

$$\begin{aligned} P(\xi|\epsilon)\mathrm{d}\xi &= L(\xi|G_\rho x)\mu(\xi)\mathrm{d}\xi \\ &= L(G_\rho^{-1}\xi|x)\mu(G_\rho^{-1}\xi)\mathrm{d}G_\rho^{-1}\xi \\ &= L(G_\rho^{-1}\xi|x)\mu(\xi)\mathrm{d}\xi\,. \end{aligned} \tag{6.52}$$

Here, the second line results from the symmetry property of P, and the third line from the symmetry of μ. The maximum of the likelihood function is at $G_\rho^{-1}\xi = \epsilon$ or $\xi^{\mathrm{ML}} = G_\rho\epsilon = \rho$. The consequence of shifting the event from x to $G_\rho x$ is to shift ξ^{ML} from ϵ to ρ. At the same time the distribution P of ξ is shifted such that ξ is replaced by $G_\rho^{-1}\xi$.

Alternatively, this argument can be phrased as follows. Let x be an arbitrary event foreseen by the model $p(x|\xi)$. It leads to the ML estimator $\xi^{\mathrm{ML}}(x)$ which is generally not equal to ϵ. With the help of the likelihood function the posterior distribution is written as in Eq. (6.51). The symmetry property

$$L(\xi|x) = L(G_\rho\xi|G_\rho x)\,, \qquad \text{for } G_\rho \in \mathcal{G}\,, \tag{6.53}$$

leads to

$$P(\xi|x)\mathrm{d}\xi = L(G^{-1}_{\xi^{\mathrm{ML}}(x)}\xi|G^{-1}_{\xi^{\mathrm{ML}}(x)}x)\mu(\xi)\mathrm{d}\xi\,. \tag{6.54}$$

We transform x to Tx defined in Eq. (6.49). The resulting likelihood function is called L_T. This gives

$$P(\xi|x)\mathrm{d}\xi = L_T\Big(G^{-1}_{\xi^{\mathrm{ML}}(x)}\xi|\epsilon, x^{\epsilon}(x)\Big)\mu(\xi)\mathrm{d}\xi \tag{6.55}$$

and shows that the posterior depends on ξ, albeit "shifted" via the transformation $G^{-1}_{\xi^{\mathrm{ML}}(x)}$. This is a generalisation of the model $w(x-\xi)$ considered in Eqs. (2.15) and (6.26).

In summary: for all events x that lead to one and the same estimator ξ^{ML}, the posterior distribution is the same. Two events leading to different estimators produce the same form of the posterior; only the positions of the maxima of the likelihood function differ. Hence, the knowledge of $\xi^{\mathrm{ML}}(x)$ suffices to know the posterior distribution; the ML estimator is the sufficient statistic.

6.8 An Invariant Version of the Shannon Information

The Shannon information has been introduced in Sect. 2.6. We are interested in the information on the parameter ξ given by the event x in the framework of the model $p(x|\xi)$. Thus we are interested in the information conveyed by the distribution $P(\xi|x)$. In the context of the present book ξ is a continuous variable. Unfortunately Shannon's expression

$$S = \int \mathrm{d}\xi\, P(\xi|x) \ln P(\xi|x)$$

changes its value when the integration variable ξ is transformed to $T\xi$. This happens because $\ln P(\xi|x)$ does not transform like a function; see Sect. 2.2.

The Shannon information remains invariant under transformations of ξ when the above expression is replaced by

$$S = \int \mathrm{d}\xi\, P(\xi|x) \ln \frac{P(\xi|x)}{\mu(\xi)}\,, \tag{6.56}$$

where μ is the measure in the space of ξ. The interested reader is asked to show this. In the sequel Equation (6.56) is considered to define the Shannon information.

This renders S invariant under transformations of ξ ; it does not well define the absolute value of S. As long as the factor $\mu(\epsilon)$ in Eq. (6.33) remains free, the information S in Eq. (6.56) is defined only up to the additive constant $-\ln\mu(\epsilon)$. For the posterior distribution, the factor $\mu(\epsilon)$ is immaterial. In the context of the Shannon information we make use of the geometric measure μ_g. The absolute value of the geometric measure is well defined; see Sect. 6.4; it allows us to define the absolute value of the Shannon information.

In the next chapter, examples of form-invariant models $p(x|\xi)$ are given together with the appropriate prior distributions.

References

1. J. Hartigan, Invariant prior distributions. Ann. Math. Statist. **35**, 836–845 (1964)
2. J.A. Hartigan, *Bayes Theory* (Springer, New York, 1983)
3. C.M. Stein. Approximation of improper prior measures by proper probability measures. In Neyman and Le Cam [9], pp. 217–240
4. E.T. Jaynes. Prior probabilities. IEEE Trans. Syst. Sci. Cybern., SSC-**4**(3):227–241, (1968)
5. C. Villegas. On Haar priors. In Godambe and Sprott [10], pp. 409–414
6. C. Villegas, On the representation of ignorance. J. Am. Statist. Assoc. **72**, 651–654 (1977)
7. C. Villegas, Inner statistical inference. J. Am. Statist. Assoc. **72**, 453–458 (1977)
8. C. Villegas, Inner statistical inference II. Ann. Statist. **9**, 768–776 (1981)
9. J. Neyman, L.M. Le Cam (eds.), *Bernoulli, Bayes, Laplace*. (Proceedings of an International Research Seminar. Statistical Laboratory. Springer, New York, 1965)
10. V.P. Godambe, D.A. Sprott (ed.), *Foundations of Statistical Inference. Waterloo, Ontario 1970*, Toronto, 1971. Holt, Rinehart & Winston
11. R.E. Kass, L. Wasserman, The selection of prior distributions by formal rules. J. Am. Statist. Assoc. **91**, 1343–1370 (1996)
12. W.A. Bentley, W.J. Humphreys. *Snow Crystals* (Dover, New York, 1931). Reprinted in 1962 and 1980
13. M. Eigen, R. Winkler. *Das Spiel*. Piper, München, 1975. See p. 125 for the snowflakes
14. H. Weyl. *Symmetry* (Princeton University Press, Princeton, NJ, 1952). A German edition was published 1955 by Birkhäuser, Basel, under the title of *Symmetrie*
15. M. Hamermesh. *Group Theory and its Application to Physical Problems* (Addison-Wesley, Reading, Massachusetts, 1962). Reprinted by Dover Publications, New York, 1989
16. B.G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974)
17. H. Weyl, Gruppentheorie und Quantenmechanik. Wissenschaftliche Buchgesellschaft, Darmstadt, *Reprint of the*, 2nd edn. (Hirzel, Leipzig, 1967). 1931
18. H. Weyl. *The Classical Groups: Their Invariants and Representations* (Princeton University Press, Princeton, NJ, 1953). Reprinted 1964
19. E.P. Wigner. *Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra* (Academic Press, New York, 1959). Translated from the German by J.J. Griffin.– Expanded and improved edition
20. R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley, New York, 1974)
21. J. Hilgert, K.H. Neeb, *Lie-Gruppen und Lie-Algebren* (Vieweg, Braunschweig, 1991)
22. W. Lucha, F. Schöberl, *Gruppentheorie - Eine elementare Einführung für Physiker* (B.I. Hochschultaschenbuch, Mannheim, 1993)
23. A. Haar, Der Maßbegriff in der Theorie der kontinuierlichen Gruppen. Ann. Math. **34**, 147–169 (1933)
24. B.L. Welch, H.W. Peers, On formulae for confidence points based on integrals of weighted likelihood. J. R. Statist. Soc. B **25**, 318–329 (1963)
25. M. Stone, Right Haar measure for convergence in probability to quasi posteriori distributions. Ann. Math. Statist. **36**, 440–453 (1965)
26. H.W. Peers, On confidence points and Bayesian probability points in the case of several parameters. J. R. Statist. Soc. B **27**, 9–16 (1965)
27. B.L. Welch, On comparisons between confidence points procedures in the case of a single parameter. J. Royal Statist. Soc. B **27**, 1–8 (1965)
28. H.W. Peers, Confidence properties of Bayesian interval estimators. J. R. Statist. Soc. B **30**, 535–544 (1968)
29. J.A. Hartigan, Note on confidence-prior of Welch and Peers. J. Royal Statist. Soc. B **28**, 32–44 (1966)
30. T. Chang, C. Villegas, On a theorem of Stein relating Bayesian and classical inferences in group models. Can. J. Statist. **14**(4), 289–296 (1986)

31. R. Mukerjee, D.K. Dey, Frequentist validity of posteriori quantiles in the presence of nuisance parameter: Higher order asymptotic. Biometrika **80**(3), 499–505 (1993)
32. A. Nicolaou, Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. J. R. Statist. Soc. B **55**, 377–390 (1993)
33. G.S. Datta, J.K. Gosh, On prior providing frequentist validity for bayesian inference. Biometrika **82**, 37–45 (1995)
34. O.A. Al-Hujaj. Objektive Bayessche Statistik. Theorie und Anwendung. Master's thesis, Fakultät für Physik und Astronomie der Universität Heidelberg, Max-Planck-Institut für Kernphysik, D-69029 Heidelberg, 1997
35. O.-A. Al-Hujaj, H.L. Harney, Objective Bayesian statistics. Technical report, Max-Planck-Institut für Kernphysik, 1997. See arXiv:physics/9706025
36. R.E. Kass, *The Riemannian Structure of Model Spaces: A Geometrical Approach to Inference* (PhD thesis, University of Chicago, 1980). See especially pp. 94–95
37. R.E. Kass, The geometry of asymptotic inference. Stat. Sci. **4**, 188–219 (1989)
38. S.I. Amari, *Differential Geometrical Methods in Statistics*, vol. 28, Lecture Notes in Statistics (Springer, Heidelberg, 1985)

Chapter 7
Examples of Invariant Measures

In the present chapter, the symmetry groups of several form-invariant models $p(x|\xi)$ are given together with their invariant measures. We restrict ourselves to probability densities of the kind described in Chap. 4. Symmetry groups of models with discrete events are discussed in Chap. 11. Section A.7 gives the solutions to the problems suggested to the reader.

7.1 Form Invariance Under Translations

Translational invariance has been used in Chap. 2 to convey the idea of form invariance. Models with the structure

$$p(x|\xi) = w(x - \xi) \ , \qquad -\infty < x < \infty \, , \tag{7.1}$$

have the symmetry group of the translations

$$G_\xi \, x = x + \xi \ , \qquad -\infty < \xi < \infty \ . \tag{7.2}$$

Figure 7.1 illustrates the essential features of this symmetry. The multiplication function of the group is

$$\Phi(\xi; \xi') = \xi + \xi' \, , \tag{7.3}$$

and therefore the invariant measure (6.33) is uniform, that is,

$$\mu(\xi) \equiv \mu(\epsilon) \, . \tag{7.4}$$

This is also true if x and ξ are n-dimensional vectors of variables.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_7

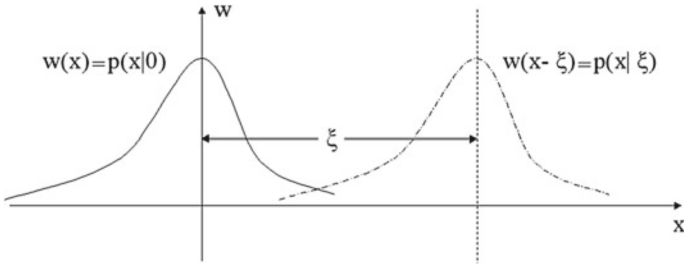


Fig. 7.1 Translational symmetry. The model $p(x|\xi)$ is obtained from the group of translations of the common form w

Form invariance under translations in one direction is found in the Gaussian model (2.16) and the Cauchy model

$$p(x, \xi) = \frac{\Gamma}{2\pi} \frac{1}{(x-\xi)^2 + \Gamma^2/4} \,. \tag{7.5}$$

For the n-dimensional vectors

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} , \qquad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} , \tag{7.6}$$

the n-dimensional Gaussian distribution

$$p(x|\xi) = (2\pi)^{-n/2} \det C^{-1/2} \exp\left(-(x-\xi)^\dagger (2C)^{-1} (x-\xi)\right) \tag{7.7}$$

is form invariant under the translations in n directions. This is an n-fold repetition of (7.2) and it is an Abelian symmetry group. The correlation matrix C is discussed in Sect. 4.1.2.

7.2 Form Invariance Under Dilations

Models of the structure

$$p(x|\sigma) = \sigma^{-1} w\left(\frac{x}{\sigma}\right), \qquad 0 < x < \infty \,, \tag{7.8}$$

are often said to be scale invariant. They have the symmetry group of the dilations

$$G_\sigma x = \sigma x \,, \qquad 0 < \xi < \infty \,. \tag{7.9}$$

The multiplication function is

$$\Phi(\sigma, \sigma') = \sigma\sigma' , \tag{7.10}$$

see (6.13). From (6.33), this yields the invariant measure

$$\mu(\sigma) = \mu(1)\, \sigma^{-1} . \tag{7.11}$$

Examples are the Gaussian model (3.3) and the exponential model (4.40).

It was shown in Sect. 3.2.2 that the model (7.8) can be reparameterised so that it takes the form (7.1). Hence, on an abstract level, form invariance under translations and form invariance under dilations have the same symmetry. It is, however, not always possible to express form invariance as translational invariance with respect to all parameters. The example in the next section shows this.

7.3 Form Invariance Under the Combination of Translation and Dilation

The model

$$p(x|\xi, \sigma) = \sigma^{-1}\, w\left(\frac{x-\xi}{\sigma}\right) , \qquad -\infty < x < \infty , \tag{7.12}$$

is form invariant under the group of transformations

$$G_{\xi,\sigma}\, x = \xi + \sigma x \tag{7.13}$$

These transformations combine a translation by the value of ξ with a dilation by the value of σ. The parameters are defined within the domains $-\infty < \xi < \infty$ and $0 < \sigma < \infty$. The translation is performed first and is followed by the dilation. The group has been introduced in Eq. (6.14). Note that

$$G^{-1}_{\xi,\sigma}\, x = \frac{x-\xi}{\sigma} \tag{7.14}$$

is the inverse of the transformation (7.13). This exemplifies a non-Abelian group. The multiplication function is

$$\Phi(\xi', \sigma'; \xi, \sigma) = (\xi' + \xi\sigma'; \sigma\sigma') , \tag{7.15}$$

see (6.19). It is left to the reader to show that the invariant measure (6.33) is

$$\mu(\xi, \sigma) = \mu(0, 1)\, \sigma^{-1} . \tag{7.16}$$

This prior distribution[1] should not be interpreted as the product of the measures (7.4) and (7.11) of the one-dimensional models of translation (7.1) and dilation (7.8): the measure of a non-Abelian group need not be equal to the product of measures of its one-dimensional subgroups (cf. Sect. 12.2). The expectation that the measure $\mu(\xi, \sigma)$ be given by the measures of the subgroups has led to controversy [1, 2] in the published literature.

Examples of the present symmetry are the Gaussian model

$$p(x|\xi, \sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\xi)^2}{\sigma^2}\right) \tag{7.17}$$

as well as the model

$$p(x|\xi, \sigma) = B^{-1}\sigma^{-1}\left(1 + \left(\frac{x-\xi}{\sigma}\right)^2\right)^{-\nu} \; . \tag{7.18}$$

The latter is the Student's t-distribution (see Eq. (4.43)). The normalising factor is the Beta function

$$B = B\left(\frac{1}{2}, \nu - \frac{1}{2}\right) , \tag{7.19}$$

see Sect. B.5. With $\nu = 1$, expression (7.18) becomes a Cauchy distribution; see Sect. 4.3. The Gaussian (7.17) is studied in Chap. 10.

It is impossible to reparameterise the model (7.12) such that its symmetry is the group of translations in two directions. A reparameterisation maps the symmetry group onto an isomorphic one. The group of two-dimensional translations is Abelian whereas the present symmetry group is non-Abelian; these two groups are not isomorphic. Hence, the present symmetry is really different from the preceding ones.

7.4 A Rotational Invariance

Let us consider the two-dimensional Gaussian distribution

$$w(x) = (2\pi\sigma_1\sigma_2)^{-1} \exp\left(-x^\dagger\,(2C)^{-1}\,x\right) , \tag{7.20}$$

where x is the vector

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} , \tag{7.21}$$

[1] The corresponding result in Eq. (7.17) of the first edition of the present book is not consistent with the (correct) multiplication function (7.16) given there. The error is due to incorrect multiplication functions in Eqs. (6.15) and (A.82) of that edition. In the present edition, Eqs. (6.16)–(6.18) make clear in which way and in which order the operations of translation and dilation are carried out.

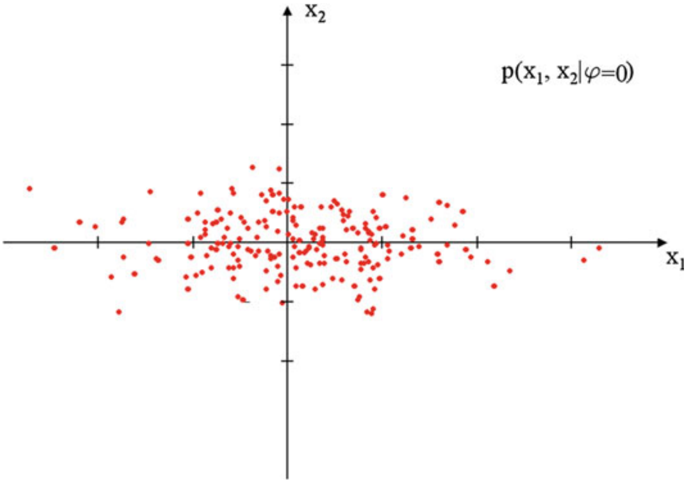


Fig. 7.2 Scatterplot of a two-dimensional Gaussian distribution without correlation between x_1 and x_2

and C is the diagonal correlation matrix

$$C = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} . \tag{7.22}$$

In this form the distribution does not correlate the variables x_1 and x_2; see Sect. 4.1.2. It is illustrated in Fig. 7.2. We introduce a correlation via an orthogonal transformation of the event. If we take

$$G_\phi = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} , \tag{7.23}$$

the model

$$\begin{aligned} p(x|\phi) &= w(G_\phi^{-1}x) \\ &= (2\pi\sigma_1\sigma_2)^{-1} \exp\left(-x^\dagger \left(2G_\phi C G_\phi^\dagger\right)^{-1} x\right) \end{aligned} \tag{7.24}$$

is constructed. Note that the Jacobian of the orthogonal transformation is unity. The parameter ϕ rotates the distribution (7.20), as is illustrated by Fig. 7.3. We have seen in Sect. 6.1 that the rotations (7.23) form a group for $0 \le \phi < 2\pi$. The model (7.24) is form invariant. The invariant measure $\mu(\phi)$ is uniform, by the discussion in Sect. 6.4.

In the present case, the prior distribution is proper; that is, the volume

$$\begin{aligned} V &= \int_0^{2\pi} \mu(\phi)\, \mathrm{d}\phi \\ &= 2\pi\, \mu(0) \end{aligned} \tag{7.25}$$

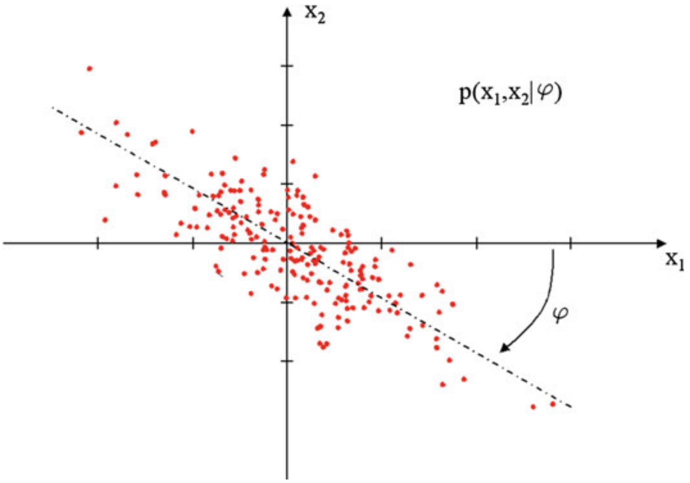
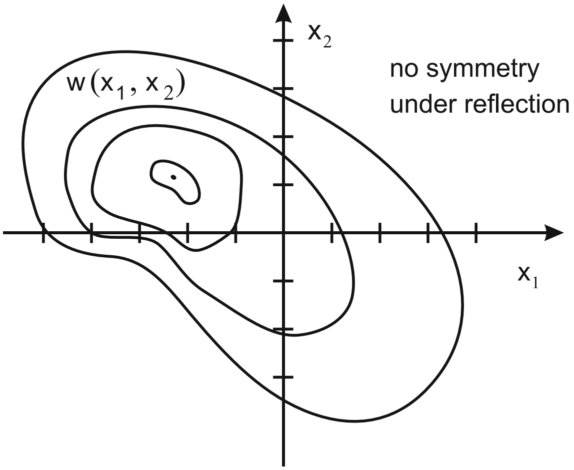


Fig. 7.3 Scatterplot of a two-dimensional Gaussian with correlated x_1, x_2

Fig. 7.4 A two-dimensional distribution that - in contrast to the Gaussian model (7.24) - is not symmetric under reflection at the origin



of the whole space of ξ exists. The interested reader should show that the volume of the space does not depend on the parameterisation. If the integral (7.25) does not exist, then this is again true for every parameterisation.

The domain of definition of ϕ is only $0 \leq \phi < \pi$, because the distribution (7.24) does not change when it is rotated by π. However, the transformations (7.23) do not form a group under this restriction of ϕ; see Problem A.6.2. Rotational invariance is truly present only if w is as sketched in Fig. 7.4, which requires $0 \leq \phi < 2\pi$. This shows that the symmetry of form invariance is an idealisation, and in practical cases, it is necessary to define the measure μ in a more general way than via the symmetry group. In Chap. 9, we define it as a geometric measure.

7.5 Special Triangular Matrices

A correlation can be introduced in a way other than that chosen in the preceding section. Let x be distributed as in (7.20) but with the correlation matrix equal to the unit matrix; that is,

$$C = \mathbf{1} \ . \tag{7.26}$$

Again there is no correlation between x_1 and x_2. However, the variables

$$\begin{aligned} y_1 &= x_1 + \gamma x_2 \,, \\ y_2 &= x_2 \end{aligned} \tag{7.27}$$

are correlated, and the correlation equals γ. We write this transformation in the form

$$y = G_\gamma x \ . \tag{7.28}$$

The transformation is achieved by the triangular matrix

$$G_\gamma = \begin{pmatrix} 1 & \gamma \\ 0 & 1 \end{pmatrix} \ , \qquad -\infty < \gamma < \infty \ . \tag{7.29}$$

Let us construct the model p by transforming the events of the model (7.20) with (7.26) according to

$$p(x|\gamma) = w(G_\gamma^{-1} x) \left| \frac{\partial G_\gamma^{-1} x}{\partial x} \right| \ . \tag{7.30}$$

The Jacobian in this equation is the determinant of G_γ^{-1}; that is, it is unity. Therefore one obtains

$$p(x|\gamma) = (2\pi)^{-1} \exp\left(-x^\dagger \left(2G_\gamma G_\gamma^\dagger\right)^{-1} x\right) \ . \tag{7.31}$$

One can easily verify that the matrices (7.29) form an Abelian group and have the multiplication function

$$\Phi(\gamma', \gamma) = \gamma + \gamma' \ . \tag{7.32}$$

Hence, the model (7.31) is form invariant and the invariant measure $\mu(\gamma)$ is uniform.

7.6 Triangular Matrices

In order to find the general two-dimensional correlation matrix, let us replace the special triangular matrix (7.29) by

$$G_\xi = \begin{pmatrix} \alpha & \gamma \\ 0 & \beta \end{pmatrix} \tag{7.33}$$

and replace the model (7.30) by

$$p(x|\xi) = w(G_\xi^{-1}\, x) \left| \frac{\partial G_\xi^{-1}\, x}{\partial x} \right| , \tag{7.34}$$

where $w(x)$ is given by (7.20) with (7.26). Here, ξ stands for the three variables in the matrix (7.33): that is, $\xi = (\alpha, \gamma, \beta)$ and $G_\xi = G_{\alpha,\gamma,\beta}$. The order of the parameters says that the transformation labelled α is applied first; it is followed by the transformations labelled γ and finally β. These three transformations are

$$\begin{aligned} G_{1,0,\beta} &= \begin{pmatrix} 1\,, & 0 \\ 0\,, & \beta \end{pmatrix} , \qquad \beta > 0\,, \\ G_{1,\gamma,1} &= \begin{pmatrix} 1\,, & \gamma \\ 0\,, & 1 \end{pmatrix} , \qquad -\infty < \gamma < \infty\,, \\ G_{\alpha,0,1} &= \begin{pmatrix} \alpha\,, & 0 \\ 0\,, & 1 \end{pmatrix} , \qquad \alpha > 0\,. \end{aligned} \tag{7.35}$$

Each of these matrices forms a one-dimensional group of transformations. Their combination

$$\begin{aligned} G_{\alpha,\gamma,\beta} &= G_{1,0,\beta}\, G_{1,\gamma,1}\, G_{\alpha,0,1} \\ &= \begin{pmatrix} \alpha\,, & \gamma \\ 0\,, & \beta \end{pmatrix} \end{aligned} \tag{7.36}$$

forms a three-dimensional group $\mathcal{G}$. The combination is a group because the product

$$\begin{aligned} G_{\alpha,\gamma,\beta} G_{\alpha',\gamma',\beta'} &= \begin{pmatrix} \alpha\,, & \gamma \\ 0\,, & \beta \end{pmatrix} \begin{pmatrix} \alpha'\,, & \gamma' \\ 0\,, & \beta' \end{pmatrix} \\ &= \begin{pmatrix} \alpha\alpha'\,, & \alpha\gamma' + \gamma\beta' \\ 0\,, & \beta\beta' \end{pmatrix} \end{aligned} \tag{7.37}$$

is contained in $\mathcal{G}$. The unit element of has the index

$$\epsilon = (1, 0, 1)\,. \tag{7.38}$$

The inverse of $G_{\alpha,\gamma,\beta}$ is given by

$$G^{-1}_{\alpha,\gamma,\beta} = G_{\alpha^{-1},-\gamma(\alpha\beta)^{-1},\beta^{-1}}\,. \tag{7.39}$$

The interested reader is asked to confirm this. Thus the set of transformations (7.33) is a group and therefore the model (7.34) with (7.33) is form invariant.

By Eq. (7.37) the multiplication function of the group is

$$\Phi(\alpha', \gamma', \beta'\,; \alpha, \gamma\beta) = (\alpha\alpha', \alpha\gamma' + \gamma\beta', \beta\beta')\,. \tag{7.40}$$

From this we obtain the invariant measure

$$\begin{aligned}\mu(\xi) &= \mu(\epsilon)\begin{vmatrix}\alpha\,, & 0\,, & 0\\ 0\,, & \alpha\,, & \gamma\\ 0\,, & 0\,, & \beta\end{vmatrix}^{-1}\\ &= \mu(\epsilon)\,(\alpha^2\beta)^{-1}\end{aligned} \tag{7.41}$$

according to Eq. (6.33). The interested reader is asked to verify this result. It agrees with Formula (2.7) of [3].

This group is not Abelian. One recognises this in the second entry on the r.h.s. of the multiplication function (7.40). This entry is altered when the primed quantities are interchanged with the unprimed ones.

The Jacobian in (7.34) is the determinant of the matrix G_ξ^{-1} and is thus equal to $(\alpha\beta)^{-1}$. Therefore one obtains

$$p(x|\xi) = (2\pi\alpha\beta)^{-1}\exp\left(-x^\dagger\left(2G_\xi G_\xi^\dagger\right)^{-1}x\right)\,. \tag{7.42}$$

The correlation matrix is

$$\begin{aligned}C_\xi &= G_\xi G_\xi^\dagger\\ &= \begin{pmatrix}\alpha^2+\gamma^2, & \beta\gamma\\ \beta\gamma, & \beta^2\end{pmatrix}\,.\end{aligned} \tag{7.43}$$

It is not difficult to convince oneself that it is positive definite; that is, it has positive eigenvalues for $0 < \alpha,\ \beta < \infty$ and $-\infty < \gamma < \infty$.

The parameters α and β essentially control the variances of x_1 and x_2 and γ controls their correlation. If one combines the rotation (7.23) with the diagonal matrices in (7.35) - depending on α and β - one does not obtain a form-invariant model. This combination does not yield a group. We aim at form-invariant models because the posterior will have the same form for every event x; only the position of the ML estimator will depend on x. See Chap. 6. This allows us to think that one and the same property is extracted from the data although the quantity of that property varies.

References

1. H. Jeffreys, An invariant form of the prior probability in estimation problems. Proc. Roy. Soc. A **186**, 453–461 (1946)
2. J.M. Bernardo, Unacceptable implications of the left Haar measure in a standard normal theory inference problem. Trab. Est. Invest. Operat. **29**, 3–9 (1978)
3. C. Villegas, On Haar priors, in Godambe and Sprott [4], pp. 409–414
4. V.P. Godambe, D.A. Sprott (eds.), *Foundations of Statistical Inference* (Holt, Rinehart & Winston, Waterloo, 1970)

Chapter 8
A Linear Representation of Form Invariance

According to the principles of group theory, every group $\mathcal{G}$ of transformations G_ρ can be represented by an isomorphic group $\mathcal{G}_L$ of linear transformations $\mathbf{G}_\rho$ of a vector space. The transformations of event and parameter, introduced in Chap. 6, are nonlinear. The probability distributions p and w are not elements of a vector space. Is it possible to define form invariance in terms of linear transformations of a vector space? This would give the possibility of distinguishing classes of transformations, such as orthogonal or unitary ones, that can occur in form invariance, from other ones that cannot occur. Section 6.4 has given a hint of the way of constructing linear representations: one must express probabilities by probability amplitudes. The present chapter pursues this route and defines linear representations of models with the symmetries of translation and dilation.

It is not a mathematical glass bead game[1] to look for a linear representation of the symmetry of form invariance. The results of the present chapter prepare us for the generalisation of form invariance needed in Chap. 11. Indeed, form invariance as defined in Sect. 6.2 can be found for probability densities only: it requires that the event variable x be continuous. In this case there are transformations G_ξ of x which are arbitrarily close to the identity G_ϵ. They allow one to shift x by an infinitesimally small amount. If the event variable is discrete, it cannot be shifted infinitesimally. Then one must look for a more general notion of form invariance. It is found in the form invariance of probability amplitudes.

Technically a linear representation of a group $\mathcal{G}$ is a one-to-one mapping of the elements G_ξ in $\mathcal{G}$ onto a group $\mathcal{G}_L$ of linear transformations $\mathbf{G}_\xi$ of a vector space such that the multiplication function remains the same. In other words: a linear representation is an isomorphism between $\mathcal{G}$ and a group $\mathcal{G}_L$ of linear transformations.

In Sect. 8.1, we take a short look at linear transformations of a space of functions in analogy to transformations of a vector space. In Sect. 8.2, orthogonal transformations of probability amplitudes are discussed. In Sect. 8.3, the linear representation of form invariance is described in a general way. Sections 8.4 to 8.6 treat the examples of

[1] See the introductory chapter of Hermann Hesse's novel.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_8

symmetry under translation, dilation, and the combination of both. Section A.8 gives the solutions to the problems suggested to the reader.

8.1 A Vector Space of Functions

The concept of a vector space of functions is introduced here together with linear operators acting on it. We do not enter into the mathematical foundations but, rather, pragmatically use the notions mentioned below, as mathematics is often used by physicists.

Consider the space of real square integrable functions that are defined on the domain of definition of the event variable x. We may write the function f of x not in the usual way, $f(x)$, but rather as f_x, inasmuch as we consider it as a vector with the components f_x labelled by x. The space is endowed with the inner product

$$f^\dagger g = \int dx\ f_x g_x \ . \tag{8.1}$$

Note that we use the superscript † to denote the adjoint of an element of the space as well as that of an operator, even if these objects are real and no complex conjugation is implied. From (8.1), it is therefore not obvious why the dagger is used at all to denote the inner product. In fact, the dagger distinguishes the inner product $f^\dagger g$ from the dyadic product $g f^\dagger$. The first product is a number; the second one is an operator.

A linear operator $\mathbf{T}$ acting on the elements of the space can be characterised by its "matrix of coefficients" $\mathbf{T}_{xx'}$. This matrix is usually called an integral kernel because the action of $\mathbf{T}$ on the function f in the space is defined as the integral

$$(\mathbf{T}f)_x = \int dx'\, \mathbf{T}_{xx'} f_{x'} \ . \tag{8.2}$$

Here, $(\mathbf{T}f)_x$ is the element x of the vector that results from the action of $\mathbf{T}$ on f. The kernel must be such that $\mathbf{T}f$ belongs to the function space when f is an element of that space. We do not enter into the details of this question. The kernel of the dyadic product is

$$(g f^\dagger)_{xx'} = g_x f_{x'} \ , \tag{8.3}$$

and the integral kernel of the product of two operators $\mathbf{T}$ and $\mathbf{S}$ is

$$(\mathbf{TS})_{xx'} = \int dy\, \mathbf{T}_{xy} \mathbf{S}_{yx'} \ . \tag{8.4}$$

The adjoint $\mathbf{T}^\dagger$ has the integral kernel

$$(\mathbf{T}^\dagger)_{xx'} = \mathbf{T}_{x'x} \ . \tag{8.5}$$

An orthogonal operator $\mathbf{O}$ has the property that it conserves the inner product of any pair of real functions f, g which means

$$\begin{aligned} f^\dagger g &= (\mathbf{O}f)^\dagger\, \mathbf{O}g \\ &= \int dy \int dx\, \mathbf{O}_{yx} f(x) \int dx'\, \mathbf{O}_{yx'} g(x') \\ &= f^\dagger \mathbf{O}^\dagger \mathbf{O} g\ . \end{aligned} \tag{8.6}$$

One proceeds from the second to the third line of this equation by changing the order of the integrations. Equation (8.6) holds for any pair of functions f, g if and only if $\mathbf{O}^\dagger\mathbf{O}$ is equal to the unit operator $\mathbf{1}$. The integral kernel of the unit operator is Dirac's[2] δ-distribution[3]; that is,

$$\mathbf{1}_{xx'} = \delta(x - x')\ . \tag{8.7}$$

Hence, the definition of an orthogonal operator $\mathbf{O}$ is equivalent to

$$(\mathbf{O}^\dagger\mathbf{O})_{xx'} = \delta(x - x')\ . \tag{8.8}$$

All of this generalises the notion of an N-dimensional vector space, where the variable x labels the components of the vectors. In that case, x is discrete and runs over N values. In the present section, x is continuous and runs over an uncountable set of values.

8.2 An Orthogonal Transformation of the Function Space

We now introduce an orthogonal transformation $\mathbf{T}$ of the function space.

Let f be an element of the space of functions considered in the preceding section, and let T be an arbitrary—not necessarily linear—transformation of x. We introduce the mapping of f onto $\mathbf{T}f$ such that

$$\begin{aligned} (\mathbf{T}f)_x &= \int \mathrm{d}x'\, \mathbf{T}_{xx'} f_{x'} \\ &= f(T^{-1}x) \left| \frac{\partial T^{-1}x}{\partial x} \right|^{1/2}\ . \end{aligned} \tag{8.9}$$

This means that $(\mathbf{T}f)$ emerges from f by shifting the component f_x to the place $x' = T^{-1}x$. The function f could be, for example, the square root of the common

[2] Paul A. M. Dirac, 1902–1984, British physicist and Nobel laureate. He contributed to the foundation of quantum mechanics. He found a relativistically invariant form of it.

[3] The δ distribution is often called the δ function. However, under a reparameterisation, it behaves as a distribution, not as a function.

form w of Eq. (6.24). Given a transformation acting on x, its linear representation acts on f and must produce the result of Eq. (8.9).

The interested reader should show that $\mathbf{T}f$ is a square integrable function and thus belongs to the space under consideration. The mapping is linear, because one obviously has

$$\mathbf{T}(f+g) = \mathbf{T}f + \mathbf{T}g \tag{8.10}$$

as well as

$$\mathbf{T}(cf) = c\mathbf{T}f \ , \tag{8.11}$$

when f and g are square integrable functions and c is the number. The operation $\mathbf{T}$ can be inverted. One can see this by replacing T^{-1} with T in the definition (8.9) to obtain a linear transformation, which we call $\mathbf{T}^{-1}$. Applying this transformation to $\mathbf{T}f$ yields

$$\begin{aligned} \mathbf{T}^{-1}\mathbf{T}f &= f(TT^{-1}x)\left|\frac{\partial TT^{-1}x}{\partial T^{-1}x}\right|^{1/2}\left|\frac{\partial T^{-1}x}{\partial x}\right|^{1/2} \\ &= f(x) \ . \end{aligned} \tag{8.12}$$

Hence, $\mathbf{T}^{-1}$ is indeed the inverse of $\mathbf{T}$. It follows that $\mathbf{T}$ is a transformation of the function space.

Furthermore, $\mathbf{T}$ is an orthogonal transformation. In order to see this, the reader may verify that its integral kernel is

$$\mathbf{T}_{xx'} = \delta(x' - T^{-1}x)\left|\frac{\partial T^{-1}x}{\partial x}\right|^{1/2} \ . \tag{8.13}$$

By replacing T^{-1} with T in this formula, we can show that $\mathbf{T}^{-1}$ is equal to the adjoint of $\mathbf{T}$. In doing so, one must observe that the δ-distribution is transformed according to (2.9). This yields

$$\begin{aligned} (\mathbf{T}^{-1})_{xx'} &= \delta(x' - Tx)\left|\frac{\partial Tx}{\partial x}\right|^{1/2} \\ &= \delta(T^{-1}x' - x)\left|\frac{\partial T^{-1}x'}{\partial x'}\right|\left|\frac{\partial Tx}{\partial x}\right|^{1/2} \\ &= \delta(T^{-1}x' - x)\left|\frac{\partial T^{-1}x'}{\partial x'}\right|^{1/2} \\ &= \mathbf{T}_{x'x} \ . \end{aligned} \tag{8.14}$$

The step from the first to the second line is performed by a transformation of the δ distribution. Observing that the delta distribution enforces x to equal $T^{-1}x'$ and by consequence Tx to equal x', one proceeds from the second to the third line. The result of the third line is obtained from (8.13) by interchanging x with x'. The last line is equivalent to (8.8) and, hence, $\mathbf{T}$ is orthogonal.

We show that a group of transformations G_ξ acting on x leads, via Eq. (8.9), to a group of linear transformations $\mathbf{G}_\xi$ acting on f.

8.3 The Linear Representation of the Symmetry Groups

Let $\mathcal{G}$ be a group of transformations G_ξ (generally nonlinear) of the domain where the event x is defined. The set $\mathcal{G}_L$ of the orthogonal transformations

$$\left(\mathbf{G}_\xi f\right)_x = f(G_\xi^{-1} x) \left| \frac{\partial G_\xi^{-1} x}{\partial x} \right|^{1/2} , \tag{8.15}$$

obtained by letting ξ run over all values of the group parameter of $\mathcal{G}$, is a group of linear transformations of the function space. The group $\mathcal{G}_L$ is isomorphic to the group $\mathcal{G}$.

We verify that $\mathcal{G}_L$ is a group by reviewing the axioms of Sect. 6.1. The product $\mathbf{G}_\xi \mathbf{G}_{\xi'}$ is obtained by applying $\mathbf{G}_\xi$ to $\mathbf{G}_{\xi'} f$. When doing so, note especially that the operation G_ξ^{-1} is applied to the variable x. We find

$$\begin{aligned}
\left(\mathbf{G}_\xi \mathbf{G}_{\xi'} f\right)_x &= \mathbf{G}_\xi\, f(G_{\xi'}^{-1} x) \left| \frac{\partial G_{\xi'}^{-1} x}{\partial x} \right|^{1/2} \\
&= f(G_{\xi'}^{-1} G_\xi^{-1} x) \left| \frac{\partial G_{\xi'}^{-1} G_\xi^{-1} x}{\partial G_\xi^{-1} x} \right|^{1/2} \left| \frac{\partial G_\xi^{-1} x}{\partial x} \right|^{1/2} \\
&= f(G_{\xi'}^{-1} G_\xi^{-1} x) \left| \frac{\partial G_{\xi'}^{-1} G_\xi^{-1} x}{\partial x} \right|^{1/2} \\
&= f((G_\xi G_{\xi'})^{-1} x) \left| \frac{\partial (G_\xi G_{\xi'})^{-1} x}{\partial x} \right|^{1/2} .
\end{aligned} \tag{8.16}$$

This transformation is an element of $\mathcal{G}_L$ because $G_\xi G_{\xi'}$ is in $\mathcal{G}$. Products of the linear transformations $\mathbf{G}_\xi$ are associative because products of the transformations G_ξ are associative. It was shown in Sect. 8.2 that $\mathbf{G}_\xi^{-1}$ exists together with $\mathbf{G}_\xi$. The unit element is in $\mathcal{G}_L$ and has the parameter ϵ. Hence $\mathcal{G}_L$ is a group. The last version of Eq. (8.16) shows that $\mathcal{G}_L$ and $\mathcal{G}$ have the same multiplication function, whence both groups are isomorphic and $\mathcal{G}_L$ is a linear representation of $\mathcal{G}$.

Let $w(x)$ be the common form of the model (6.23). We consider the function

$$f_x = \sqrt{w(x)} . \tag{8.17}$$

This function is square integrable, because w is normalised and thus f belongs to the function space. By applying $\mathbf{G}_\xi$, one obtains

$$\begin{aligned}\left(\mathbf{G}_\xi f\right)_x &= \sqrt{w(G_\xi^{-1}x)}\left|\frac{\partial G_\xi^{-1}x}{\partial x}\right|^{1/2} \\ &= \sqrt{p(x|\xi)}\,. \end{aligned} \tag{8.18}$$

This is a probability amplitude. It deserves a notation of its own; throughout the present book, we write

$$a_x(\xi) = \sqrt{p(x|\xi)}\,. \tag{8.19}$$

The element of the function space that has these components is called $a(\xi)$.

The probability amplitude is an object with richer properties than the probability; the amplitude may have both signs. It may therefore change sign at the places where p vanishes. One may even consider complex probability amplitudes and thus turn the function space into a Hilbert space.

The definition of $a(\xi)$ allows us to write (8.18) in the form

$$a(\xi) = \mathbf{G}_\xi\, a(\epsilon)\,. \tag{8.20}$$

The symmetry defined by (6.23) is therefore equivalent to the fact that the amplitude vectors $a(\xi)$ are obtained from a common form $a(\epsilon)$ by a group of linear transformations $\mathbf{G}_\xi$ according to (8.15) and (8.20). Thus the symmetry of form invariance is a symmetry of the probability amplitudes as well as a symmetry of the probabilities. The consequence of this is explored in Chap. 11.

The action of an element $\mathbf{G}_\rho$ of $\mathcal{G}_L$ on $a(\xi)$ can be expressed as

$$\begin{aligned}\mathbf{G}_\rho\, a(\xi) &= \mathbf{G}_\rho \mathbf{G}_\xi\, a(\epsilon) \\ &= a(\Phi(\xi;\rho))\,, \end{aligned} \tag{8.21}$$

where Φ is the multiplication function introduced in Sect. 6.1. Using the transformation of ξ defined by (6.21), one brings the last equation into the form

$$\mathbf{G}_\rho\, a(\xi) = a(G_\rho \xi)\,. \tag{8.22}$$

This relation expresses the form invariance of the probability amplitudes.

We have shown that the symmetry defined by (6.23) and (6.24) can be expressed by (8.20). The reverse is not true. One cannot express every group of orthogonal transformations of an amplitude vector as a group of transformations of the event variable x. When x is discrete, this cannot be done. Thus the concept of probability amplitudes allows us to generalise form invariance beyond the concept of Chap. 6. We define: the model $p(x|\xi)$ is called form invariant if the amplitude vector $a(\xi)$ emerges from a common form $a(\epsilon)$ by a group of orthogonal linear transformations

$\mathbf{G}_\xi$. This generalisation is important in Chap. 11. In any case, the prior distribution of form-invariant models is the invariant measure of the symmetry group.

In the following sections we especially study the linear representations of translation and dilation. We show that their linear operators $\mathbf{G}_\xi$ can be expressed by an exponential function of a "generating operator" **g**.

8.4 The Linear Representation of Translation

The simplest example of form invariance was considered in Sects. 2.3 and 6.2; it exhibits translational symmetry such as

$$p(x|\xi) = w(x-\xi)\,, \qquad -\infty < x,\, \xi < \infty\,. \tag{8.23}$$

The transformation of the probability amplitude $f_x = \sqrt{w(x)}$ to $\sqrt{w(x-\xi)}$ can be written as the Taylor expansion

$$\sqrt{w(x-\xi)} = \sum_{k=0}^{\infty} \frac{(-\xi)^k}{k!} \frac{\partial^k}{\partial x^k} f_x\,. \tag{8.24}$$

The shift from x to $x-\xi$ is an operation on x like the operations T^{-1} in Eq. (8.9) and G_ξ^{-1} in Eq. (8.15). The Taylor expansion (8.24), however, can be understood - and shall be understood - as an operation on the function f as a whole; at every place x it replaces the value f_x by the value at $x-\xi$ as is illustrated in Fig. 7.1.

We write the sum on the r.h.s. of the last equation as an exponential of the differential operator $\partial/\partial x$; that is, we introduce

$$\begin{aligned}\mathbf{G}_\xi &= \exp\left(-\xi\frac{\partial}{\partial x}\right)\\ &= \sum_{k=0}^{\infty}\frac{(-\xi)^k}{k!}\frac{\partial^k}{\partial x^k}\,.\end{aligned} \tag{8.25}$$

This is a linear operator because every differentiation in the sum (8.25) is a linear operation. In fact, this is an explicit formula for the linear operator evoked in Eq. (8.15). It shifts the events of the model with the common form $w(x) = p(x|\epsilon)$ from x to $x-\xi$; that is, it changes $p(x|\epsilon)$ to $p(x|\xi)$. Note that the translation implies $\partial G_\xi^{-1}x/\partial x \equiv 1$.

From group theory it is expected that a linear representation of a (one-parametric) Lie group can be characterised by one single operator, the generator of the group. Every element G_ξ of the group is given by the exponential function of ξ times the generator **g**. Equation (8.25) says that, for the translation, the generator is given by the differential operator $\partial/\partial x$. However, it is not simply equal to this; we must take care of the following finesse. The translation is an orthogonal operator; when applied to f it conserves the norm $f^\dagger f$. One knows from the theory of linear operators that a unitary operator U can be written as the exponential function

$$U = \exp(iH) \tag{8.26}$$

of a Hermitian matrix H multiplied with the imaginary unit i. This means that an orthogonal operator (which is real) requires H to be purely imaginary and therefore antisymmetric. The operator $\partial/\partial x$ is antisymmetric in the present context. Translational symmetry requires x and ξ to be defined on the entire real axis. Whence, the normalisation

$$\int_{-\infty}^{\infty} \mathrm{d}x\, f_x^2 = 1 \tag{8.27}$$

requires f_x to vanish with $x \to \pm\infty$. This is true for any other probability amplitude g_x as well. Therefore we obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{d}x\, f_x \frac{\partial}{\partial x} g_x &= \left[f_x g_x \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \mathrm{d}x \left(\frac{\partial}{\partial x} f_x \right) g_x \\ &= - \int_{-\infty}^{\infty} \mathrm{d}x\, g_x \frac{\partial}{\partial x} f_x \,. \end{aligned} \tag{8.28}$$

For this reason the generator of translation is the Hermitian operator

$$\mathbf{g}_t = i \frac{\partial}{\partial x} \,, \tag{8.29}$$

and we write G_ξ in the form

$$G_\xi = \exp(i\xi \mathbf{g}_t) \,. \tag{8.30}$$

The generator $\mathbf{g}$ is also called the "infinitesimal operator" of the group because it yields the approximation

$$\mathbf{G}_\xi \approx \mathbf{1} + i\xi \mathbf{g} \tag{8.31}$$

to the first order in ξ.

In the following section, the linear representation of the symmetry of dilation is derived.

8.5 The Linear Representation of Dilation

We can use the linear operator of translation to construct the linear operator of dilation. Consider again the model with the structure of (3.9),

$$p(x|\sigma) = \sigma^{-1} w \left(\frac{x}{\sigma} \right) , \qquad 0 < \sigma \,. \tag{8.32}$$

To begin we assume the event variable x to be positive; this assumption can be dropped later on.

We want to bring the linear transformation of the probability amplitude (8.17) of this model into a form similar to Eq. (8.30). We show in Chap. 9 that this requires the measure μ of the parameter to be uniform, whence σ must be transformed according to

$$\eta = \ln \sigma \,, \tag{8.33}$$

see Eq. (2.14). This brings the model (8.32) into the form

$$p(x|\eta) = \exp(-\eta) w(x \exp(-\eta)), \qquad -\infty < \eta < \infty \,, \tag{8.34}$$

and for $\eta = 0$ the probability amplitude is

$$\begin{aligned} f_x &= a_x(\epsilon) \\ &= \Big(w(x)\Big)^{1/2} . \end{aligned} \tag{8.35}$$

For $G_\eta x = x\, e^\eta$ the linear transformation $\mathbf{G}_\eta$ has to respect Eq. (8.18); that is,

$$\begin{aligned} \Big(\mathbf{G}_\eta\Big)_x &= \Big(w(G_\eta^{-1}x)\Big)^{1/2} \frac{\partial G_\eta^{-1}x}{\partial x} \\ &= \Big(w(G_\eta^{-1}x)\Big)^{1/2} \exp(-\eta/2) \,. \end{aligned} \tag{8.36}$$

The linear operator $\mathbf{G}_\eta$ acts on x. One can transform x such that $\mathbf{G}_\eta$ becomes the operator of translation considered in Sect. 8.4. For this, x must be transformed to y such that the measure $m(y)$ becomes uniform and equal to the measure $\mu(\eta)$. The measure m of the event variable has been defined in Eq. (2.5). In the present case the required transformation is

$$y = \ln x \,. \tag{8.37}$$

Then the model (8.33) becomes

$$\tilde{p}(y|\eta) = \exp(y - \eta) w\left(\exp(y - \eta)\right) \tag{8.38}$$

and its probability amplitude $\tilde{f}_y = \tilde{a}_y(\epsilon)$ reads

$$\tilde{f}_y = \exp(y/2) \Big(w(\exp y)\Big)^{1/2} . \tag{8.39}$$

In this parameterisation, the linear operator $\tilde{\mathbf{G}}_\eta$ is just the operator that translates y to $y - \eta$; that is,

$$\tilde{\mathbf{G}}_\eta = \exp\left(i\eta\frac{\partial}{\partial y}\right) \tag{8.40}$$

as considered in Sect. 8.4.

Now we can express the generator $i\frac{\partial}{\partial y}$ of Eq. (8.40) by a differential operator that acts on the variable x. The variable y is a unique function of x. Therefore the amplitude $\tilde{f}$ implicitly is the function f of x. Its differentiation with respect to x is

$$\begin{aligned}\frac{\partial}{\partial y}\tilde{f} &= \frac{\mathrm{d}x}{\mathrm{d}y}\frac{\partial}{\partial x}f \\ &= \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^{-1}\frac{\partial}{\partial x}f \\ &= x\frac{\partial}{\partial x}f\,. \end{aligned} \tag{8.41}$$

This means that

$$\mathbf{g}_d = ix\frac{\partial}{\partial x} \tag{8.42}$$

is the generator of the linear representation of the transformation of the amplitude (8.35). Thus we obtain

$$\begin{aligned} a_x(\eta) &= \tilde{\mathbf{G}}_\eta\, a_x(\epsilon) \\ &= \exp(i\eta\mathbf{g}_d)\, a_x(\epsilon) \\ &= \left(w(x\,e^{-\eta})\right)^{1/2}. \end{aligned} \tag{8.43}$$

From Eq. (8.42) one sees that the generator $\mathbf{g}_d$ remains well defined if the event variable x changes sign. Therefore the requirement $x > 0$ can be dropped.

8.6 Linear Representation of Translation Combined with Dilation

The combination of translation with dilation means "translation followed by dilation". A model which is invariant under this combination is also invariant under a dilation followed by a translation. However, the two ways of looking at the symmetry group imply different multiplication functions Φ und thus different measures μ; see Chap. 6. In either of the two cases the set of combined transformations is a mathematical group $\mathcal{G}$. One therefore expects that the product $\mathbf{g}_d\mathbf{g}_t$ as well as the product $\mathbf{g}_t\mathbf{g}_d$ can be expressed by the generators $\mathbf{g}_t$ of translation and $\mathbf{g}_d$ of dilation. This means that the first product differs from the second one by a linear combination of the generators $\mathbf{g}_t$ and $\mathbf{g}_d$ multiplied by the imaginary unit i; in other words: the commutator

$$[\mathbf{g}_t, \mathbf{g}_d] = \mathbf{g}_t\mathbf{g}_d - \mathbf{g}_d\mathbf{g}_t \tag{8.44}$$

must be a linear combination of $\mathbf{g}_t$ and $\mathbf{g}_d$ multiplied by i. The factor i is due to Eq. (8.26).

This is indeed the case for the generators (8.29) of translation and (8.42) of dilation, as is shown below. Consider again the model (2.16). It is now conditioned by both parameters, x and ξ,

$$p(x|\xi, \sigma) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right), \tag{8.45}$$

where $-\infty < x < \infty$ and $0 < \sigma < \infty$. To obtain the linear operator of the combined group, each parameter (when considered to be the only parameter of the model) must be defined such that its measure μ is uniform. For the parameter ξ of translation this is already the case. The parameter σ of dilation must be transformed according to Eq. (8.37), as is explained in Sect. 6.5. This brings (8.45) into the form

$$\begin{aligned} p(x|\xi, \eta) &= (2\pi)^{-1/2}\, e^{-\eta} \exp\left(-\left((x-\xi)e^{-\eta}\right)^2/2\right) \\ &= (2\pi)^{-1/2}\, e^{-\eta} \exp\left(-\left(G^{-1}_{\xi,\eta}x\right)^2/2\right) \end{aligned} \tag{8.46}$$

The operation G acting on x is

$$G_{\xi,\eta}\, x = x\, e^{\eta} + \xi\,, \tag{8.47}$$

and the index of the unit operator is

$$\epsilon = (0, 0)\,. \tag{8.48}$$

Equation (8.46) yields the amplitude

$$\begin{aligned} f_x &= a_x(\epsilon) \\ &= (2\pi)^{-1/4} \exp(-x^2/4)\,. \end{aligned} \tag{8.49}$$

The linear transformation of the amplitude is

$$\begin{aligned} G_{\xi,\eta} &= G_{\eta,0} G_{0,\xi}\, x \\ &= \exp(i\eta\mathbf{g}_d) \exp(i\xi\mathbf{g}_t)\, x\,. \end{aligned} \tag{8.50}$$

The set of operators $G_{\xi,\eta}$, obtained when ξ and η run over the real axis, forms a group because the commutator of the generators is

$$\begin{aligned}[\mathbf{g}_t, \mathbf{g}_d] &= -\frac{\partial}{\partial x}\, x\, \frac{\partial}{\partial x} + x \frac{\partial}{\partial x}\, \frac{\partial}{\partial x} \\ &= -\frac{\partial}{\partial x} - x\, \frac{\partial^2}{\partial x^2} + x\, \frac{\partial^2}{\partial x^2} \\ &= i\mathbf{g}_t \end{aligned} \tag{8.51}$$

which indeed is a linear combination of the generators $\mathbf{g}_t,\ \mathbf{g}_d$ multiplied by i. We note that a combination of groups, the generators of which commute, certainly forms a group.

In the next chapter, the prior distribution is defined for models that lack form invariance.

Chapter 9
Going Beyond Form Invariance: The Geometric Prior

There is a formula that yields the invariant measure of a form-invariant model $p(x|\xi)$ in a straightforward way without analysis of the symmetry group, in particular without knowledge of the multiplication function. This is useful because the analysis of the symmetry group may be difficult. This is even of basic importance, because the formula allows one to generalise the definition of the prior μ to cases where form invariance does not exist or is not known to exist. Thus we do not require form invariance for the application of Bayes' theorem.

In Sect. 9.1, the formula is given and we show that it does yield the invariant measure (in case that form invariance exists). The formula is shown in Sects. 9.2 and 9.3 to be proportional to the geometric measure μ_g on a curved surface. Examples of geometric measures are given in Sect. 9.4. Section A.9 gives the solutions to the problems suggested to the reader.

Differential geometry has been introduced into statistics by Kass [1, 2], Amari [3], and others [4–9]; see also the collection [12].

9.1 Jeffreys' Rule

The measure $\mu(\xi)$ in the parameter space of a model p is given by the determinant of the so-called Fisher matrix [13] F such that

$$\mu(\xi) = (\det F)^{1/2} \ . \tag{9.1}$$

This is called Jeffreys' rule.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_9

For n-dimensional $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$, as in (4.15), the Fisher matrix is n-dimensional. Its elements[1] estimate the second derivatives of $\ln p$,

$$\begin{aligned} F_{\nu,\nu'} &= -\int \mathrm{d}x \; p(x|\boldsymbol{\xi}) \frac{\partial^2}{\partial\xi_\nu \partial\xi_{\nu'}} \ln p(x|\boldsymbol{\xi}) \\ &= \int \mathrm{d}x \; p(x|\boldsymbol{\xi}) \left(\frac{\partial}{\partial\xi_\nu} \ln p(x|\boldsymbol{\xi}) \right) \frac{\partial}{\partial\xi_{\nu'}} \ln p(x|\boldsymbol{\xi}) \; . \end{aligned} \tag{9.2}$$

The two lines of this equation are equal to each other because p is normalised to unity for every ξ; see Sect. D.1. Equation (9.1) was suggested by Jeffreys [14, 15] because under reparameterisations it behaves as a density. The proof is left to the reader. Jeffreys did not refer to a symmetry of p; he emphasised that (9.1) can be interpreted as a geometric measure.

Equation (6.41) allows us to write the Fisher matrix in terms of the vector $a(\xi)$ of probability amplitudes

$$a_x(\xi) = \sqrt{p(x|\xi)} \, , \tag{9.3}$$

see also Sect. D.1. We find

$$\begin{aligned} \frac{1}{4} F_{\nu,\nu'} &= \int \mathrm{d}x \, \left(\frac{\partial}{\partial\xi_\nu} a_x(\xi) \right) \frac{\partial}{\partial\xi_{\nu'}} a_x(\xi) \\ &= \left(\frac{\partial}{\partial\xi_\nu} a(\xi) \right)^\dagger \left(\frac{\partial}{\partial\xi_{\nu'}} a(\xi) \right) . \end{aligned} \tag{9.4}$$

The last expression will yield the geometric measure as is explained in Sect. 9.2. The interested reader should show that the eigenvalues of F are nonnegative. Hence, the square root in (9.1) is real. If ξ is one-dimensional, $F^{1/2}$ equals the inverse statistical error that is conventionally assigned to the estimated value ξ^{ML}.

The measure (9.1) is invariant [1] under the transformations G_ξ in the symmetry group $\mathcal{G}$ if the model p is form invariant. We prove this. Some notational conventions are needed. The Jacobian matrix

$$\frac{\partial G_\rho \xi}{\partial \xi}$$

of derivatives is introduced. It has the elements

$$\left(\frac{\partial G_\rho \xi}{\partial \xi} \right)_{\nu\nu'} = \frac{\partial (G_\rho \xi)_{\nu'}}{\partial \xi_\nu} \, . \tag{9.5}$$

[1] The present definition (9.2) of the Fisher matrix differs from the definition given in Eq. (9.2) in the first edition of this book. The present definition is commonly used; it defines F to be larger by a factor of 4 than the definition in the first edition.

Furthermore we use the vector of partial derivatives

$$\left(\frac{\partial}{\partial\xi}\right) = \begin{pmatrix} \partial/\partial\xi_1 \\ \vdots \\ \partial/\partial\xi_n \end{pmatrix} . \tag{9.6}$$

The dyadic product

$$\left(\frac{\partial}{\partial\xi}\right)\left(\frac{\partial}{\partial\xi}\right)^\dagger$$

is the matrix of second derivatives

$$\left(\left(\frac{\partial}{\partial\xi}\right)\left(\frac{\partial}{\partial\xi}\right)^\dagger\right)_{\nu\nu'} = \frac{\partial}{\partial\xi_\nu}\,\frac{\partial}{\partial\xi_{\nu'}} . \tag{9.7}$$

By help of these notations the Fisher matrix (9.4) can be written

$$\frac{1}{4}F = \left(\frac{\partial}{\partial\xi}\right)\left(\frac{\partial}{\partial\xi'}\right)^\dagger \; a^\dagger(\xi)a(\xi')\big|_{\xi=\xi'} . \tag{9.8}$$

An implicit derivation with respect to the components of ξ is expressed as

$$\left(\frac{\partial}{\partial_\xi}\right) = \frac{\partial G_\rho\xi}{\partial\xi}\left(\frac{\partial}{\partial_{G_\rho\xi}}\right) . \tag{9.9}$$

With these conventions the invariance of (9.1) under the group $\mathcal{G}$ can be shown in a few steps.

$$\begin{aligned} \mu(\xi) &= \det\left(\left(\frac{\partial}{\partial_\xi}\right)\left(\frac{\partial}{\partial_{\xi'}}\right)^\dagger a^\dagger(\xi)\mathbf{G}_\rho^\dagger\mathbf{G}_\rho a(\xi')\right)^{1/2}\Bigg|_{\xi=\xi'} \\ &= \det\left(\left(\frac{\partial}{\partial_\xi}\right)\left(\frac{\partial}{\partial_{\xi'}}\right)^\dagger a^\dagger(G_\rho\xi)a(G_\rho\xi')\right)^{1/2}\Bigg|_{\xi=\xi'} \\ &= \det\left(\left(\frac{\partial}{\partial_{G_\rho\xi}}\right)\left(\frac{\partial}{\partial_{G_\rho\xi'}}\right)^\dagger a^\dagger(G_\rho\xi)a(G_\rho\xi')\right)^{1/2}\Bigg|_{\xi=\xi'} \det\left(\frac{\partial G_\rho\xi}{\partial\xi}\right) \\ &= \mu(G_\rho\xi)\left|\frac{\partial G_\rho\xi}{\partial\xi}\right| . \end{aligned} \tag{9.10}$$

In all of this chain of equations, ξ is set equal to ξ' after the differentiations. The first line of (9.10) is obtained from (9.8) because the transformation $\mathbf{G}_\rho$ is orthogonal. For the second line, form invariance, as formulated in (8.21), has been used. Then (9.9) has been used. The notation $|A|$ means the absolute value of the determinant of the matrix A. As a result, μ is invariant under all transformations in $\mathcal{G}$. Therefore it is the invariant measure μ of the symmetry group of a form-invariant model p, compare Refs. [1, 16].

Expression (9.1) uses only the local properties of p with respect to ξ, not the global properties of a group; see the discussion in Sect. 7.4. There, a definition of the prior distribution was asked for that would not require form invariance but would be compatible with the invariant measure if form invariance existed. Expression (9.1) achieves this.

We note that the Fisher matrix (9.2) is not only used to define the prior distribution, it is also used to define the correlation matrix of the Gaussian approximation to the posterior of $p(x|\boldsymbol{\xi})$. When the posterior can be considered Gaussian then the Fisher matrix is

$$\begin{aligned} F_{\nu,\nu'} &= \int \mathrm{d}x\; p(x|\xi) \frac{\partial^2}{\partial \xi_\nu \partial \xi'_{\nu'}} \left((x-\xi)^\dagger (2C)^{-1} (x-\xi') \right) \Big|_{\xi=\xi'} \\ &= \left(C^{-1} \right)_{\nu,\nu'} \;, \end{aligned} \tag{9.11}$$

whence, F then is the inverse of the correlation matrix C.

9.2 Geometric Interpretation of the Prior μ

Equation (9.8) is a generalisation of the geometric measure (6.38). The two expressions are identical if a is two-dimensional and depends on a one-dimensional parameter ξ. In order to justify the generalisation, we use some ideas of differential geometry in the present section.

Whether a is an element of a function space or an element of a vector space is immaterial in what follows. In the case of a vector space, the variable x that labels the components of a, is discrete. In connection with geometry we prefer to speak of the vector or even of the point a because these are familiar notions of analytical geometry.

The set of points $a(\xi)$ that is obtained when ξ runs over all its values, is a curve for one-dimensional ξ, and it is a surface for two-dimensional ξ, and it is a hypersurface for n-dimensional ξ. We call it a surface in all cases here. It is embedded in the space wherein the vectors a are defined.

To derive the measure on the surface we consider, in the space of the parameters ξ, the cube

$$[\xi_\nu,\ \xi_\nu + \mathrm{d}\xi_\nu]\,, \quad \nu = 1, \ldots, n\,.$$

It is imaged onto an n-dimensional parallelepiped by the mapping $\xi \to a(\xi)$. Let us define the infinitesimal shift

$$\Delta_\nu = (0, \dots 0, \mathrm{d}\xi_\nu, 0, \dots 0)\,, \quad \nu = 1, \dots, n\,.$$

The parallelepiped is spanned by the vectors

$$\begin{aligned} a(\xi + \Delta_\nu) - a(\xi) \approx \frac{\partial a(\xi)}{\partial \xi_\nu}\mathrm{d}\xi_\nu\ , \\ \nu = 1, \dots, n\ . \end{aligned} \tag{9.12}$$

These n vectors are tangential to the surface. They span the tangent space at $a(\xi)$. We want to calculate the volume of the parallelepiped. When the dimension of the space in which $a(\xi)$ is defined, were n, then the vectors (9.12) could form an n-dimensional matrix J, and its determinant

$$|J|\,\mathrm{d}\xi_1 \dots \mathrm{d}\xi_n$$

would be equal to the volume of the parallelepiped. Here, the vectors

$$\frac{\partial a(\xi)}{\partial \xi_\nu}\ , \quad \nu = 1, \dots, n\ .$$

would be the columns of the matrix J. However, the dimension of $a(\xi)$ is higher than n; it is infinite. In this case we cannot write down the matrix J but we can obtain the matrix $JJ^\dagger$ by the following arguments.

Let us introduce a system of orthonormal vectors e_ν , $\nu = 1, \dots, n$, that spans the tangent space at $a(\xi)$. One then has

$$\begin{aligned} \frac{\partial a}{\partial \xi_\nu} = \sum_{\nu'} e_{\nu'} e^\dagger_{\nu'} \frac{\partial a}{\partial \xi_\nu}\ , \\ \nu = 1, \dots, n\,, \end{aligned} \tag{9.13}$$

which means that the ν'th expansion coefficient $J_{\nu\nu'}$ of the νth tangent vector is

$$J_{\nu\nu'} = e^\dagger_{\nu'} \frac{\partial a}{\partial \xi_\nu}\ . \tag{9.14}$$

From this follows

$$\begin{aligned} \left(JJ^\dagger\right)_{\nu\nu'} &= \sum_{\nu''} \frac{\partial a^\dagger}{\partial \xi_\nu} e_{\nu''} e^\dagger_{\nu''} \frac{\partial a}{\partial \xi_{\nu'}} \\ &= \frac{\partial a^\dagger}{\partial \xi_\nu} \frac{\partial a}{\partial \xi_{\nu'}}\ . \end{aligned} \tag{9.15}$$

For the first line of this equation, we have used the identity $a^\dagger e = e^\dagger a$ in order to rewrite the inner product of two real vectors. The second line results from the fact that

$$\sum_\nu e_\nu e_\nu^\dagger = \mathbf{1}_n \tag{9.16}$$

is the unit operator in the n-dimensional tangent space at $a(\xi)$. The matrix with the elements (9.15) takes the form

$$\begin{aligned} J J^\dagger &= \left(\frac{\partial}{\partial_\xi}\right)\left(\frac{\partial}{\partial_{\xi'}}\right)^\dagger a^\dagger(\xi) a(\xi')\Bigg|_{\xi=\xi'} \\ &= \frac{1}{4} F \end{aligned} \tag{9.17}$$

in the notation used in (9.8). The volume of the parallelepiped is

$$\left(\det(J J^\dagger)\right)^{1/2} \mathrm{d}\xi_1 \ldots \mathrm{d}\xi_n \,,$$

and the measure on the surface consisting of the points $a(\xi)$ is

$$\begin{aligned} \mu_g(\xi) &= \left(\det(J J^\dagger)\right)^{1/2} \\ &= \det\left(\frac{1}{4} F\right)^{1/2} \\ &= \det\left(\frac{\partial}{\partial \xi} a^\dagger(\xi) \frac{\partial}{\partial \xi'} a^{(\xi')}\Bigg|_{\xi=\xi'}\right)^{1/2} . \end{aligned} \tag{9.18}$$

By Eq. (9.17) it is consistent with Eq. (9.1). In this sense we confirm Jeffreys' claim of a geometric measure [1, 2, 8] on the surface with the parameter representation $a(\xi)$.

Note that the geometric measure is exactly determined; there is no arbitrary factor as there is in the invariant measure (6.33). This is due to the basis vectors e_ν of the tangent space which are normalised to unity. We have made use of the absolute definition of the geometric measure in connection with Fig. 4.3.

As an example, consider the representation $a(\omega)$ of the surface of a sphere of unit radius in an M-dimensional space; that is,

$$\begin{aligned} a_1 &= \xi_1 \,, \\ a_2 &= \xi_2 \,, \\ &\vdots \\ a_M &= \left(1 - \sum_{k=1}^{M-1} \xi_k^2\right)^{1/2} . \end{aligned} \tag{9.19}$$

Here, the parameter $\boldsymbol{\xi}$ has the $M-1$ components $\xi_1, \dots, \xi_{M-1}$. The measure on the surface of the sphere is

$$\mu(\boldsymbol{\xi}) = \left(1 - \sum_{k=1}^{M-1} \xi_k^2\right)^{-1/2} . \tag{9.20}$$

The interested reader may prove this result. A determinant given in Sect. D.2 is helpful. We find this measure as the prior distribution of the models treated in Sect. 9.4. The surface of a sphere is finite, and (9.20) is a proper measure.

Often the measure on the surface of a sphere appears in another parameterisation. Let the set of parameters be

$$\begin{aligned} \eta_k &= \xi_k^2 \\ k &= 1, \dots, M-1 . \end{aligned} \tag{9.21}$$

Through this transformation, the measure (9.20) becomes

$$\mu_T(\eta) = 2 \prod_{k=1}^{M} \left(\frac{1}{2}\eta_k^{-1/2}\right) , \tag{9.22}$$

where

$$\eta_M = 1 - \sum_{k=1}^{M-1} \eta_k . \tag{9.23}$$

The interested reader should derive this form of the measure.

9.3 On the Geometric Prior Distribution

We have taken the form-invariant models as the starting point and have identified the prior distribution with the invariant measure of the symmetry group. The prior of a model without form invariance must be a generalisation of this: it shall merge into the invariant measure when form invariance exists. This is achieved by using (9.1) to define the prior.

Still, the discussion of form invariance is necessary. The notion of geometric measure is not well defined by itself. It is a measure on a curved surface. We need to define the space in which this surface is embedded. Form invariance has led us to conclude that this surface consists of the "points" $a(\xi)$ in the linear space of square integrable functions.

From now on we do not ask whether the prior distribution is an invariant measure or a geometric measure, unless the context requires this distinction. Expression (9.1) is the general definition of the prior.

In Chap. 2, we required that the model $p(x|\xi)$ be proper so that it is normalised to unity for every ξ. This ensures that the geometric measure exists. Indeed, it follows that $a^\dagger(\xi)a(\xi')$ exists. This entails the existence of the Fisher matrix (9.8) when we take for granted that the amplitude $a_x(\xi)$ is a regular or even analytical function of the parameters ξ.

From Sect. 6.4, we know that the invariant measure does not vanish. When there is no form invariance, we must require that the geometric measure not vanish, neither identically nor at isolated points. At such a point it would not depend on all of its parameters.

Form invariance favors the existence of the ML estimator $\sigma^{\mathrm{ML}}(x)$ for every x foreseen by the model p. Whether or not p is form invariant, we require the existence of $\sigma^{\mathrm{ML}}(x)$.

Jeffreys' rule yields the geometric measure of models with continuous event variable x and of models with discrete x. Equation (9.8) is the geometric measure on the surface $a(\xi)$ irrespective of whether the inner product $a^\dagger a$ involves an integration or a summation over x. For continuous x we have shown in Sect. 9.1 that Jeffreys rule agrees with the invariant measure when form invariance is given. In the case of discrete x, we - in Chap. 11 - define form invariance such that expression (9.8) again yields the invariant measure.

9.4 Examples of Geometric Priors

The geometric priors of two models are calculated. The models are similar in that both of them contain an expansion of a probability amplitude in terms of orthogonal vectors. The expansion coefficients ω_ν are the hypothesis parameters. The models are different in that the first one has a continuous event variable x, whereas the second one has a discrete x.

9.4.1 An Expansion in Terms of Orthogonal Functions

Let the model

$$p(x|\omega) = a_x^2(\omega)\,, \quad x \text{ real}\,,\ \omega = (\omega_1, \dots, \omega_n)\,, \tag{9.24}$$

be defined as the square of the amplitude

$$a_x(\omega) = \sum_{\nu=1}^{n} \omega_\nu c_x(\nu)\ . \tag{9.25}$$

Here, x is continuous. Equation (9.25) is an expansion of a into the n basis functions $c(\nu)$. The expansion coefficients ω_ν are inferred from the events. One can think of an expansion in terms of trigonometric functions or orthogonal polynomials. One cannot infer an infinity of parameters; therefore an infinite basis of the expansion does not make sense.

The parameters ω_ν depend on each other because they are normalised according to

$$\sum_{\nu=1}^{n} \omega_\nu^2 = 1 \tag{9.26}$$

which guarantees the normalisation of $p(x|\omega)$ provided that the basis functions $c(\nu)$ are orthogonal to each other.

The model (9.24), (9.25) is not form invariant. The prior is the geometric measure (9.18). The Fisher matrix is evaluated in Sect. D.3 with the result

$$\begin{aligned} \frac{1}{4} F_{\nu\nu'} &= \delta_{\nu\nu'} + \frac{\omega_\nu \omega_{\nu'}}{\omega_n^2} \\ \nu\,, \nu' &= 1, \dots, n-1\ . \end{aligned} \tag{9.27}$$

This is the sum of the unit matrix and a dyadic product. With the help of a rule established in Sect. D.2, one obtains

$$\mu(\omega_1, \dots, \omega_{n-1}) = \left(1 - \sum_{\nu=1}^{n-1} \omega_\nu^2 \right)^{-1/2}\ . \tag{9.28}$$

This is the geometric measure on the surface of a sphere in n-dimensional space; see (9.20). The result is not surprising in view of the normalisation (9.26). Note that (9.28) is the measure on a sphere of unit radius in n-dimensional space when we interpret $\omega_1, \dots, \omega_n$ as the Cartesian coordinates of the points on the sphere. The ω_ν are restricted to the domain $|\omega_\nu| \leq 1$. When we look for a linear representation of the present model then Eq. (8.30) lets us expect that there should not be any such restriction of the domain of definition of the parameters. Indeed, the ω_ν can be expressed by trigonometric functions of a set of angles. Their periodicity removes the above restriction although it respects the restriction in the sense that shifting an angular parameter by one period leaves the model unaltered.

The next subsection shows that the prior of the multinomial distribution is formally identical with this result.

9.4.2 The Multinomial Model

The multinomial model, introduced in Sect. 5.2, is an M-fold alternative. It gives the joint probability

$$p(x_1, \ldots, x_{M-1}|\eta_1, \ldots, \eta_{M-1}) = N! \prod_{k=1}^{M} \frac{\eta_k^{x_k}}{x_k!} \tag{9.29}$$

to find the kth state realised x_k times, $k = 1, \ldots, M$, when N events are collected. The hypothesis η specifies the probabilities η_k, $k = 1, \ldots, \eta_{M-1}$, for an event to realise the state k. One of the M choices must be realised so that (5.10) holds. This relation defines η_M.

We introduce the amplitudes

$$\begin{aligned} \beta_k &= \sqrt{\eta_k} \\ k &= 1, \ldots, M \ . \end{aligned} \tag{9.30}$$

These amplitudes form the M-dimensional vector β. In analogy with the last subsection, we expand β in a system of $n \leq M$ basis vectors $c(\nu)$, $\nu = 1, \ldots, n$, so that

$$\beta_k = \sum_{\nu=1}^{n} \omega_\nu c_k(\nu) \ . \tag{9.31}$$

Hence, the model now reads

$$p(x|\omega) = N! \prod_{k=1}^{M} \frac{\beta_k^{2x_k}}{x_k!} \ , \tag{9.32}$$

where the β_k are linear functions of the hypothesis parameters according to (9.31). The analogy with the model (9.24), (9.25) is obvious.

The system of basis vectors $c(\nu)$ is required to be orthonormal; that is,

$$\begin{aligned} c^\dagger(\nu)c(\nu) &= \delta_{\nu\nu'} \ , \\ \nu, \nu' &= 1, \ldots, n \ , \end{aligned} \tag{9.33}$$

so the normalisation (5.10) is ensured when (9.26) holds. This relation defines ω_n. The quantities $\omega = (\omega_1, \ldots, \omega_{n-1})$ are considered as the parameters of (9.32).

This model is not form invariant. The prior distribution is given by the geometric measure (9.18). The Fisher matrix F is found in Sect. D.4 to have the elements

$$\begin{aligned} \frac{1}{4} F_{\nu\nu'} &= N \left(\delta_{\nu\nu'} + \frac{\omega_\nu \omega_{\nu'}}{\omega_n^2} \right) \ , \\ \nu, \nu' &= 1, \ldots, n-1 \ . \end{aligned} \tag{9.34}$$

Using the determinant calculated in Sect. D.2, we obtain the prior

$$\mu(\omega) = \left(1 - \sum_{\nu=1}^{n-1} \omega_\nu^2\right)^{-1/2} \tag{9.35}$$

in formal agreement with the result of the last subsection. The similarity lies in the expansions (9.25), (9.31), and the normalisation condition (9.26).

In the special case where all the probabilities for the M basic states of the alternative shall be inferred, one obtains the prior distribution (9.35) with $n = M$. When this is transformed to the parameters η of Eq. (9.29), one obtains (9.22).

In the following chapter, we turn again to a form-invariant model.

References

1. R.E. Kass, The Riemannian Structure of Model Spaces: A Geometrical Approach to Inference. Ph.D. thesis, University of Chicago, 1980. See especially pp. 94–95
2. R.E. Kass, The geometry of asymptotic inference. Stat. Sci. **4**, 188–219 (1989)
3. S.I. Amari, *Differential Geometrical Methods in Statistics*, Lecture Notes in Statistics, vol. 28 (Springer, Heidelberg, 1985)
4. C. Radakrishna Rao, Differential metrics in probability spaces based on entropy and divergence measures. Technical Report, Report No. AD-A160301, AFOSR-TR-85-0864, Air Force Office of Scientific Research, Bolling AFB, DC (1985)
5. R.D. Levine, Geometry in classical statistical thermodynamics. J. Chem. Phys. **84**, 910–916 (1986)
6. C. Radakrishna Rao, Differential metrics in probability spaces, in Shun ichi Amari [12], pp. 218–240
7. C.C. Rodriguez, Objective Bayesianism and geometry, in Fougère [10], pp. 31–39
8. C. Villegas, Bayesian inference in models with Euclidean structures. J. Am. Stat. Assoc. **85**(412), 1159–1164 (1990)
9. C.C. Rodriguez, Are we cruising a hypothesis space? in von der Linden et al. [11], pp. 131–139
10. P.F. Fougère (ed.), *Maximum Entropy and Bayesian Methods, Dartmouth, 1989* (Dordrecht, Kluwer,1990)
11. W. von der Linden, V. Dose, R. Fischer, R. Preuss (eds.), *Maximum Entropy and Bayesian Methods, Garching, 1998* (Dordrecht, Kluwer, 1999)
12. S.I. Amari (ed.), *Differential Geometry in Statistical Inference*. Lecture Notes and Monograph Series, vol. 10 (Institute of Mathematical Statistics, Hayward, California, 1987)
13. R.A. Fisher, Theory of statistical information. Proc. Camb. Philos. Soc. **22**, 700–725 (1925)
14. H. Jeffreys, *Theory of Probability* (Oxford University Press, Oxford, 1939) (2nd ed. 1948; 3rd ed. 1961, here Jeffreys' rule is found in III §3.10)
15. H. Jeffreys, An invariant form of the prior probability in estimation problems. Proc. R. Soc. A **186**, 453–461 (1946)
16. T. Chang, C. Villegas, On a theorem of Stein relating Bayesian and classical inferences in group models. Can. J. Stat. **14**(4), 289–296 (1986)

Chapter 10
Inferring the Mean or the Standard Deviation

We assume that a set of events $x_1, \dots, x_N$ is given. Neither the mean ξ nor the standard deviation σ of the data is known. They are to be inferred. It is assumed that the Gaussian model

$$q(x|\xi, \sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) \tag{10.1}$$

of Eq. (4.1) applies for the distribution of every x_k. In Sect. 10.1 of the present chapter, inference of both parameters is described. Often, however, one is not interested in both parameters of the model but in only one of them, say the mean ξ. The model p, however, depends on σ too. The data allow us to determine ξ and σ, but our interest is focused on ξ, whatever the value of σ. This case is treated in Sect. 10.2. In Sect. 10.3, the standard deviation σ is inferred, whatever the value of ξ. Section 10.4 treats an argument intended to reveal an inconsistency of the ML estimator [1]. In the framework of Bayesian statistics the inconsistency disappears.

10.1 Inferring Both Parameters

Figure 10.1 shows some data from nuclear physics [2, 3]; compare Sect. 3.2.2. Longitudinally polarised neutrons have been captured into a long-lived state, a resonance. The absorption cross-section is found to depend on the sign of the polarisation. This violates parity conservation and is therefore interesting. The asymmetry

$$x = \frac{a^+ - a^-}{a^+ + a^-}$$

between the absorption cross-sections $a^\pm$ for the two polarisations is given at 24 resonances. The abscissa specifies the neutron energies at which the resonances are found. It was expected that x would have a mean value of zero [4]. Only the variance

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_10

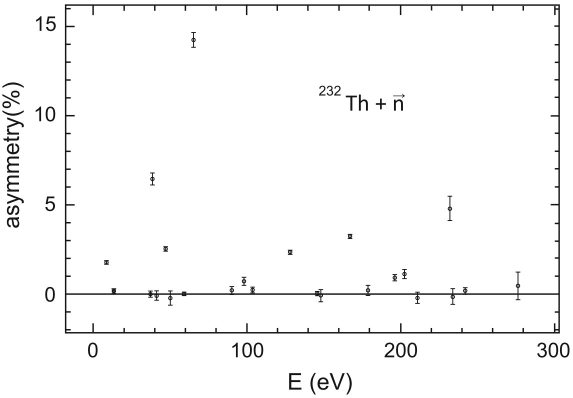


Fig. 10.1 The parity violating asymmetry x for the absorption cross-sections of longitudinally polarised neutrons [2]. According to the statistical theory of nuclear reactions, the 24 events should be described by the Gaussian model (10.1). The error bars refer to systematic errors; they are not considered in the present context

seemed interesting. However, the figure shows that x is more frequently positive than negative. This caused a discussion about both parameters. The data are not analysed here, but the formalism is described.

According to Chap. 6, the model (10.1) is form invariant. The symmetry group is a non-Abelian combination of the translation and the dilation groups. In Chap. 7, the invariant measure was found to be

$$\mu(\xi, \sigma) \propto \sigma^{-1} . \tag{10.2}$$

As in (2.23), the distribution of N events is written in the form

$$\begin{aligned} p(\boldsymbol{x}|\xi, \sigma) &= \prod_{k=1}^{N} q(x_k|\xi, \sigma) \\ &= (2\pi\sigma^2)^{-N/2} \\ &\quad \times \exp\left(-\frac{N}{2\sigma^2}\left[(\xi - \langle x \rangle)^2 + V\right]\right) , \end{aligned} \tag{10.3}$$

where

$$V = \langle x^2 \rangle - \langle x \rangle^2 . \tag{10.4}$$

Here, $\langle x^2 \rangle$ and $\langle x \rangle$ are averages over the x_k^2 and x_k as defined in Eq. (2.21). This gives the posterior distribution

$$\begin{aligned} P(\xi, \sigma|\boldsymbol{x}) = {} & \frac{N^{(N+1)/2}}{\pi^{1/2}\, 2^{(N-1)/2}\, \Gamma(N/2)} V^{N/2} \\ & \times \sigma^{-N-1} \exp\left(-\frac{N}{2\sigma^2}\left[(\xi - \langle x \rangle)^2 + V\right]\right) . \end{aligned} \tag{10.5}$$

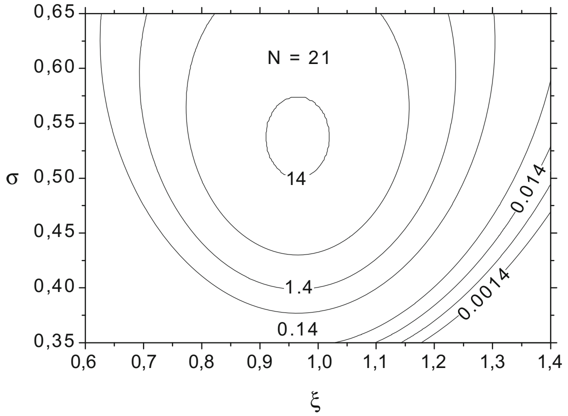


Fig. 10.2 Posterior of $N = 21$ events drawn from the distribution (10.1). A random number generator was used with the true values of ξ and σ set to 1 and 0.5, respectively

The normalisation is verified in Sect. E.1 as well as Problem A.3.9. The properties of the Γ function can be found in Sect. B.4.

We have simulated the model (10.3) by a computer experiment. A random number generator was used to draw 101 events from the distribution (10.1) with ξ set equal to 1 and σ equal to 0.5. Of course, in a real experiment, these true values are unknown. They are to be inferred. From the first $N = 21$ events of the computer experiment, we find the posterior distribution (10.5) displayed in Fig. 10.2. The result from the first $N = 51$ events is shown in Fig. 10.3. The posterior obtained from all $N = 101$ events is given in Fig. 10.4. These three figures show how the result improves with increasing N in a case of two-parameter inference. The distribution contracts in both dimensions. It tends towards a two-dimensional δ function at the position of the true parameter values. From the result for $N = 21$ events, one clearly sees that for finite N, the distribution is not necessarily centered at the true values. It is centered at the coordinates $(\xi^{\rm ML}, \sigma^{\rm ML})$. This describes the maximum of the likelihood function.

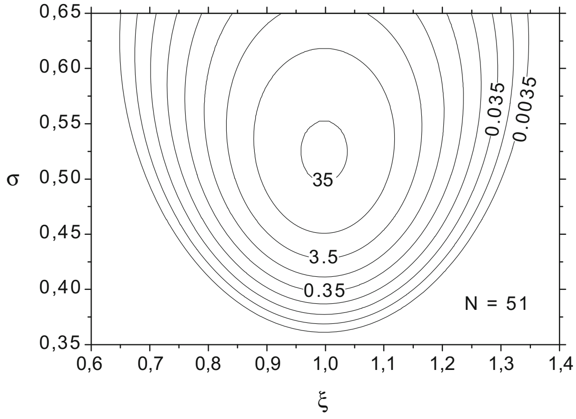


Fig. 10.3 Posterior of $N = 51$ events drawn from the distribution (10.1) as in Fig. 10.2

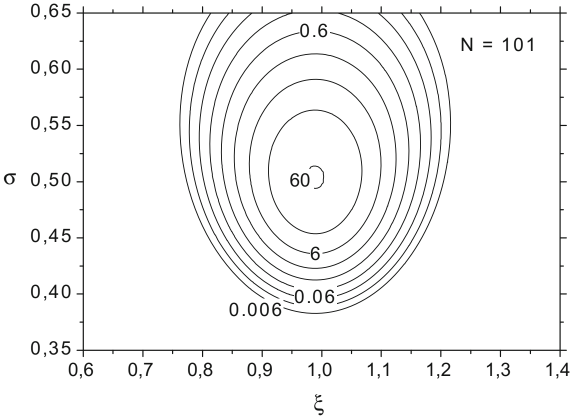


Fig. 10.4 Posterior of $N = 101$ events drawn from the distribution (10.1) as in Fig. 10.2

The likelihood function L is defined in Eq. (6.46). Together with Eqs. (10.2), (10.3) we obtain

$$L(\xi, \sigma|x) \propto \sigma^{-N} \exp\left(-\frac{N}{2\sigma^2}\left((\xi - \langle x\rangle)^2 + V\right)\right) \tag{10.6}$$

for the model (10.3). The system of ML equations is

$$\begin{aligned} 0 &= \frac{\partial}{\partial \xi} \ln L\,, \\ 0 &= \frac{\partial}{\partial \sigma} \ln L\,. \end{aligned} \tag{10.7}$$

Their solution determines the two-dimensional ML estimator $(\xi^{\mathrm{ML}}, \sigma^{\mathrm{ML}})$. We find the set of ML equations

$$\begin{aligned} 0 &= \frac{N}{\sigma^2}(\xi - \langle x\rangle)\,, \\ 0 &= -\frac{N}{\sigma} + \frac{NV}{\sigma^3} \end{aligned} \tag{10.8}$$

which yields

$$\begin{aligned} \xi^{\mathrm{ML}} &= \langle x\rangle\,, \\ (\sigma^{\mathrm{ML}})^2 &= V\,. \end{aligned} \tag{10.9}$$

The Bayesian area - defined in Chap. 3 - is bordered by a contour line of the likelihood function $L = P/\mu$. Hence, in any one of the present cases, the border of the Bayesian area will be close to one of the ellipsoidal curves in Figs. 10.2, 10.3 and 10.4. This means that in higher dimensions, the error “interval” is a rather complicated object. A rectangle of the kind of $[\xi^{\mathrm{ML}} \pm \Delta\xi, \sigma^{\mathrm{ML}} \pm \Delta\sigma]$ is too rough an approximation.

Often one is not interested in knowing both parameters but instead only one of them, say ξ; then one can think of projecting the two-dimensional distribution P onto the ξ-axis. This projection is called the "marginal distribution"

$$P^{\downarrow}(\xi|\boldsymbol{x}) = \int \mathrm{d}\sigma \, P(\xi, \sigma|\boldsymbol{x}) \, . \tag{10.10}$$

This usually is hard to obtain. However, for a sufficiently large number N of events a Gaussian approximation to the posterior helps. Its marginal distributions can be obtained by use of the "simple rule" stated in Sect. 4.1.2 and proven in Sect. B.1. Thus multidimensional distributions can be handled analytically when they are approximated by a Gaussian. This is described in what follows.

10.1.1 Multidimensional Gaussian Approximation to a Posterior Distribution

The Gaussian approximation to a one- or more-dimensional posterior distribution is given by setting the inverse of the correlation matrix C equal to the Fisher matrix F of the model. When F depends on the parameters of the model then these are set equal to their ML estimators and F is considered constant for the sake of the approximation. The ML estimator has been introduced in Sect. 6.6 and discussed in Sect. 6.7.

The Fisher matrix has been defined in Sect. 9.1. It generalises the Fisher information introduced by Eq. (4.13). The latter one estimates the inverse variance of a one-dimensional distribution as explained in Sect. 4.1.2. The estimation is valid for sufficiently large values of the number N of events. When there is more than one parameter the Fisher matrix estimates the inverse C^{-1} of the correlation matrix of the Gaussian approximation to the distribution. Again the approximation is valid for sufficiently large N.

The elements of the Fisher matrix equal the expectation values of the negative second derivatives of $\ln q(x|\xi, \sigma)$ with respect to the parameters. Here, q is the elementary model which for the present case is given in Eq. (10.1). The expectation values are taken with respect to the event variable x.

For the case at hand we find

$$-\ln q = \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{(x-\xi)^2}{2\sigma^2} \, . \tag{10.11}$$

The first derivatives are

$$\begin{aligned} -\frac{\partial}{\partial\xi}\ln q &= \frac{\xi - x}{\sigma^2}\,, \\ -\frac{\partial}{\partial\sigma}\ln q\,, &= \frac{1}{\sigma} - \frac{(x-\xi)^2}{\sigma^3}\,. \end{aligned} \tag{10.12}$$

The second derivatives are given by

$$\begin{aligned} -\frac{\partial^2}{\partial\xi^2}\ln q &= \frac{1}{\sigma^2}\,, \\ -\frac{\partial^2}{\partial\xi\partial\sigma}\ln q &= \frac{2}{\sigma^3}(x-\xi)\,, \\ -\frac{\partial^2}{\partial\sigma^2}\ln q &= -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}(x-\xi)^2\,. \end{aligned} \tag{10.13}$$

The expectation values of the second derivatives turn out to be

$$\begin{aligned} -\overline{\frac{\partial^2}{\partial\xi^2}\ln q} &= \frac{1}{\sigma^2}\,, \\ -\overline{\frac{\partial^2}{\partial\xi\partial\sigma}\ln q} &= 0\,, \\ -\overline{\frac{\partial^2}{\partial\sigma^2}\ln q} &= \frac{2}{\sigma^2}\,. \end{aligned} \tag{10.14}$$

Thus the Fisher matrix of the elementary model q is

$$\begin{aligned} F &= -\int \mathrm{d}x\, q(x|\xi,\sigma) \begin{pmatrix} \frac{\partial^2}{\partial\xi^2}\ln q\,, & \frac{\partial^2}{\partial\xi\partial\sigma}\ln q \\ \frac{\partial^2}{\partial\xi\partial\sigma}\ln q\,, & \frac{\partial^2}{\partial\sigma^2}\ln q \end{pmatrix} \\ &= \frac{1}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\,. \end{aligned} \tag{10.15}$$

The Fisher matrix of the model (10.3) is N times this result. For this reason the a posteriori distribution contracts in the directions of all its parameters with growing N.

Because $F = F(\sigma)$ is not constant but rather depends on σ, one sets $F(\sigma) = F(\sigma^{\mathrm{ML}})$ and, for the Gaussian approximation, one considers F as independent of the variables ξ, σ.

$$P(\xi,\sigma|\boldsymbol{x}) \propto exp\left(-(\xi-\xi^{\mathrm{ML}}, \sigma-\sigma^{\mathrm{ML}})\left(\frac{1}{2}NF(\sigma^{\mathrm{ML}}\right)\begin{pmatrix}\xi-\xi^{\mathrm{ML}} \\ \sigma-\sigma^{\mathrm{ML}}\end{pmatrix}\right)\,. \tag{10.16}$$

Note that, by Eq. (9.1), the prior distribution μ is constant when the Fisher matrix is constant. Then μ drops out of the posterior; see Eq. (2.6).

Inverting NF leads to the correlation matrix

$$C = \frac{(\sigma^{\mathrm{ML}})^2}{N} \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix} \tag{10.17}$$

for the Gaussian (10.16).

10.2 Inferring the Mean Only

In the sequel the "interesting" parameter is called structural and the "uninteresting" one is called incidental. With this nomenclature we follow previous publications on the subject, such as Refs. [1, 5]. For the present section let ξ be the structural parameter of the model (10.3) and σ the incidental one.

The posterior of ξ is obtained as the marginal distribution (10.10) of the Gaussian approximation (10.16) to $P(\boldsymbol{\xi}, \sigma|\boldsymbol{x})$. In this way the problem of the so-called marginalisation paradox is avoided, a problem encountered in the literature [6–10] as well as in the first edition of the present book. With the help of the rule given in Sect. 4.1.1 we obtain

$$P^{\downarrow}(\xi|\boldsymbol{x}) \approx (2\pi)^{-1/2} \frac{N^{1/2}}{\sigma^{\mathrm{ML}}} \exp\left(-\frac{(\xi - \xi^{\mathrm{ML}})^2}{2(\sigma^{\mathrm{ML}})^2} \right) . \tag{10.18}$$

The ML estimators are given in Eq. (10.9).

In the next section we interchange the roles of the parameters ξ and σ.

10.3 Inferring the Standard Deviation Only

Let now σ be the structural and ξ the incidental parameter. We project the Gaussian approximation (10.16) onto the σ-axis and obtain the posterior distribution

$$P^{\downarrow}(\sigma|\boldsymbol{x}) \approx \pi^{-1/2} \frac{N^{1/2}}{\sigma^{\mathrm{ML}}} \exp\left(-\frac{N(\sigma - \sigma^{\mathrm{ML}})^2}{(\sigma^{\mathrm{ML}})^2} \right) \tag{10.19}$$

of σ.

This approximation to the posterior of σ can be improved. We can transform σ such that the diagonal element $\overline{\frac{\partial^2}{\partial\sigma^2} \ln q}$ becomes independent of the (transformed) parameter. Then the Gaussian approximation will be useful already at a lower value of N than before: the distribution (10.19) must not be as narrow as to make the dependence of F on σ negligible. In other words: $P^{\downarrow}$ is already Gaussian before we replace σ by σ^{ML}.

The necessary transformation is

$$\sigma = e^{\eta} \,, \tag{10.20}$$

compare Sect. 2.2. Note that η is defined on the entire real axis, whereas σ is positive. Thus η serves a Gaussian approximation more easily than does σ.

The transformation leads to the second derivatives

$$\begin{aligned} -\frac{\partial^2}{\partial\xi^2} \ln q &= e^{-2\eta} \,, \\ -\frac{\partial^2}{\partial\xi\partial\eta} \ln q &= 2(\xi - x)e^{-2\eta} \,, \\ -\frac{\partial^2}{\partial\eta^2} \ln q &= 2(x-\xi)^2 e^{-2\eta} \end{aligned} \tag{10.21}$$

and further to the Fisher matrix

$$F = \begin{pmatrix} e^{-2\eta} & 0 \\ 0 & 2 \end{pmatrix} \,, \tag{10.22}$$

where the second diagonal element is independent of the parameters. The interested reader is asked to verify this.

There exists no transformation that would render the Fisher matrix independent of both parameters, ξ and σ (or η), inasmuch as the symmetry group of the model (10.1) is not Abelian. The best one can achieve is to make every diagonal element of the Fisher matrix independent of the parameter whose second derivative enters into it. For the model at hand this is achieved in Eq. (10.22).

10.4 The Argument of Neyman and Scott Against ML Estimation

Neyman and Scott [1] have raised a widely discussed argument against the ML estimator; see Ref. [11]. The argument is indended to show that an ML estimator does not necessarily converge towards the true value of its parameter, when the number of observations increases indefinitely. The argument is as follows.

Consider N series $i = 1, \ldots, N$ of observations $x_{i,j}$. Every series i comprises n observations $x_{i,j}$ with $j = 1, \ldots, n$ from a Gaussian distribution

$$q(x_{i,j}|\xi_i, \sigma) = (2\pi)^{-1/2}\sigma^{-1} \exp\Big(-(x_{i,j} - \xi_i)^2/2\sigma^2 \Big) \,. \tag{10.23}$$

Different series $i \neq i'$ have been obtained in different experiments with possibly different $\xi_i \neq \xi_{i'}$. Yet every series is conditioned by the same value of the variance

σ^2. This is the structural parameter. The centers ξ_i of the distributions q are the incidental parameters. The joint distribution of all $x_{i,j}$ is the product

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\xi},\sigma) &= \prod_{i=1}^{N}\prod_{j=1}^{n} q(x_{i,j}|\xi_i,\sigma) \\ &= (2\pi)^{-Nn/2}\sigma^{-Nn} \\ &\quad\times \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big((\xi_i-\langle x\rangle_i)^2+V_i\Big)\right). \end{aligned} \tag{10.24}$$

Section E.2 shows how to bring the product in the first line of this equation into the form of the second line.

The likelihood function of the model p is

$$L(\boldsymbol{\xi},\sigma|\boldsymbol{x}) \;\propto\; \sigma^{-Nn}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\Big((\xi_i-\langle x\rangle_i)^2+V_i\Big)\right). \tag{10.25}$$

Here, we have set

$$V_i = \langle x_{i,j}^2\rangle - \langle x_{i,j}\rangle^2 \tag{10.26}$$

and used averages taken for a given i,

$$\begin{aligned} \langle x\rangle_i &= n^{-1}\sum_{j=1}^{n} x_{i,j} \\ \langle x^2\rangle_i &= n^{-1}\sum_{j=1}^{n} x_{i,j}^2\,. \end{aligned} \tag{10.27}$$

The ML equations require that the logarithmic derivatives of L with respect to the parameter vanish; that is,

$$\begin{aligned} 0 &= \xi_i - \langle x\rangle_i\,, \qquad i=1,\dots,N\,, \\ 0 &= -\frac{Nn}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{N}\big((\xi_i-\langle x\rangle_i)^2+V_i\big)\,. \end{aligned} \tag{10.28}$$

From these ML equations one obtains

$$\begin{aligned} \xi_i^{\mathrm{ML}} &= \langle x\rangle_i\,, \qquad i=1,\dots,N\,, \\ (\sigma^{\mathrm{ML}})^2 &= \frac{1}{nN}\sum_{i=1}^{N}(\langle x^2\rangle_i - \langle x\rangle_i^2)\,. \end{aligned} \tag{10.29}$$

The authors of Ref. [1] calculated the expectation values of these estimators with respect to the distribution p of $\boldsymbol{x}$. For the structural parameter they find

$$\overline{(\sigma^{\text{ML}})^2} = \sigma^2(1 - 1/n)\,, \tag{10.30}$$

when σ is the true value. This result is independent of the number N of experiments. From every series of observations the ML value of σ^2 seems to underestimate the true value by the factor of $1 - 1/n$. This bias is not removed for any number N of experiments. Therefore the ML estimation of σ was called inconsistent in many publications; see the review [11]. The interested reader may confirm Eq. (10.30). In doing so it is useful to observe that the distribution (10.23) implies that

$$\begin{aligned} \sigma^2 &= \overline{x_{i,j}^2} - \overline{x_{i,j}}^{\,2}\,, \\ &= \overline{x_{i,j}^2} - \xi_i^2\,. \end{aligned} \tag{10.31}$$

Our point of view is: the expectation value (10.30) of $(\sigma^{\text{ML}})^2$ need not be equal to the true value of σ^2. The bias noticed by Neyman and Scott is not a valid argument against ML estimation.

We have seen in Sect. 10.3 that the width of the Bayesian posterior of σ shrinks with increasing N. The likelihood function (10.25) explains why this is so: the sum over i in the exponent is of the order of N.

The quantity σ^{ML} does not run away with $N \to \infty$; in other words the Bayesian procedure converges. It is not bound to converge towards the expectation value (10.30). The difference between σ^{ML} and the expectation value (10.30) changes when σ is transformed, for example, via Eq. (10.20). One can transform it in such a way that the difference vanishes. There is no “natural” parameterisation of σ, thus one cannot require σ^{ML} to converge towards any given expectation value. By definition, the Bayesian posterior distribution is the probability density of the place, where the true value lies. The scale of the posterior is free to transformations. Thus Bayesian statistics requires that σ^{ML} converges with $N \to \infty$. It does not require that σ^{ML} converges towards its expectation value. The value to which σ^{ML} converges is the true value by definition. In order to show that an ML estimator is not consistent, one would have to show that it does not converge for $N \to \infty$.

References

1. J. Neyman, E.L. Scott, Consistent estimates based on partially consistent observations. Econometrica **16**, 1–32 (1948)
2. S.L. Stephenson, J.D. Bowman, B.E. Crawford, P.P.J. Delheij, C.M. Frankle, M. Iinuma, J.N. Knudson, L.Y. Lowie, A. Masaike, Y. Matsuda, G.E. Mitchell, S.I. Pentilä, H. Postma, N.R. Roberson, S.J. Seestrom, E.I. Sharapov, Y.-F. Yen, V.W. Yuan, Parity nonconservation in neutron resonances in ^{232}Th. Phys. Rev. C **58**, 1236–1246 (1998)

3. G.E. Mitchell, J.D. Bowman, S.I. Penttilä, E.I. Sharapov, Parity violation in compound nuclei: experimental methods and recent results. Phys. Rep. — Rev. Sec. Phys. Lett. **354**:157–241 (2001) (The original publication on parity violation in ^{115}In is [12])
4. G.E. Mitchell, J.D. Bowman, H.A. Weidenmüller, Parity violation in the compound nucleus. Rev. Mod. Phys. **71**, 445–457 (1999)
5. J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat. **27**, 887–906 (1956)
6. A.P. Dawid, N. Stone, J.V. Zidek, Marginalisation paradoxes in Bayesian and structural inference. J. R. Stat. Soc. B **35**, 189–233 (1973)
7. J.M. Bernardo, Reference posterior distributions for Bayesian inference. J. R. Stat. Soc. B **41**, 113–147 (1979)
8. E.T. Jaynes, Marginalization and prior probabilities, in Zellner [13], pp. 43–87
9. A. Nicolaou, Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. J. R. Stat. Soc. B **55**, 377–390 (1993)
10. R. Mukerjee, D.K. Dey, Frequentist validity of posteriori quantiles in the presence of nuisance parameter: higher order asymptotic. Biometrika **80**(3), 499–505 (1993)
11. T. Lancaster, The incidental parameter problem since 1948. J. Econ. **95**, 391–413 (2000)
12. S.L. Stephenson, J.D. Bowman, F. Corvi, B.E. Crawford, P.P.J. Delheij,C.M. Frankle, M. Iinuma, J.N. Knudson, L.Y. Lowie, A. Masaike,Y. Masuda, Y. Matsuda, G.E. Mitchell, S.I. Pentilä, H. Postma, N.R. Roberson, S.J. Seestrom, E.I. Sharapov, H.M. Shimizu, Y.-F. Yen, V.W. Yuan, L. Zanini, Parity violation in neutron resonances in ^{115}in. Phys. Rev. C **61**, 045501/1–11 (2000)
13. A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics: Essays in Honour of H. Jeffreys*. Studies in Bayesian Econometrics, vol. 1 (North Holland, Amsterdam, 1980)

Chapter 11
Form Invariance II: Natural *x*

In the present chapter, the symmetry of form invariance - introduced in Chap. 6 - is generalised to discrete event variables x. In Chap. 6 it was assumed that both x and the hypothesis parameter ξ were real. If x is discrete and ξ is real, form invariance cannot exist in the sense of (6.23) because there is no transformation of x that can account for an infinitesimal transformation of ξ. Hence, the definition of form invariance given in Chap. 6 does not cover the case of discrete x. The linear representation (8.20) of form invariance found in Chap. 8 provides the starting point for the more general definition.

Let the model $p(x|\xi)$ be given with discrete x and real (continuous) ξ. Let $a(\xi)$ be the vector with the components

$$a_x(\xi) = \sqrt{p(x|\xi)}\ . \tag{11.1}$$

The model is form invariant if $a(\xi)$ is obtained from $a(\epsilon)$ by way of a linear transformation $\mathbf{G}_\xi$; that is,

$$a(\xi) = \mathbf{G}_\xi\, a(\epsilon)\ , \tag{11.2}$$

and the set of $\mathbf{G}_\xi$ is a group. The invariant measure of the group is the prior $\mu(\xi)$.

Can one give examples of models $p(x|\xi)$ that are form invariant in this sense? Yes! We discuss two examples, the binomial model in Sect. 11.1 and the Poisson distribution in Sect. 11.2. The latter one is a limiting case of the binomial model. Section A.11 gives the solutions to the problems suggested to the reader.

There is a field of research where results from the binomial model have found widespread attention; this is item response theory (IRT). It is the conceptual basis of the so-called PISA studies. We devote Chap. 12 to IRT.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_11

11.1 The Binomial Model

The binomial distribution was introduced in Chap. 5. It describes simple alternatives. Each of N observations yields an answer out of two possible ones such as *right* or *wrong*, *up* or *down*. In Sect. 11.1.1 the basic binomial model is discussed, where one answer to a simple alternative is observed: how to define the parameter ξ of the model? What is the measure on the scale of ξ? In Sect. 11.1.2 the more general case is treated when N answers are observed to the same alternative. The "same" means that it is conditioned by the same value of ξ. Of course, N is a natural number.

11.1.1 The Basic Binomial Model

The event variable x takes two values $x = 0, 1$. The result $x = 1$ shall be encountered with probability η, the value $x = 0$ with probability $1 - \eta$; see Sect. 5.1. Setting

$$\begin{aligned} \eta &= \sin^2 \xi \, , \\ 1 - \eta &= \cos^2 \xi \, , \end{aligned} \tag{11.3}$$

the binomial model reads

$$q(x|\xi) = \sin^{2x} \xi \, \cos^{2(x-1)} \, , \qquad x = 0, 1 \, . \tag{11.4}$$

The two-dimensional vector of amplitudes is

$$a(\xi) = \begin{pmatrix} \sin \xi \\ \cos \xi \end{pmatrix} ; \tag{11.5}$$

the amplitude a_0 for the event $x = 1$ is $\sin \xi$ and a_1 for $x = 0$ is $\cos \xi$.

The value of ϵ in Eq. (11.2) is $\xi = 0$; that is,

$$a(\epsilon) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} . \tag{11.6}$$

One obtains Eq. (11.2) with the linear transformation

$$\mathbf{G}_\xi = \begin{pmatrix} \cos \xi & \sin \xi \\ - \sin \xi & \cos \xi \end{pmatrix} ; \tag{11.7}$$

and the matrix $\mathbf{G}_\xi$ can be expressed as

$$\mathbf{G}_\xi = \exp(i \xi \mathbf{g}) \, , \tag{11.8}$$

where

$$\mathbf{g} = i \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{11.9}$$

is the generating operator. This one is Hermitian, whence $\mathbf{G}_\xi$ is unitary. Because $i\xi\mathbf{g}$ is real, $\mathbf{G}_\xi$ is real and orthogonal. The interested reader may show that the expressions (11.7) and (11.8) agree. For this the power series of the exponential function as well as those of the trigonometric functions are used.

The expressions (11.7) and (11.8) show that the binomial model with $N = 1$ is form invariant. The reader is asked to state the reasons. The multiplication function of the group of matrices $\mathbf{G}_\xi$ is

$$\Phi(\xi; \xi') = \xi + \xi' \tag{11.10}$$

as the interested reader may show. Hence, the invariant measure on the scale of ξ is uniform.

The geometric measure of Eq. (9.18) is

$$\begin{aligned} \mu_g(\xi) &= \left[\sum_{x=0}^{1} \left(\frac{\partial}{\partial \xi} a_x(\xi) \right)^2 \right]^{1/2} \\ &= \left[\sin^2 \xi + \cos^2 \xi \right]^{1/2} \\ &\equiv 1 \,. \end{aligned} \tag{11.11}$$

11.1.2 The Binomial Model for *N* Observations

Now N observations $\boldsymbol{x} = (x_1, \dots x_N)$ are given by the binomial model (11.4). Every x_k equals either 0 or 1. All values are conditioned by one and the same value of the parameter ξ. The distribution of the set of x_k is the product

$$p(\boldsymbol{x}|\xi) = \prod_{k=1}^{N} q(x_k|\xi) \tag{11.12}$$

of values given by the model q.

Let the event $\boldsymbol{x}$ contain n times the value $x_k = 1$ and $(N - n)$ times the value $x_k = 0$. The order of the x_k is of no importance because they are all conditioned by the same ξ. By renumbering them one reaches that

$$\begin{aligned} x_k &= 1 \quad \text{for } k = 1, \dots, n \,, \\ x_k &= 0 \quad \text{for } k = n + 1, \dots, N \,. \end{aligned} \tag{11.13}$$

In this way one finds

$$\binom{N}{n} - \text{times}$$

the value of $\sin^{2n}\xi$ for the product in Eq. (11.12) as well as

$$\binom{N}{N-n} - \text{times}$$

the value of $\cos^{2(N-n)}\xi$. Given the number N of observations the event can be characterised by n and the distribution p can be written

$$p(n|\xi) = \binom{N}{n} \sin^{2n}\xi \cos^{2(N-n)}\xi, \quad n = 0, 1, \ldots, N\,. \tag{11.14}$$

The sum over all n is

$$\begin{aligned}\sum_{n=0}^{N} p(n|\xi) &= \binom{N}{n} \sin^{2n}\xi \cos^{2(N-n)}\xi \\ &= \left(\sin^2\xi + \cos^2\xi\right)^N \\ &= 1\end{aligned} \tag{11.15}$$

as it should be.

The ML estimator $\xi^{\text{ML}}(n)$ of the event n of Eq. (11.13) is the solution of the equation

$$\frac{\partial}{\partial\xi} \ln p(n|\xi)\bigg|_{\xi=\xi^{\text{ML}}(n)} = 0\,. \tag{11.16}$$

We place the estimator into the interval $0 < \xi^{\text{ML}} < \pi/2$. Then the logarithm

$$\ln p(n|\xi) = \ln\binom{N}{n} + 2n\ln\sin\xi\; + 2(N-n)\ln\cos\xi \tag{11.17}$$

exists and the ML Eq. (11.16) becomes

$$0 = n\,\frac{\cos\xi}{\sin\xi} - (N-n)\ln\frac{\sin\xi}{\cos\xi}\,. \tag{11.18}$$

The solution is

$$\sin^2\xi^{\text{ML}}(n) = \frac{n}{N}\,. \tag{11.19}$$

The interested reader is asked to check this.

The sin function is periodic and we have to define the interval in which the condition ξ of the model (11.4) is considered. The ML estimator (11.19) is well defined for $n = 0, 1, \dots, N$, where ξ takes values between 0 and $\pi/2$. The parameter η of Sect. 5.1 lies in the interval $0 \le \eta < 1$ which seems to restrict ξ to $0 \le \xi < \pi/2$. Especially, negative values of ξ do not seem to be required inasmuch as the sign of ξ cannot be inferred by counting events. Yet we insist on the probability amplitudes (11.5) which can be analytically continued to negative values of ξ (and even to any value of the real parameter ξ). What can be the use of this? This is useful when we do not infer only one interesting parameter but several or many that are located at different places on the scale of ξ. The item response theory described in Chap. 12 provides an example. When ξ quantifies some phenomenon which is circular by its nature then ξ is an angle and it is defined between zero and 2π. Negative values of $\sin\xi$ are not inferred by counting events at one setting of the observing apparatus but by relating the results from several or many settings. To measure an angle, the apparatus will be moved in small steps of an angle. In this way even negative values of ξ are inferred, by analytical continuation. The latter is an expression from the mathematics of functions that can be expanded into power series. In practical measurement it means to trust into a strong continuity of the results obtained at distinct settings of the apparatus. Then even negative values of a parameter ξ become meaningful.

It happens that one infers parameters requiring an infinite continuous scale although they are based on the binomial model with its circular symmetry. This happens, for example, when at every setting of the measuring apparatus one counts answers to simple alternatives, although the apparatus is moved along a distance without an end. Attempts to measure human competence provide an example; see Chap. 12.

11.2 The Poisson Model as a Limit of the Binomial Model for $N \gg n$

11.2.1 The Prior and Posterior of the Poisson Model

Let N events be collected that follow the binomial model and all are conditioned by one and the same parameter ξ which then ranges within $0 < \xi < \pi/2$. With growing N, ever smaller values of ξ^{ML} become accessible. The corresponding events are characterised by $N \gg n$. We restrict ourselves to such events. More precisely, we let n/N become small by increasing N and construct the distribution of n in this limit. This corresponds to the transition from the binomial model to the Poisson model as described in Sect. 5.3. Thus the desired model is

$$p(n|\lambda) = \frac{\lambda^n}{n!} \exp(-\lambda) , \qquad n = 0, 1, 2, \dots ; \ \lambda > 0 . \tag{11.20}$$

Note that $0! = 1$. We use the results of Sect. 5.3 to obtain the measure on the λ scale and then transform to the parameter ξ such that

$$\xi^2 = \lambda \, . \tag{11.21}$$

With respect to ξ the measure will be constant. Put differently, the λ scale is a transformation of the ξ scale: we distinguish the distributions of λ by the index T.

The amplitudes of the model (11.20) are

$$a_n = \frac{\lambda^{n/2}}{\sqrt{n!}} \exp(-\lambda/2) \, . \tag{11.22}$$

According to Eq. (9.18) the geometric measure is given by

$$(\mu_T(\lambda))^2 = \sum_{n=0}^{\infty} \left(\frac{\partial}{\partial\lambda} a_n(\lambda) \right)^2 \, . \tag{11.23}$$

This can be rewritten as

$$\begin{aligned}(\mu_T(\lambda))^2 &= \frac{1}{4} \sum_{n=0}^{\infty} p(n|\lambda)\, (n/\lambda - 1)^2 \\ &= \frac{1}{4} \overline{(n/\lambda - 1)^2} \, .\end{aligned} \tag{11.24}$$

The overline in the second line of this equation denotes the expectation value with respect to n. The interested reader may confirm this result. The expectation values given in Eqs. (5.21) and (5.22) lead to the measure

$$\mu_T(\lambda) = \frac{1}{2} \lambda^{-1/2} \, . \tag{11.25}$$

This yields the posterior

$$\begin{aligned}P_T(\lambda|n) &= \frac{1}{\Gamma(n+1/2)} \lambda^{n-1/2} \exp(-\lambda) \\ &= \chi_f^{\mathrm{sq}}(\lambda|\tau = 1)\end{aligned} \tag{11.26}$$

of (11.21). It is a chi-squared distribution with $f = 2n + 1$ degrees of freedom; compare Eq. (4.34). Figure 11.1 shows two examples of $P_T(\lambda|n)$.

The announced transformation from λ to ξ according to Eq. (11.21) gives

$$\begin{aligned}\mu(\xi) &= \mu_T(\lambda) \left| \frac{\mathrm{d}\lambda}{\mathrm{d}\xi} \right| \\ &\equiv 1 \, .\end{aligned} \tag{11.27}$$

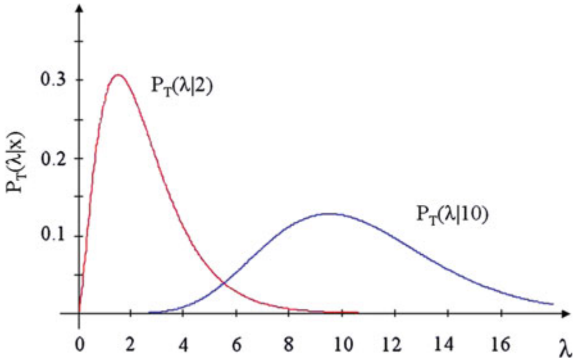


Fig. 11.1 The posterior (11.26) of the Poisson model for $n = 2$ counts and for $n = 10$ counts. The ML estimator is equal to n. The prior distribution is not constant, whence the maximum of the distribution is shifted with respect to λ^{ML}

This agrees with Eq. (11.11) which is satisfying because the Poisson model (11.20) has been obtained as a limiting case of the binomial model (11.4).

Using the parameter ξ, the model (11.20) reads

$$p(n|\xi) = \frac{\xi^{2n}}{n!} \exp(-\xi^2) , \qquad -\infty < \xi < \infty, \ n = 0, 1, 2, \dots . \tag{11.28}$$

Its posterior distribution is

$$P(\xi|n) = \frac{\xi^{2n}}{\Gamma(n + 1/2)} \exp(-\xi^2) . \tag{11.29}$$

This version of the chi-squared distribution with $2n + 1$ degrees of freedom is mirror symmetric with respect to $\xi = 0$ as is the posterior of the basic binomial model (11.4). See Fig. 11.2.

The value of $\xi = 0$ is contained in the domain of definition of ξ; and when the event $n = 0$ is given, one obtains the estimator $\xi^{\text{ML}} = 0$. Thus the estimator recommends $\xi = 0$ which, when given as a condition, entails

$$p(n|0) = \delta_{n,0} , \tag{11.30}$$

that is, $n = 0$ for sure. The interested reader may verify Eq. (11.30). However, the event $n = 0$ does not entail $\xi = 0$ because no Bayesian interval contains the point $\xi = 0$; see Fig. 11.2. This makes sense. When one has observed N times the event $n = 0$ there is still the possibility to find $n = 1$ in a further observation. Even a very large N is not infinite. Every measurement is finite.

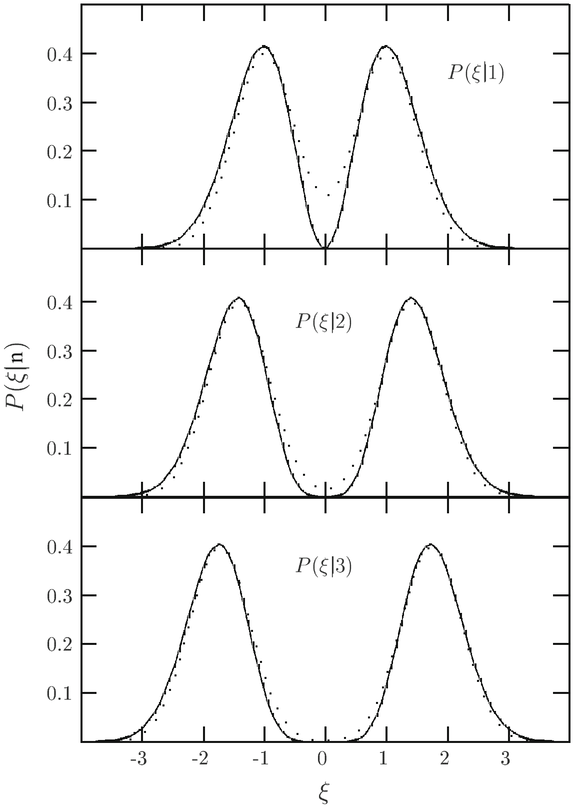


Fig. 11.2 The posterior (11.29) of the Poisson model for $n = 1, 2, 3$ - *full lines* - together with a Gaussian approximation given as *dotted lines*. This graph demonstrates the superior quality of the Gaussian approximation when applied on the ξ scale which has a uniform measure; compare $P(\xi|2)$ to $P_T(\lambda|2)$ in Fig. 11.1

11.2.2 The Poisson Model is Form Invariant

Very much as in Sect. 8.4, we can define a matrix $\mathbf{g}$ which generates a group of linear transformations of the amplitude vector a of the Poisson model. This procedure shows that the Poisson model is form invariant.

We start from the parameterisation (11.28) of the Poisson model. The amplitudes are

$$a_n(\xi) = \frac{\xi^n}{\sqrt{n!}} \exp(-\xi^2/2)\,, \qquad -\infty < \xi < \infty\,,\; n = 0, 1, 2, \ldots\,. \tag{11.31}$$

The derivative of a_n can be written as a linear combination of a_{n-1} and a_{n+1}; one finds

$$\begin{aligned}
\frac{\partial}{\partial \xi} a_n(\xi) &= \frac{1}{\sqrt{n!}}\Big(n\xi^{n-1} - \xi^{n+1}\Big)\exp(-\xi^2/2) \\
&= \left(\sqrt{n}\frac{\xi^{n-1}}{\sqrt{(n-1)!}} - \sqrt{n+1}\frac{\xi^{n+1}}{(n+1!)}\right)\exp(-\xi^2/2) \\
&= \sqrt{n}\, a_{n-1} - \sqrt{n+1}\, a_{n+1} \, . \qquad (11.32)
\end{aligned}$$

This linear combination is obtained by applying the Hermitian matrix

$$\mathbf{g} = i \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & \sqrt{2} & 0 & 0 & 0 \\ 0 & -\sqrt{2} & 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & -\sqrt{3} & 0 & \sqrt{4} & 0 \\ & & \ddots & -\sqrt{4} & \ddots & \ddots \end{pmatrix} \qquad (11.33)$$

to the vector $a(\xi)$. Thus we obtain

$$\frac{\partial}{\partial \xi} a_x(\xi) = i\mathbf{g}\, a(\xi) \, . \qquad (11.34)$$

This operator allows for an infinitesimal shift of the parameter ξ because

$$\begin{aligned}
a(\xi + \mathrm{d}\xi) &= a(\xi) + \frac{\partial}{\partial \xi} a(\xi)\mathrm{d}\xi \\
&= a(\xi) + i\mathbf{g}\, a(\xi)\mathrm{d}\xi \; . \qquad (11.35)
\end{aligned}$$

The operators

$$\mathbf{G}_\xi = \exp\,(i\xi\mathbf{g}) \qquad (11.36)$$

shift the parameter value of a given a to any other one when ξ runs over all real numbers; compare Sect. 8.4. The operator (11.36) is orthogonal, because the generator $\mathbf{g}$ is Hermitian. Therefore the operator (11.36) conserves the norm of the vector to which it is applied. The dimension of the vector $a(\xi)$ as well as the dimension of the operators $\mathbf{g}$ and $\mathbf{G}_\xi$ is infinite. The operators $\mathbf{G}_\xi$ form a group $\mathcal{G}$ and the invariant measure on the scale of ξ is uniform. We conclude that the Poisson model is form invariant and that every amplitude vector $a(\xi)$ is obtained exactly once by applying all the elements $\mathbf{G}_\xi$ of the group $\mathcal{G}$ to the element

$$a(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \, . \qquad (11.37)$$

This vector predicts zero counts with probability 1. One can call it the “vacuum” of events. In Section F it is shown that the generator $\mathbf{g}$ is the difference between a “creation operator” $\mathbf{A}^\dagger$ and the corresponding destruction operator $\mathbf{A}$. With the help of their commutation relation one can show that the nth component of the vector

$$a(\xi) = \mathbf{G}_\xi \, a(0) \tag{11.38}$$

is indeed[1] given by Eq. (11.31).

With the uniform prior one obtains the posterior

$$P(\xi|x) = \frac{\xi^{2x}}{\Gamma(x+1/2)} \exp(-\xi^2)\,, \tag{11.39}$$

of the Poisson model. The normalisation can be verified with the help of Section B.1. The distribution P is symmetric under reflection at $\xi = 0$. Bayesian intervals have the same property. They contain two equivalent ML estimators at $\pm\sqrt{x}$. For every event $n > 0$ the posterior P vanishes at $\xi = 0$. Therefore this value is excluded from any possible Bayesian interval. Hence, when any event has been recorded, the vacuum state (11.37) is excluded.

In the following chapter a widely used application of the binomial model is described. To test the competence of persons one asks questions (called “items” in that context) and registers whether the answers are right or wrong. From sufficiently many of these alternatives a parameter of competence is deduced. Within the PISA studies this serves to rank different nations according to the competence of their students.

[1] The fact that $\mathbf{g}$ is related to the creation and destruction operators of events from the Poisson model (11.28) had been treated quite extensively in the first edition of the present book. Here as well as in Section F, it is mentioned more as a remarkable curiosity. It would require entering into a new realm of mathematics, not really needed to explain form invariance of distributions.

Chapter 12
Item Response Theory

Item response theory[1] IRT tries to construct a statistical model of measurement in psychology. The most frequently used model of IRT carries the name of Georg Rasch[2] [1–6]. He discovered the principle of "specific objectivity". This requires assigning ability parameters $\boldsymbol{\theta}$ to persons and difficulty parameters $\boldsymbol{\sigma}$ to the questions asked in a test such that differences of abilities do not depend on the specific questions, and vice versa such that differences of difficulties do not depend on the persons who establish the difficulty parameters. This requirement attempts to define the notion of "measurement" in general.

In Sect. 12.1 specific objectivity is described more precisely and some terminology of IRT is introduced. In Sect. 12.2 the idea of specific objectivity is used to construct a form-invariant version of IRT, the trigonometric model, in the framework of Bayesian inference. In Sect. 12.3 "typical" data are analysed and discussed. Section 12.4 establishes the statistical errors of the ML estimators. Section A.12 gives the solution to the problem suggested to the reader.

[1] In the first edition of the present book, the place of the present chapter was used to discuss the independence of parameters in the context of form invariance with a non-Abelian symmetry group. Such symmetries make it difficult to integrate the posterior distribution over incidental parameters. Here, we solve the problem with the help of a Gaussian approximation to the posterior; see Chap. 10. Every posterior can be approximated by a Gaussian if the number N of events is large enough. Marginal distributions of a Gaussian are obtained easily; see Sect. B.2. The Gaussian approximation requires, however, that the ML estimators of all parameters exist for every event. This is one of the premises here. The considerations in the former chapter twelve are no longer needed.

[2] Georg Rasch, 1901–1980, Danish mathematician, professor at the University of Copenhagen.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_12

12.1 The Idea of and Some Notions Akin to Item Response Theory

In the framework of IRT the questions that compose a test are simply called "items". A test that allows comparing person parameters as well as item parameters independently of its specific items and specific persons, is called specifically objective by G. Rasch [1, 3, 5, 6] and G.H. Fischer [7–9]. Both authors see specific objectivity realised when (i) the model "compares" every competence parameter to every difficulty parameter and when (ii) one of the parameter classes can be eliminated from the model and the other class is estimated. Both requirements concern the Rasch model. In its framework the comparison between competence and difficulty parameters is achieved by introducing the differences (and only the differences) of the parameters into the model. In the framework of the Rasch model the elimination can be achieved by considering a subset of the data such that the sufficient statistic for the eliminated parameters is constant. The sufficient statistics of the Rasch model are the scores (defined below in Sect. 12.3.1) of both the persons and the items. The sufficient statistic for a person parameter is held constant by considering data with a given person score. Under this condition Andersen [10–12] determines the estimators of the item parameters. His method is called an estimation under "conditional maximum likelihood" (CML). It is described, for example, in the book by Rost [13], p. 126, and the Ph.D. thesis by Fuhrmann [14], Sect. 3.1.

The Rasch model has found a large-scale application of even political relevance in the so-called PISA[3] studies that ranks school education almost worldwide; see, for example, [15–20].

Details of the Rasch model are not described in the present work. A comparison - in terms of philosophical as well as mathematical arguments - between the Rasch and trigonometric models is given in the Ph.D. thesis [14] by Fuhrmann. The formalism of the trigonometric model is described here in Sect. 12.2.

The present approach aims at specific objectivity, too. In the framework of Bayesian inference we consider a model specifically objective if (i) it depends on the differences between person and item parameters and (ii) the measure on the scale of the parameters is uniform. Otherwise the meaning of a difference would change from one place on the scale to another one. A constant measure was requested in References [21, 22], too. The subject of a measure on the parameter scale does not appear in Rasch's work because he did not use Bayesian inference. This holds also for [7–12].

We derive the measure from the property of form invariance. It is defined by its invariance under a group of linear transformations of the probability amplitudes; see Chap. 8 and especially Eq. (8.20). One can say that Rasch's scale maps relations

[3] PISA means a "Programme for International Student Assessment" set up by the OECD, the Organisation for Economic Co-Operation and Development in Paris. Since the year 2000 the competence of high school participants has been measured and compared in most member states of the OECD as well as a number of partner states.

between conditional probabilities whereas the trigonometric scale maps relations between probability amplitudes.[4]

The binomial model

$$q(x_{p,i}|\theta_p,\sigma_i) = \left[R(\theta_p-\sigma_i)\right]^{x_{p,i}} \left[1-R(\theta_p-\sigma_i)\right]^{1-x_{p,i}} . \qquad (12.1)$$

is the basic element of IRT because IRT is based on decisions within binary alternatives. The function q is the probability of obtaining the answer $x_{p,i}$ of the pth person to the ith item. There are two possible answers $x_{p,i}=0,1$ which are interpreted as the alternative of false and true. The interested reader is asked to obtain the expectation values $\overline{x^2}$, $\overline{x(1-x)}$ and $\overline{(1-x)^2}$ from the binomial model.

An item response model is defined by the product

$$p(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\sigma}) = \prod_{p=1}^{N_P}\prod_{i=1}^{N_I} q(x_{p,i}|\theta_p,\sigma_i) \qquad (12.2)$$

of the binomial models for N_P persons and N_I items. The fact that p factorises into the elementary models q expresses the assumption that $x_{p,i}$ is statistically independent of $x_{p',i'}$ for $(p,i)\neq(p',i')$. We use the notation

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{N_P} \end{pmatrix}, \qquad \boldsymbol{\sigma} = \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_{N_I} \end{pmatrix} \qquad (12.3)$$

for the vectors of the person and item parameters as well as

$$\boldsymbol{x} = (x_{p,i}) ; \qquad p=1,\ldots,N_P; \qquad i=1,\ldots,N_I \qquad (12.4)$$

for the matrix of the answers $x_{p,i}$. When R is given, the model (12.2) depends parametrically on the person parameters $\boldsymbol{\theta}$ and the item parameters $\boldsymbol{\sigma}$. This allows us to infer the parameters from the data matrix $\boldsymbol{x}$. The function R is called the item reponse function (IRF).

Inasmuch as the model depends on the differences $\theta_p-\sigma_i$, it remains unchanged if one adds the same constant to each of its parameters. This freedom is removed by a convention; we set

$$\sum_{i=1}^{N_I}\sigma_i = 0 . \qquad (12.5)$$

[4] The author is indebted to Prof. Andreas Müller at the Université de Genève for his proof that probability amplitudes allow a geometric representation of the symmetry of a statistical model. This is described in Sects. 4.13 and 4.14 of Ref. [14]. Here in Chap. 9, we have seen that the amplitudes provide the geometric measure.

12.2 The Trigonometric Model of Item Response Theory

The item response function R in Eq. (12.1) completes the definition of a model of IRT. Rasch [1] used the so-called logistic function

$$R^{(\mathrm{logi})}(\theta_p - \sigma_i) = \frac{\exp(\theta_p - \sigma_i)}{1 + \exp(\theta_p - \sigma_i)}, \tag{12.6}$$

and the trigonometric model is defined by introducing the IRF

$$R(\theta_p - \sigma_i) = \sin^2(\pi/4 + \theta_p - \sigma_i) \tag{12.7}$$

into the model of Eqs. (12.2) and (12.1). As in Sect. 4.3 of [14], a shift of $\pi/4$ in the argument of the $\sin^2$ function places the point $\theta_p - \sigma_i = 0$ at the center of the region of monotonic increase of the IRF. A trigonometric IRF was first suggested in [23, 24].

The function (12.7) turns the binomial model (12.1) into

$$q(x_{p,i}|\theta_p, \sigma_i) = \left[\sin^2(\pi/4 + \theta_p - \sigma_i)\right]^{x_{p,i}} \left[\cos^2(\pi/4 + \theta_p - \sigma_i)\right]^{1-x_{p,i}} \tag{12.8}$$

so that the trigonometric model is

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\sigma}) = \prod_{p=1}^{N_P} \prod_{i=1}^{N_I} \left[\sin^2(\pi/4 + \theta_p - \sigma_i)\right]^{x_{p,i}} \left[\cos^2(\pi/4 + \theta_p - \sigma_i)\right]^{1-x_{p,i}} . \tag{12.9}$$

The vector a of probability amplitudes of the binomial model is now

$$a(\xi) = \begin{pmatrix} \sin(\pi/4 + \theta_p - \sigma_i) \\ \cos(\pi/4 + \theta_p - \sigma_i) \end{pmatrix} . \tag{12.10}$$

In Chap. 11 we have seen that this parameterisation guarantees a uniform measure. As is shown by Eq. (11.11), the geometric measure is identically equal to unity.

The trigonometric model yields the ML estimators as solutions of the system of equations

$$\begin{aligned} 0 &= \sum_{i=1}^{N_I} \left[x_{p,i} \cot\Big(\pi/4 + \theta_p - \sigma_i\Big) - (1 - x_{p,i}) \tan\Big(\pi/4 + \theta_p - \sigma_i\Big)\right] , \\ p &= 1, \ldots, N_P ; \\ 0 &= \sum_{p=1}^{N_P} \left[x_{p,i} \cot\Big(\pi/4 + \theta_p - \sigma_i\Big) - (1 - x_{p,i}) \tan\Big(\pi/4 + (\theta_p - \sigma_i\Big)\right] , \\ i &= 1, \ldots, N_I \end{aligned} \tag{12.11}$$

obtained by requiring the logarithmic derivatives of p to vanish.

On the trigonometric scale the estimators roughly follow the scores of the persons as well as the items. On the whole, but not in the details, we shall find: the smaller the number of correctly solved items (i.e. the person score), the smaller is the competence parameter of a person. The analyses of (artificially generated) data in Sect. 12.3 will show it. They also show that the estimators are all placed on the increasing part of the sin function.

12.3 Analysing Data with the Trigonometric Model

12.3.1 The Guttman Scheme

The most schematic set of data that one can imagine to arise from a competence test is given by a data matrix $\boldsymbol{x}$ which allows us to order persons and items according to the numbers of correct answers obtained. An example for $N = N_P = N_I = 10$ is given in Table 12.1. The persons $p = 1, \ldots, 10$ are ordered according to the decreasing number of correct answers. This number is called the score of a person. The items $i = 1, \ldots, 10$ are ordered with an increasing number of correct solutions. This number is called the score of an item. The data matrix of Table 12.1 is called a Guttman scheme [25]. The possibility of ordering persons and items according to their scores lets a sharp transition from false to correct answers appear at the diagonal of $\boldsymbol{x}$. Below the diagonal one finds false answers, above it the correct answers. Such a data matrix may be a very improbable result from a test. We discuss it because one can analytically write down the ML estimators that it entails. The estimators are linear functions of the respective scores. This is shown by Fuhrmann in Sect. 4.8.2 of his work [14]. There is more than that. Even for tests with a gradual instead of a sharp transition from false to correct, the ML estimators approximately follow the linear functions of the scores obtained from Guttman data. The numerical iteration procedure that solves the ML Eq. (12.11) can be started with the estimators of a Guttman scheme; see Sect. 4.8.3 of [14]. This allows a fast and safe solution of the nonlinear Eq. (12.11).

Given the data of Table 12.1, the ML estimators are

$$\begin{aligned} \theta_p^{\mathrm{ML}} &= \left(t_p - \frac{N}{2}\right)\Delta\,, \qquad \text{for } p = 1, \ldots, N\,, \\ \sigma_i^{\mathrm{ML}} &= \left(\frac{N+1}{2} - s_i\right)\Delta\,, \qquad \text{for } i = 1, \ldots, N\,. \end{aligned} \tag{12.12}$$

Here,

$$\begin{aligned} t_p &= N + 1 - p\,, \\ s_i &= i \end{aligned} \tag{12.13}$$

Table 12.1 Guttman ordered $\boldsymbol{x}$ with dimension $N = 10$. The estimators of the person parameters are given in the last column; the estimators of the item parameters are listed in the last line

	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10	θ_p^{ML}
p=1	1	1	1	1	1	1	1	1	1	1	0.79
p=2	0	1	1	1	1	1	1	1	1	1	0.63
p=3	0	0	1	1	1	1	1	1	1	1	0.47
p=4	0	0	0	1	1	1	1	1	1	1	0.31
p=5	0	0	0	0	1	1	1	1	1	1	0.16
p=6	0	0	0	0	0	1	1	1	1	1	0.00
p=7	0	0	0	0	0	0	1	1	1	1	−0.16
p=8	0	0	0	0	0	0	0	1	1	1	−0.31
p=9	0	0	0	0	0	0	0	0	1	1	−0.47
p=10	0	0	0	0	0	0	0	0	0	1	−0.63
σ_i^{ML}	0.71	0.55	0.39	0.24	0.079	−0.079	−0.24	−0.39	−0.55	−0.71	

are the scores of the persons and items, respectively. The step width

$$\Delta = \frac{\pi}{2N} \tag{12.14}$$

is given by the length $\pi/2$ of the monotonically increasing part of the sin function and the number $N = 10$ of available scores. See Sect. G.1 for the proof that the expressions (12.12) solve the ML Eq. (12.11).

It is possible to modify the Guttman scheme such that $N_P \gg N_I$ whereas the data can still be characterised entirely by the scores.

We note that the Rasch model - whose sufficient statistics are the scores - does not allow analysis of the Guttman scheme, whose data are characterised by the scores. See [24] and Sect. 3.1.7 of [14].

12.3.2 A Monte Carlo Game

We have said that a data matrix $\boldsymbol{x}$ showing a gradual transition from false to correct answers leads to ML estimators that approximately follow linear functions of the scores. The trigonometric model generates such a gradual transition. The deviations of the estimators from those linear functions are now studied. For this a data matrix $\boldsymbol{x}$ has been generated by a Monte Carlo simulation of the trigonometric model.

The simulation comprises $N_P = 500$ persons and $N_I = 20$ items. The "true values" of both parameters, θ_p and σ_i, were chosen between $-\pi/2$ and $+\pi/2$. In this interval the sin function is monotonically increasing. The parameters were given at constant steps of $\theta_{p+1} - \theta_p$ and constant steps of $\sigma_{i+1} - \sigma_i$. These values were fed into a random number generator[5] which produced the $x_{p,i}$ according to the distribution (12.8).

Figure 12.1 visualises one fifth of the simulated data matrix, the part with the most competent persons. There is a black square for a correct answer $x_{p,i} = 1$ and an empty square for a false answer $x_{p,i} = 0$. The persons p label the rows of the matrix. They are ordered according to the values of the estimators θ_p^{ML}. The most competent person is given in the top row. Thus the persons are ordered such that the ML estimators decrease with increasing index p. The items label the columns. The most difficult item is given in the left-hand column, whence the items are ordered such that their ML estimators again decrease with increasing index i. The ML estimators have been found by numerically solving[6] the ML Eq. (12.11).

One observes that the order of the ML estimators basically corresponds to the order of the scores. The person $p = 1$ with the highest parameter solves all items correctly. There is no "hole" within the black of the topmost row of Fig. 12.1. The following rows $p = 2, \dots, 12$ each contain one hole. These persons all have the score $t_p = 19$. They are followed by 17 persons $p = 13, \dots, 29$ where one sees two holes: that is, score $t_p = 18$, and so on.

Looking at the total of Fig. 12.1 we observe the characteristic difference to the Guttman scheme treated in Sect. 12.3.1. There is no sharp boundary between the black area of correct answers and the white area of false answers. The trigonometric model entails a diffuse transition from one area to the other one.

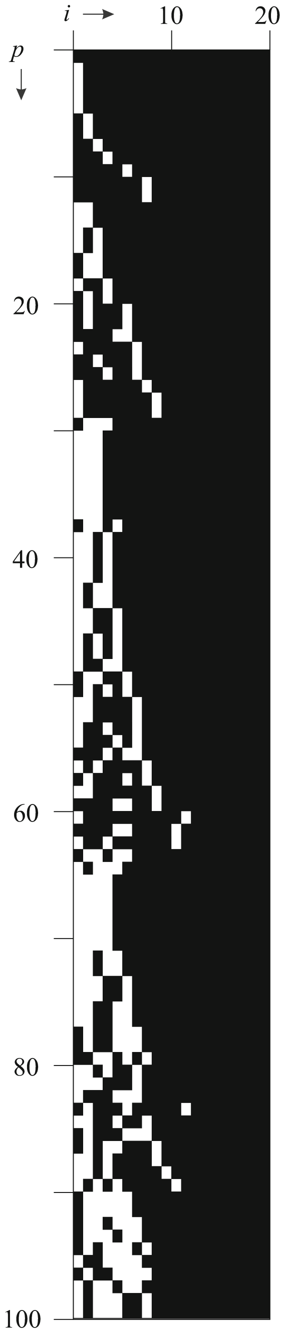
The diffuseness is not entirely structureless; one recognises systematic features within the region of transition from false to correct answers. Look at the persons within a group of the same score, for example, the group of persons $p = 2, \dots, 12$ with $t_p = 19$. A sequence of white holes appears within the black area such that the holes move deeper and deeper into the black as p increases. In other words, the person score remaining constant, the person parameter decreases as the unsolved item becomes easier. This still allows us to interpret the ML estimator as an ability parameter. A similar structure is observed for the two holes that characterise the group $p = 13, \dots, 29$ with $t_p = 18$, and so on. Thus the ability estimator, although it essentially is a monotonic function of the score, is fine-tuned by the difficulty of the items: the more difficult the solved items, the higher is the estimator. The trigonometric model unravels the order beyond the score which is present in the individual patterns of the data matrix.

Sometimes the "fine-tuning" overrules the order of the scores. One finds "intruders", that is, persons lying within a group of the neighboring score. The person $p = 61$ with two holes, that is, score 18, is found within the group of three holes,

[5] The Monte Carlo simulation has been executed in the framework of EXCEL 2003. This system provides the function RAND to generate random numbers. It is described under http://support.microsoft.com/kb/828795/en-us. The author is endebted to Dr. Henrik Bernshausen, University of Siegen (Germany), Fachbereich Didaktik der Physik, for carrying out the Monte Carlo simulation.

[6] The solution of the ML equations has been obtained with the help of the "Euler Math Toolbox". It provides the same routines as "R" or "STATA" to deal with matrices. Details are given in Sections B and C of the Ph.D. thesis [14]. The author is indebted to Dr. Christoph Fuhrmann, Bergische Universität Wuppertal (Germany), Institut für Bildungsforschung, for the numerical calculations.

Fig. 12.1 Data matrix given by a Monte Carlo simulation. A *black* (*empty*) *square* denotes a correct (false) answer. The test comprises $N_P = 500$ persons and $N_I = 20$ items. The most competent 100 persons are displayed (i.e. one fifth of the entire data matrix is shown). Both, person and item parameters, decrease with increasing index



that is, score 17. One of the items unsolved by $p = 61$ lies beyond the center of the black area. This holds also for the intruder $p = 84$ having score $t_{83} = 17$ within the group of score 16. The existence of the intruders shows that the trigonometric model does not only provide information in addition to that given by the scores; it may modify the order of the scores.

We conclude: the ML estimators of the trigonometric model depend on the individual patterns of answers in a way that allows for a meaningful interpretation of the person parameters as measures of ability.

The item parameters cannot be discussed likewise because, in the present example, their number N_I is much smaller than the number N_P of persons. Whence, there is more than one person for every possible person score, and the possible item scores do not all occur even once.

It is obvious from Fig. 12.1 that the scores are not the sufficient statistics of the trigonometric model. Nevertheless one finds the estimators θ_p^{ML} to be approximately given by a linear function of the scores. The difference between this linear function and the actual value of θ_p^{ML} is not merely random; it contains information on the individual strengths or weaknesses of the person p.

12.4 The Statistical Errors of the ML Estimators of Item Response Theory

In Sect. 10.1.1 we have explained that the uncertainty of estimated parameters is derived from a Gaussian approximation to the posterior distribution of these parameters. The statistical model yields the Fisher matrix F. Its inverse is the correlation matrix $C = F^{-1}$ of the Gaussian. The diagonal elements of C are the variances of the parameters. The square root of a variance is the “root mean square value” or Gaussian error of the respective parameter. The present section is devoted to the calculation of the Gaussian errors of both the person and the item parameters of the trigonometric item response theory.

In Sect. G.2 the Fisher matrix is derived under the assumption that the convention (12.5) may be replaced by setting the item parameter σ_{N_I} equal to zero. This is valid because we show that the statistical errors do not depend on the values of the parameters θ_p, σ_i. Thus the errors do not change when one shifts the system of parameters, obtained under the condition $\sigma_{N_I} = 0$, such that the convention (12.5) is met.

In this way we show in Sect. G.2 that the Fisher matrix of the trigonometric model is an $(N_P + N_I - 1)$-dimensional matrix with the structure

$$F = 4 \begin{pmatrix} a\,, & s \\ r\,, & b \end{pmatrix}\,, \tag{12.15}$$

where the submatrices a and b are diagonal and refer to the person and item parameters, respectively. The diagonal matrix

$$a = N_I \, \mathbf{1}_{N_P} \tag{12.16}$$

is proportional to the unit matrix in N_P dimensions. The matrix

$$b = N_P \, \mathbf{1}_{N_I - 1} \tag{12.17}$$

is proportional to the unit matrix in $N_I - 1$ dimensions. The reduction from N_I to $N_I - 1$ dimensions is due to the fact that one of the item parameters is given independently of the data. The matrix s is rectangular; it has N_P rows and $N_I - 1$ columns. Its elements

$$s = \begin{pmatrix} -1, \dots, -1 \\ \vdots \qquad \vdots \\ -1, \dots, -1 \end{pmatrix} \tag{12.18}$$

are all equal to -1. The matrix

$$r = s^{\dagger} \tag{12.19}$$

is the transpose of s.

Note that all elements of F are independent of the parameters $(\boldsymbol{\theta}, \boldsymbol{\sigma})$ of the trigonometric model. Therefore the prior distribution of every parameter is uniform; that is, the measure on its scale is uniform. This is what we had required in the beginning of Sect. 12.1 for a statistical model of IRT. This requirement has led to the trigonometric IRF (12.7).

Let us calculate the determinant of F. It is needed to obtain the inverse of F as well as to check whether the ML estimators of all the parameters of the trigonometric model are well defined. We start by rewriting the matrix in Eq. (12.15) as

$$\begin{pmatrix} a, & s \\ r, & b \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{N_P}, & s \\ 0, & b \end{pmatrix} \begin{pmatrix} \mathbf{1}_{N_P} - sb^{-1}ra^{-1}, & 0 \\ b^{-1}ra^{-1}, & \mathbf{1}_{N_I} \end{pmatrix} \begin{pmatrix} a, & 0 \\ 0, & \mathbf{1}_{N_I} \end{pmatrix} . \tag{12.20}$$

The determinant of a triangular block matrix equals the product of the determinants of the blocks found in the diagonal. This rule yields the determinant of F as

$$\begin{aligned} \det(F) &= 4^{N_P + N_I - 1} \det(a) \det(b) \det\left(\mathbf{1}_{N_P} - sb^{-1}ra^{-1}\right) \\ &= 4^{N_P + N_I - 1} N_I^{N_P} N_P^{N_I - 1} \det\left(\mathbf{1}_{N_P} - N_I^{-1} N_P^{-1} \, sr\right) . \end{aligned} \tag{12.21}$$

By introducing the N_P-dimensional vector

$$|e\rangle = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{12.22}$$

as well as its transpose

$$|e\rangle^\dagger = \langle e| \, , \tag{12.23}$$

one sees that

$$N_I^{-1} N_P^{-1} \, sr = N_I^{-1} N_P^{-1} \, |e\rangle\langle e| \, . \tag{12.24}$$

Here, $|e\rangle\langle e|$ is the dyadic product of the vector (12.22) with itself. The determinant of a combination of the unit matrix with a dyadic product is

$$\det \Big(\mathbf{1} - |e\rangle\langle f| \Big) = 1 - \langle f|e\rangle \, ; \tag{12.25}$$

whence we obtain

$$\begin{aligned} \det(F) &= 4^{N_P+N_I-1} N_I^{N_P} N_P^{N_I-1} \Big(1 - N_I^{-1} N_P^{-1} (N_I - 1) N_P \Big) \\ &= 4^{N_P+N_I-1} N_I^{N_P-1} N_P^{N_I-1} \, . \end{aligned} \tag{12.26}$$

This does not vanish. Hence, all $N_P + N_I - 1$ parameters of the trigonometric model can be estimated and the inverse of F exists.

To write down the Gaussian approximation to the posterior of the model of Eqs. (12.2)–(12.3), we construct the $(N_P + N_I - 1)$-dimensional vector

$$\boldsymbol{\zeta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{N_P} \\ \sigma_1 \\ \vdots \\ \sigma_{N_I-1} \end{pmatrix} \tag{12.27}$$

of all parameters of the model. Then the Gaussian approximation is given by the expression

$$\begin{aligned} P(\boldsymbol{\zeta}|\boldsymbol{x}) \approx {} & (2\pi)^{-(N_P+N_I-1)/2} (\det C)^{-1/2} \\ & \times \exp\left(-(\boldsymbol{\zeta} - \boldsymbol{\zeta}^{\mathrm{ML}})^\dagger (2C)^{-1} (\boldsymbol{\zeta} - \boldsymbol{\zeta}^{\mathrm{ML}}) \right) \, . \end{aligned} \tag{12.28}$$

To obtain the Gaussian errors

$$\Delta\zeta_k = \sqrt{C_{k,k}} \, , \qquad k = 1, \ldots, (N_P + N_I - 1) \tag{12.29}$$

of the estimators, we have to know the diagonal elements $C_{k,k}$ of the correlation matrix $C = F^{-1}$; that is, we have to invert the Fisher matrix.

The diagonal elements of F^{-1} are given by

$$C_{k,k} = \frac{\det(F^{(k,k)})}{\det(F)} . \qquad (12.30)$$

Here, the matrix $F^{(k,k)}$ is obtained by omitting the kth row and the kth line from F; see, for example, Sects. 13.116, 13.117, and 14.13 of [26]. In Sect. G.3 it is shown that

$$\det(F^{(p,p)}) = 4^{N_P+N_I-2}\, N_I^{N_P-2} N_P^{N_I-2}(N_I+N_P-1) , \qquad p = 1, \dots, N_P \quad (12.31)$$

for the elements that refer to the person parameters. Together with Eq. (12.26) one obtains

$$C_{p,p} = \frac{1}{4}\left(\frac{1}{N_P} + \frac{1}{N_I} - \frac{1}{N_P N_I}\right) , \qquad p = 1, \dots, N_P . \qquad (12.32)$$

For the Gaussian approximation to hold, the numbers N_P and N_I of persons and items must be large campared to 1. Then the last result can be approximated by

$$C_{p,p} \approx \frac{1}{4}\left(\frac{1}{N_P} + \frac{1}{N_I}\right) , \qquad p = 1, \dots, N_P . \qquad (12.33)$$

When the number of persons is large compared to the number of items then this reduces to

$$C_{p,p} \approx \frac{1}{4N_I} \qquad \text{for } N_P \gg N_I \qquad (12.34)$$

which means that the Gaussian error of the person parameters depends only on the number of items answered. This is reasonable because in that limit (as we show) the error of the item parameters becomes small as compared to the error of the person parameters; that is, the item parameters can be considered to be known exactly. Equation (12.34) was obtained in Eq. (81) in Sect. 4 of Ref. [14].

The diagonal elements of C referring to the item parameters are

$$\begin{aligned} C_{N_P+i,N_P+i} &= \frac{\det(F^{(N_P+i,N_P+i)})}{\det(F)} \\ &= 4^{N_P+N_I-2}\, \frac{2N_I^{N_P-1} N_P^{N_I-2}}{\det(F)} , \quad i = 1, \dots, N_I - 1 . \end{aligned} \qquad (12.35)$$

The second line of this equation is proven in Sect. G.3. Together with (12.26) one obtains

$$C_{N_P+i,N_P+i} = \frac{1}{2N_P}\,, \qquad i = 1, \ldots, N_I - 1\,. \tag{12.36}$$

The fact that the result C_{N_P+i,N_P+i} for the item parameters is not obtained from the result $C_{p,p}$ for the person parameters by interchanging N_P with N_I, is due to the convention about one of the item parameters. Yet Eq. (12.33) agrees with (12.36) in the case where $N_P = N_I$.

In Sect. 4.4 of his work [14], Fuhrmann shows that the statistical errors of the estimators obtained from the Rasch model are larger or at best equal to the errors obtained from the trigonometric model.

References

1. G. Rasch, On general laws and the meaning of measurement in psychology, in *Fourth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, 1961), pp. 321–333
2. G. Rasch, An item analysis which takes individual differences into account. Br. J. Math. Stat. Psychol. **19**, 49–57 (1966)
3. G. Rasch, An informal report on a theory of objectivity, in *comparisons, in Measurement Theory. Proceedings of the NUFFIC International Summer Session in Science at 'Het Oude Hof'. The Hague, July 14–28, 1966*, ed. by LJTh van der Kamp, C.A.J. Vlek (University of Leiden, Leiden, 1967), pp. 1–19
4. G. Rasch, A mathematical theory of objectivity and its consequences for model contribution, in *European Meeting on Statistics, Econometrics, and Management Science* (Amsterdam, 1968)
5. G. Rasch, On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. Dan. Yearb. Philos. **14**, 58–94 (1977)
6. G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (The University of Chicago Press, Chicago, 1980)
7. G.H. Fischer, *Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells* (Psychologie Verlagsunion, Weinheim, 1988), pp. 87–111
8. G.H. Fischer, Derivations of the rasch model. In Rasch-models: foundations, recent developments and applications, New York, 1995. Workshop held at the University of Vienna (1995), pp. 15–38. 25–27 Feb. 1993
9. G.H. Fischer, *Rasch Models* (Elsevier, Amsterdam, 2007), pp. 515–585. chapter 16
10. E.B. Anderson, Asymptotic properties of conditional maximum likelihood estimators. J. R. Stat. Soc. **32**, 283–301 (1970)
11. E.B. Anderson, Conditional inference for multiple-choice questionaires. Br. J. Math. Stat. Psychol. **26**, 31–44 (1973)
12. E.B. Anderson, Sufficient statistic and latent trait models. Psychometrika **42**, 69–81 (1977)
13. J. Rost, *Lehrbuch Testtheorie—Testkonstruktion*, 2nd edn. (Hans Huber, Bern, 2004)
14. C. Fuhrmann, Eine trigonometrische Parametrisierung von Kompetenzen. Zur Methodologie der probabilistischen Bildungsforschung. Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany. In print at Springer vs. To appear in 2017
15. OECD. PISA 2000. Zusammenfassung zentraler Befunde, http://www.oecd.org/germany/33684930.pdf. Accessed 25 May 2015
16. OECD. Pisa 2003, Technical report, http://www.oecd.org/edu/school/programmeforinternationastudentassessmentpisa/35188570.pdf. Accessed 16 Sept 2013
17. OECD. The PISA 2003 assessment framework—mathematics, reading, science and problem knowledge and skills, http://www.oecd.org/edu/preschoolandschool/

programmeforinternationastudentassessmentpisa/336694881.pdf. Accessed 30 Jan, 2 Feb 2013

18. OECD. Pisa 2003 data analysis manual, spss user 2 ed, http://browse.oecdbookshop.org/oecd/pdfs/free/9809031e.pdf. Accessed on 3 Sept 2008 and 28 Oct 2012
19. OECD. PISA 2012 Ergebnisse im Fokus, http://www.oecd.org/berlin/themen/PISA-2012-Zusammenfassung.pdf. Accessed 9 July 2014
20. OECD. PISA 2009 results: Overcoming social background, equity in learning opportunities and outcomes (volume II), http://dx.doi.org/10.1787/9789264091504-en in http://www.oecd-ilibrary.org/education. Accessed 25 April 2013
21. L.L. Thurstone, *The Measurement of Values*, 3rd edn. (University of Chicago Press, Chicago, 1963), p. 195
22. X. Liu, W. Boone, *Introduction to Rasch Measurement in Science Education* (JAM Press, Maple Grove, 2006), pp. 1–22
23. F. Samejima, *Constant information model on dichotomous response level, in New Horizons in Testing* (Academic Press, New York, 1983), pp. 63–79
24. K. Harney, C. Fuhrmann, H.L. Harney, Der schiefe Turm von PISA. die logistischen Parameter des Rasch-Modells sollten revidiert werden, in *ZA-Information. Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln* **59** (2006), pp. 10–55
25. L. Guttman, *The Basis of Scalogram Analysis* (Gloucester, Smith, 1973)
26. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products* (Academic Press, New York, 2015)

Chapter 13
On the Art of Fitting

13.1 General Remarks

Fitting a set of observations $\boldsymbol{x} = (x_1, \ldots, x_N)$ means that one hopes to have a theoretical understanding of the observations and wants to see whether theory and data fit to each other. The theory is expressed by a family of distributions $p(\boldsymbol{x}|\boldsymbol{\xi})$ which parametrically depends on $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$. The parameters are "optimised" such that the theory comes as close as possible to the observations. This procedure is a subject of the present book anyway: parameters are optimised by determining their ML estimators $\boldsymbol{\xi}^{\mathrm{ML}}$. The existence of the estimators is required.

The observations are supposed to be events from a statistical model. Is this really necessary? The answer is, "Yes": a fit is immediately followed by the question, "Is the fit satisfactory?" This question is answered[1] by looking at the observations, noting their deviations from the theory and deciding whether the sum over the deviations is acceptable. When any deviation between fit and observation is tolerated then the observations must be statistically distributed, and their distribution must be known.

In Sect. 13.2 an example is discussed where the observations follow a Gaussian distribution. The decision about the validity of the fit is taken on the basis of a so-called chi-squared criterion. Section 13.3 discusses the case where the observations are given in terms of natural numbers. In this case, not the distribution of the events but rather its posterior - which is a chi-squared distribution - becomes the basis of the decision. So the decision is again obtained by a chi-squared criterion. If it were possible to approximate every distribution of one real parameter by a chi-squared distribution, then one could base every decision about the validity of a fit on a chi-squared criterion. This possibility is worked out in Sect. 13.4. Section A.13 gives the solutions to the problems suggested to the reader.

[1] The present chapter replaces and summarises Chaps. 13–16 of the first edition of this book.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_13

13.2 The Chi-Squared Criterion

Figure 13.1 illustrates the question: “Do the points and the curve fit to each other?” The graph presents results from experiments designed to verify the principle of detailed balance which says that, except for factors due to the relevant phase spaces, the probability for a certain reaction to occur equals the probability for the inverse reaction. The notion of detailed balance is taken from chemistry. The example of Fig. 13.1 is taken from nuclear physics. The fluctuating curve shows the cross-section of a (p, α) reaction. This curve can be considered as the “theory” because it has been measured to much higher precision than the points which refer to the inverse reaction (α, p). Are the cross-sections of forward and backward reaction compatible with each other?

We label the points by the index $k = 1, \dots, N$. The corresponding ordinate is the event x_k. At the same abscissa the curve shall have the ordinate ξ_k. The error bars in Fig. 13.1 give the root mean square values σ_k of the Gaussian distributions

$$
\begin{aligned}
q(x_k|\xi_k) &= (2\pi\sigma_k^2)^{-1/2} \exp\left(-\frac{(x_k-\xi_k)^2}{2\sigma_k^2}\right) \\
k &= 1, \dots, N\,.
\end{aligned}
\tag{13.1}
$$

of each point. The variances σ_k^2 shall be known.

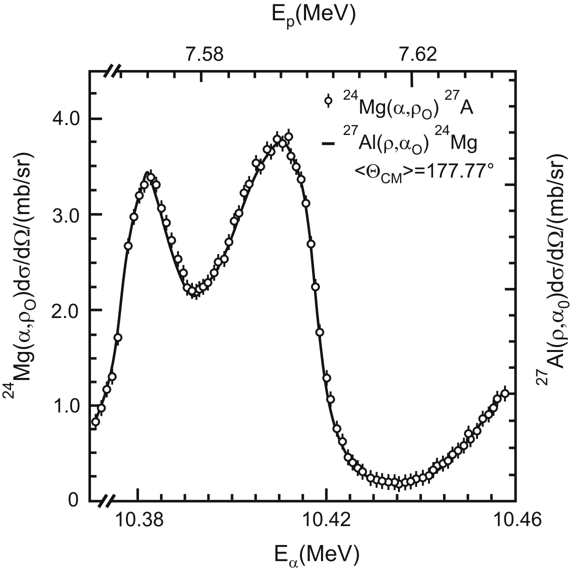


Fig. 13.1 Experimental verification of the principle of detailed balance. A piece of excitation function of the nuclear reaction ^{27}Al$(p, \alpha)^{24}$Mg is given by the curve. The reverse reaction ^{24}Mg$(\alpha, p)^{27}$Al is represented by the points with error bars. The experiment [1] showed that the cross-sections of both reactions agree to a high precision; see also [2, 3]

The sum of the squared differences, divided by σ_k^2,

$$T = \sum_{k=1}^{N} (x_k - \xi_k)^2/\sigma_k^2, \tag{13.2}$$

has the chi-squared distribution

$$\chi_N^{\mathrm{sq}}(T|\tau) = \frac{1}{\Gamma(N/2)}\tau^{-1}\left(\frac{T}{\tau}\right)^{N/2-1}\exp\left(-\frac{T}{\tau}\right) \tag{13.3}$$

with N degrees of freedom; see Sect. 4.1.3. The scaling parameter is

$$\tau = 2\,. \tag{13.4}$$

The decision whether the theory fits the data is obtained by considering the distribution of the quantity T. According to Eq. (4.35) it has the expectation value

$$\overline{T} = \frac{N}{2}\,. \tag{13.5}$$

If T is much larger than this, one shall reject the hypothesis that theory and observations fit to each other. The limit $T_>$, where the rejection occurs is chosen such that the probability $1-K$ to find $T > T_>$ is small. Then the rejection of the hypothesis is justified with the probability K close to unity. The precise value of $T_>$ depends on the choice of K left to the experimenter.

In the framework of Bayesian statistics one can just as well ask whether the value (13.4) of the parameter τ is compatible with the event T. The decision will be the same. This amounts to basing the decision on the posterior of the model (13.3) rather than the distribution of T. The possibility to interchange T with τ becomes most conspicuous when we transform the chi-squared distribution such that its symmetry becomes translational as in Eq. (2.15). One such possibility is given by the transformations

$$\begin{aligned} y &= \ln T\,, \\ \eta &= \ln \tau \end{aligned} \tag{13.6}$$

that lead to Eq. (4.39),

$$\tilde{\chi}_N^{\mathrm{sq}}(y|\eta) = \frac{1}{\Gamma(N/2)}\exp\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right). \tag{13.7}$$

The geometric measure (9.18) on the scale of η is uniform; its value is

$$\mu_g(\eta) = \left(\frac{N}{8}\right)^{1/2}, \tag{13.8}$$

see Sect. H.1. This measure is common to η and y, whence the expression (13.7) is the probability density of the difference between y and the known value

$$\eta = \ln 2\,. \tag{13.9}$$

Unfortunately, the position of the maximum of the likelihood (13.7) depends on the number of degrees of freedom. One has

$$\eta^{\mathrm{ML}} = y - \ln(N/2)\,. \tag{13.10}$$

The interested reader may want to verify this estimate.

A Gaussian approximation to the distribution (13.7) is most meaningful when η is defined such that η^{ML} does not depend on N. This is reached if one shifts the parameter η to

$$\eta' = \eta + \ln(N/2) \tag{13.11}$$

which turns the chi-squared model into

$$\begin{aligned}\tilde{\chi}_N^{\mathrm{sq}}(y|\eta') &= \frac{1}{\Gamma(N/2)}\exp\left(\frac{N}{2}[y-\eta'+\ln(N/2)] - e^{y-\eta'+\ln(N/2)}\right)\\ &= \frac{(N/2)^{N/2}}{\Gamma(N/2)}\exp\left(\frac{N}{2}[y-\eta'-e^{y-\eta'}]\right),\end{aligned} \tag{13.12}$$

and the maximum is now given

$$\eta'^{\mathrm{ML}} = y \tag{13.13}$$

for any degree of freedom.

The parameterisation (13.12) is the best possible one for a Gaussian approximation. The Fisher function (or the inverse variance of the Gaussian) is

$$F = \frac{N}{2}\,. \tag{13.14}$$

Thus the Gaussian approximation to the chi-squared model is

$$\tilde{\chi}_N^{\mathrm{sq}}(y|\eta) \approx \left(\frac{N}{4\pi}\right)^{1/2}\exp\left(-\frac{N}{4}(y-\eta')^2\right). \tag{13.15}$$

Here, η' is given by Eqs. (13.6) and (13.11), and y is given by Eqs. (13.2) and (13.6). The interested reader is asked to verify the approximation (13.15).

The distributions (13.12) or (13.15) can be used to find the Bayesian interval $\mathcal{B}(K)$ (see Chap. 3) in which the difference

$$y - \eta' = \ln(T/N) \tag{13.16}$$

is found with the probability K close to 1.

If $y - \eta'$ is in $\mathcal{B}(K)$, one accepts the hypothesis that the theory fits the data. If $y - \eta'$ is outside $\mathcal{B}(K)$, the hypothesis is rejected. This is what we call the chi-squared criterion. It resembles the popular chi-squared test, often applied to decide whether a fit is reasonable; one finds it in Chap. 11 of [4], in Sect. 9.1.2 of [5], or in Chap. 12 of [6]. However, the chi-squared criterion and the chi-squared test are not the same. For the criterion, we transform via Eqs. (13.6) and (13.11) to a scale on which the measure is uniform. This allows for the Gaussian approximation (13.15) at the lowest possible number N of events.

The Bayesian interval $\mathcal{B}(K)$ of the variable (13.16) is such that the hypothesis is accepted for $y - \eta' = 0$, where the maximum of the likelihood (13.12) occurs. For $y - \eta' > 0$ the value of T is larger than its expected value $N/2$. For $y - \eta' < 0$ the value of T is smaller than $N/2$. For $y - \eta \to -\infty$ the value of T approaches zero, where the observations x_k coincide with the theory ξ_k; see Eq. (13.2). Close to this limit one cannot reject the hypothesis that the theory fits the data. One rather would try to simplify the theory. This can be done on the basis of a philosophical argument frequently used in the natural sciences. “Occam’s razor”[2] helps to avoid unnecessary complication: among the models compatible with the observations $\boldsymbol{x}$, the one with the smallest number of parameters is considered the best.

In the following sections the chi-squared criterion is generalised to cases where the observations x_k do not follow a Gaussian distribution, especially where the x_k are natural numbers.

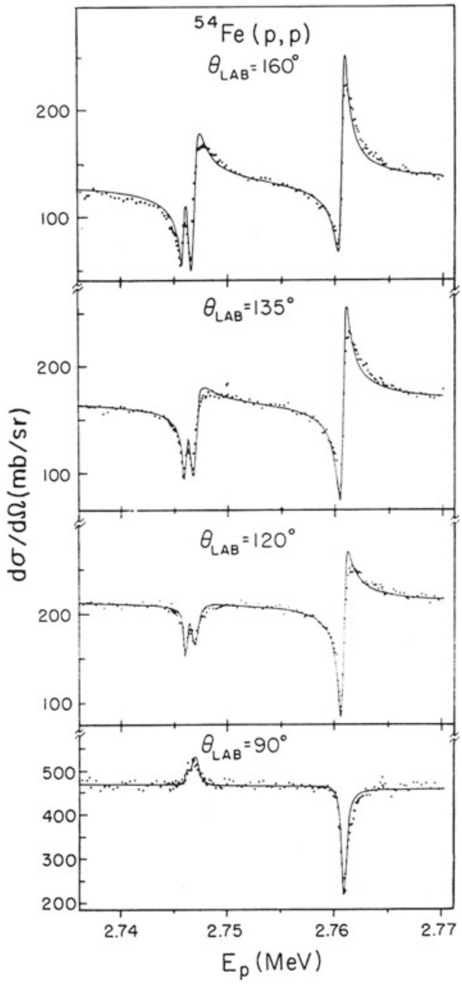
13.3 A Histogram of Counts

Figure 13.2 shows three resonances in elastic proton scattering off ^{54}Fe. The abscissa gives the energy of the protons incident on the target. The ordinate is the scattering cross-section. The points are the observed events. They stand for the counted numbers of recorded protons (before the counts are converted to the cross-section with the unit mb/sr). This example is rooted in quantum mechanics, because the intensity of the radiation comes in quanta. The number n_k recorded in the kth bin of the histogram is an event from the Poisson model

$$q(n_k|\lambda_k) = \frac{\lambda_k^{n_k}}{x_k!}\, e^{-\lambda_k}\,, \quad n_k = 0, 1, 2, \ldots; \lambda_k > 0, \tag{13.17}$$

[2] William Occam (or William of Occam), c.1285–c.1349, English philosopher, theologian, and political writer. He developed a critical epistemology that influenced modern science via Leibniz and Kant.

Fig. 13.2 Resonances in elastic proton scattering on ^{54}Fe. Three resonances appear on each part of the figure. They are viewed under different scattering angles Θ_{LAB} on the four parts. See text. The figure is from [7]



compare Sect. 11.2. The parameter λ_k is the expectation value of n_k; compare Sect. 5.3. Thus the histogram in Fig. 13.2 is distributed according to the model

$$p(\boldsymbol{n}|\boldsymbol{\lambda}) = \prod_{k=1}^{N} q(n_k|\lambda_k)$$

$$= \prod_{k=1}^{N} \frac{\lambda_k^{n_k}}{n_k!} \exp(-\lambda_k) . \tag{13.18}$$

In Fig. 13.2, the deviations between the points and the line are due to this distribution of $\boldsymbol{n}$ provided that the theory, represented by the line, fits the observed points. We want to decide whether this is the case.

For this decision we introduce the prior (11.25) of the model (13.17) and obtain its posterior

$$\begin{aligned} Q(\lambda_k|n_k) &= \frac{1}{\Gamma(n_k+1/2)}\,\lambda_k^{n_k-1/2}e^{-\lambda_k} \\ &= \chi^{\mathrm{sq}}_{f_k}(\lambda_k|\tau=1)\,. \end{aligned} \tag{13.19}$$

It is a chi-squared distribution with

$$f_k = 2n_k + 1 \tag{13.20}$$

degrees of freedom; the notation follows Eq. (11.26). For the desired decision we introduce a scaling parameter s_k and thus generalise the distribution (13.19) to the chi-squared model

$$\begin{aligned} Q(\lambda_k|n_k) &= \chi^{\mathrm{sq}}_{f_k}(\lambda_k|s_k) \\ &= \frac{1}{\Gamma(f_k/2)}s_k^{-1}\left(\frac{\lambda_k}{s_k}\right)^{f_k/2-1}\exp\left(-\frac{\lambda_k}{s_k}\right), \\ 0 &< \lambda_k, s_k < \infty, \end{aligned} \tag{13.21}$$

with f_k degrees of freedom, where f_k is given by Eq. (13.20). The scaling parameter s_k is given by the ordinate of the curve in Fig. 13.2. This curve is the result of fitting three resonances and a slowly varying background to the points in the figure. Each resonance is characterised by three parameters: its position, its width, and its strength. The resonances interfere with the background. For this reason the measured intensity may drop below the level of the background. The mathematical model of scattering theory, used to describe the three resonances plus background, has about a dozen parameters which are determined by the procedure of fitting. The number of observed points, that is, the number of bins in the histogram in Fig. 13.2, is about 100. Therefore it is by no means clear that the set of s_k fits the observations. This is decided by answering the question: “Are the ratios λ_k/s_k as a whole compatible with the numbers n_k?”

The answer is given by the fact that the sum

$$T = \sum_{k=1}^{N} \lambda_k/s_k \tag{13.22}$$

has a chi-squared distribution with

$$f^{\text{tot}} = \sum_{k=1}^{N} (2n_k + 1) \tag{13.23}$$

degrees of freedom. This is the sum over the degrees of freedom f_k assigned to every bin. The proof is given in Sect. H.2. By consequence the quantity T follows the distribution

$$P(T|\boldsymbol{n}) = \chi^{\text{sq}}_{f^{\text{tot}}}(T|\tau)\,. \tag{13.24}$$

The scaling parameter τ in this model is chosen such that the expectation value of T agrees with the expectation value of the sum in Eq. (13.22); that is,

$$\overline{T} = \sum_{k=1}^{N} \overline{\lambda_k}/s_k\,. \tag{13.25}$$

The distribution (13.21) of λ_k yields

$$\overline{\lambda_k} = \frac{f_k}{2} s_k\,. \tag{13.26}$$

Compare Eq. (4.35). By Eqs. (13.20) and (13.25) the expectation value of T is

$$\begin{aligned}\overline{T} &= \frac{1}{2}\sum_{k=1}^{N}(2n_k + 1)\\ &= \frac{f^{\text{tot}}}{2}\,.\end{aligned} \tag{13.27}$$

By a further use of Eq. (4.35) the scaling parameter is

$$\tau = 1\,. \tag{13.28}$$

Thus the distribution of T is

$$\begin{aligned}P(T|\boldsymbol{n}) &= \chi^{\text{sq}}_{f^{\text{tot}}}(T|\tau = 1)\\ &= \frac{1}{\Gamma(f^{tot}/2)}\; T^{f^{\text{tot}}/2-1}\,\exp(-T)\,.\end{aligned} \tag{13.29}$$

The number

$$f^{\text{tot}} = 2\left(\sum_{k=1}^{N} n_k\right) + N \tag{13.30}$$

of degrees of freedom can be seen as a generalisation of Eq. (13.20): the sum over the n_k is the (total) number of counts, and N is unity if only a single bin is considered.

As in Sect. 13.2 we transform T to

$$\begin{aligned} y &= \ln T \\ &= \ln\Big(\sum_{k=1}^{N} \frac{\lambda_k}{s_k}\Big) \end{aligned} \tag{13.31}$$

and τ to

$$\begin{aligned} \eta &= \ln \tau \\ &= 0\,. \end{aligned} \tag{13.32}$$

This leads to a distribution $\tilde{\chi}^{\mathrm{sq}}$ as given by Eq. (13.7); that is, the distribution of y is

$$\tilde{\chi}^{\mathrm{sq}}_{f^{\mathrm{tot}}}(y|\eta = 0) = \frac{1}{\Gamma(f^{\mathrm{tot}}/2)} \exp\Big(\frac{f^{\mathrm{tot}}}{2} y - e^{y}\Big)\,. \tag{13.33}$$

The measure μ_g on the scale of y is uniform; its value is given by Eq. (13.8); that is,

$$\mu_g = \Big(\frac{f^{\mathrm{tot}}}{8}\Big)^{1/2}\,. \tag{13.34}$$

Thus (13.32) is a likelihood function. The position of its maximum depends on f^{tot} in analogy to the result (13.10). Therefore we shift η to

$$\begin{aligned} \eta' &= \eta - \ln \frac{f^{\mathrm{tot}}}{2} \\ &= -\ln \frac{f^{\mathrm{tot}}}{2} \end{aligned} \tag{13.35}$$

in analogy to (13.11). This turns the distribution of y into

$$\chi^{\mathrm{sq}}_{f^{\mathrm{tot}}}(y|\eta') = \frac{(f^{\mathrm{tot}}/2)^{f^{\mathrm{tot}}/2}}{\Gamma(f^{\mathrm{tot}}/2)} \exp\Big(\frac{f^{\mathrm{tot}}}{2}[y - \eta' - e^{y-\eta'}]\Big)\,, \tag{13.36}$$

where Eq. (13.13) applies.

This Fisher function of (13.35) is

$$F = \frac{f^{\mathrm{tot}}}{2} \tag{13.37}$$

which yields the Gaussian approximation

$$\tilde{\chi}^{\mathrm{sq}}_{f^{\mathrm{tot}}}(y|\eta') \approx \Big(\frac{f^{\mathrm{tot}}}{4\pi}\Big)^{1/2} \exp\Big(-\frac{f^{\mathrm{tot}}}{4}(y - \eta')^2\Big) \tag{13.38}$$

in analogy to (13.15). Here, η' is given by Eq. (13.34) and y by Eq. (13.30).

As in Sect. 13.2, the distributions (13.35) or (13.37) can be used to find the Bayesian interval $\mathcal{B}(K)$ in which

$$y - \eta' = \ln \frac{T}{f^{\mathrm{tot}}} \tag{13.39}$$

is found with probability K.

13.4 The Absolute Shannon Information

The last Sect. 13.3 has shown that when one were able to approximate every (one-dimensional) posterior by a chi-squared distribution then one could devise a chi-squared criterion about the validity of every fit. In the present section we hope to present a reasonable way to achieve this approximation. The Shannon information of a given distribution will determine the number of degrees of freedom that the desired approximation must provide.

The Shannon information S has been introduced in Sect. 2.6. However, the expression (2.33) does not provide a well-defined, unique value. It depends on the parameterisation of the distribution w. A transformation of the integration variable x in Eq. (2.33) generally changes the value of S. This is due to the logarithm in the integrand. It does not transform like a function in the sense of Sect. 2.2. In Sect. 13.4.1 we modify the definition of the Shannon information such that S becomes invariant under transformations of its integration variable. In this way, an absolute value of the Shannon information is obtained. By the same token the Shannon information of the (posterior of) a form-invariant model becomes independent of its parameter value.

The absolute Shannon information for chi-squared distributions is given in Sect. 13.4.2 as a function of the number N of degrees of freedom. By numerically calculating the (absolute) Shannon information of any posterior one can assign an “effective” number of degrees of freedom to it and in this way approximate it by a chi-squared distribution.

13.4.1 Definition of the Absolute Shannon Information

Let $p(x|\eta)$ be a form-invariant model with the geometric measure $\mu_g(\eta)$ on the one-dimensional scale of η. We consider the Shannon information of the posterior $P(\eta|x)$. The conventional information S from Eq. (2.33) would change under transformations of η. In other words, S would depend on η. We suggest the definition

$$S = \int \mathrm{d}\eta \, P(\eta|x) \ln \frac{P(\eta|x)}{\mu_g(\eta)} \, . \tag{13.40}$$

The value of S remains unchanged under any transformation of the variable η; see Sect. 6.4. Then it is immaterial whether one calculates the Shannon information from the version (4.34) or the version (4.39) of the chi-squared model with N degrees of freedom. The result is the same provided that compatible measures are used in the two calculations. Remember that the invariant measure (6.33) is determined only up to an arbitrary factor. In the definition (13.40), we use the geometric measure μ_g of Eq. (9.18) which is uniquely defined. Therefore we call (13.40) the absolute Shannon information.

13.4.2 *The Shannon Information Conveyed by a Chi-Squared Distribution*

The absolute Shannon information of the chi-squared distribution (4.39),

$$\tilde{\chi}_N^{\mathrm{sq}}(y|\eta) = \frac{1}{\Gamma(N/2)} \exp\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right), \tag{13.41}$$

is now calculated. The geometric measure is found from the second line of Eq. (9.18),

$$\mu_g^2(\eta) = \frac{1}{4} F\,, \tag{13.42}$$

where F is the Fisher function of the model (13.41). This leads to

$$\begin{aligned}
\mu_g^2(\eta) &= -\frac{1}{4}\,\frac{1}{\Gamma(N/2)} \int \mathrm{d}y \, \exp\left(\frac{N}{2}(y-\eta) - e^{y-\eta}\right) \frac{\partial^2}{\partial\eta^2}\left[\frac{N}{2}(y-\eta) - e^{y-\eta}\right] \\
&= \frac{1}{4}\,\frac{1}{\Gamma(N/2)} \int \mathrm{d}y \, \exp\left(\frac{N}{2}(y-\eta) - e^{y-\eta}\right) e^{y-\eta} \\
&= \frac{1}{4}\,\frac{1}{\Gamma(N/2)} \int \mathrm{d}y \, \exp\left((\frac{N}{2}+1)y - e^{y}\right) \\
&= \frac{1}{4}\,\frac{\Gamma(N/2+1)}{\Gamma(N/2)}\,.
\end{aligned} \tag{13.43}$$

The step from the third to the last line of this equation is done by virtue of the normalisation of the distribution (13.41). The property (B.24) of the Γ function gives

$$\begin{aligned}
\mu_g^2(\eta) &= \frac{1}{4}\,\frac{N}{2} \\
&= \frac{N}{8}\,.
\end{aligned} \tag{13.44}$$

The geometric measure

$$\mu_g(\eta) = \left(\frac{N}{8}\right)^{1/2} \tag{13.45}$$

must be used in (13.40).

Then the Shannon information becomes

$$\begin{aligned} S &= \frac{1}{\Gamma(N/2)} \int \mathrm{d}y \, \exp\left(\frac{N}{2}y - e^y\right) \left[-\ln\Big(\Gamma(N/2)\Big) + \frac{N}{2}y - e^y - \frac{1}{2}\ln(N/8) \right] \\ &= -\ln\Big(\Gamma(N/2)\Big) - \frac{1}{2}\ln(N/8) + \frac{1}{\Gamma(N/2)} \int \mathrm{d}y \, \exp\left(\frac{N}{2}y - e^y\right)\left[\frac{N}{2}y - e^y\right] \\ &= -\ln\Big(\Gamma(N/2)\Big) - \frac{1}{2}\ln(N/8) - \frac{\Gamma(N/2+1)}{\Gamma(N/2)} \\ &\quad + \frac{N/2}{\Gamma(N/2)} \int \mathrm{d}y \, \exp\left(\frac{N}{2}y - e^y\right) y \end{aligned} \tag{13.46}$$

The third line of this equation is obtained by again using the normalisation of the distribution (13.41). The last integral can be obtained from [8] after the substitution

$$y = \ln x \tag{13.47}$$

which gives

$$\begin{aligned} \int \mathrm{d}y \, \exp\left(\frac{N}{2}y - e^y\right) y &= \int_0^\infty \mathrm{d}x \, x^{N/2-1} e^{-x} \ln x \\ &= \frac{\partial}{\partial \nu}\Gamma(\nu)\bigg|_{\nu=N/2} \end{aligned} \tag{13.48}$$

with the help of entry 5 in Sect. 4.358 of Ref. [8]. From Sect. 8.360 of [8] we take the definition

$$\psi(\nu) = \frac{\partial}{\partial \nu} \ln \Gamma(\nu) \tag{13.49}$$

of the ψ function. Thus the Shannon information of a chi-squared distribution with N degrees of freedom takes the form

$$S = -\ln \Gamma(N/2) - \frac{1}{2}\ln(N/8) + \frac{N}{2}\Big(\psi(N/2) - 1\Big) . \tag{13.50}$$

This expression is given as a function of N in Fig. 13.3. It diverges towards $-\infty$ for $N \to 0$, and it asymptotically approaches the value -0.7258 for $N \to \infty$. Its value is negative everywhere for $N > 0$.

This graph helps to solve the task announced in the beginning of the present section. We want to approximate any given distribution of one real variable by a chi-squared distribution. When the Shannon information has been numerically obtained

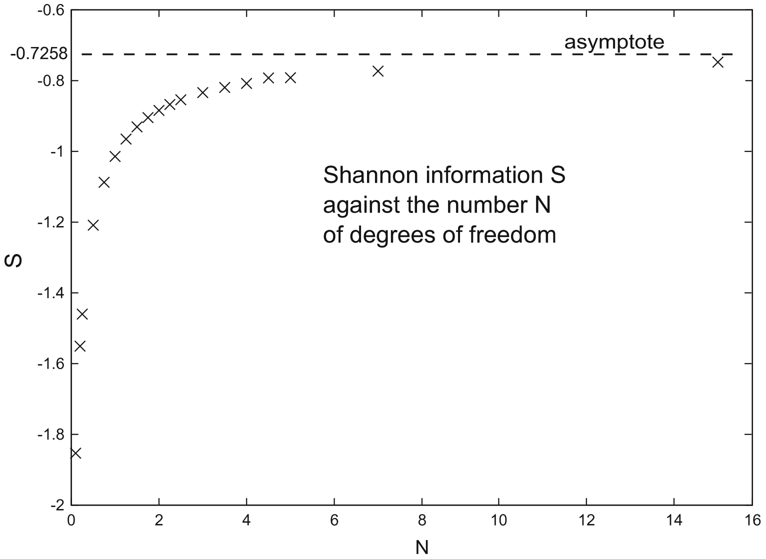


Fig. 13.3 The absolute Shannon information (13.50) for chi-squared distributions as a function of the number N of degrees of freedom. This graph helps to approximate any given distribution (of one real variable) by a suitable chi-squared distribution. For this the Shannon information is to be numerically obtained by Eq. (13.40). The suitable number of degrees of freedom can be found from the present graph. See text

via Eq. (13.40), the effective number of degrees of freedom that characterises the chi-squared distribution can be found from Fig. 13.3. We comment on this in the following section.

13.4.3 Assigning Effective Numbers of Freedom

The Gaussian distribution in Fig. 4.3,

$$P_{\mathrm{Gauss}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \tag{13.51}$$

has the Fisher function

$$F \equiv 1\,. \tag{13.52}$$

According to Eq. (9.18), the geometric measure on the scale of t is

$$\mu_g^{Gauss} \equiv \frac{1}{2}\,. \tag{13.53}$$

By Eq. (13.40) the Shannon information of the Gaussian is

$$\begin{aligned} S_{\text{Gauss}} &= \int \mathrm{d}t\; p_{\text{Gauss}}(t|\sigma=1) \ln \frac{p_{Gauss}(t|\sigma=1)}{\mu_g^{Gauss}} \\ &= -\frac{1}{2}\ln(2\pi) - \frac{1}{2} + \ln 2 \\ &= -0.7258\,. \end{aligned} \tag{13.54}$$

However, there is no chi-squared distribution which would equal a Gaussian. The distribution $\tilde{\chi}_N^{\text{sq}}$ of Eq. (13.12) tends towards a Gaussian for $N \to \infty$. Hence, the value of $S_{\text{Gauss}} = -0.7258$ is reached asymptotically in Fig. 13.3.

Without proof we mention that the Cauchy distribution, compared to the Gaussian in Fig. 4.3, has the effective number $N \approx 0.25$ of degrees of freedom.

The item response theory treated in Chap. 12, is built onto the binomial model

$$q(x|\xi) = \sin^{2x}\xi \cos^{2(x-1)}\xi\,, \qquad x = 0, 1\,, \tag{13.55}$$

introduced in Sect. 5.1; it provides the geometric measure

$$\mu_g^{\text{binom}} \equiv 1 \tag{13.56}$$

as given in Sect. 6.4. The posterior of the model (13.55) is

$$Q(\xi|x) = \frac{2}{\pi}\sin^2(\xi - \xi^{\text{ML}})\,, \qquad 0 \le \xi < \pi\,. \tag{13.57}$$

Within the given range of ξ this is normalised to unity as one confirms via partial integration.

In order to establish a chi-squared criterion on the applicability of Eq. (13.55) or the item response model (12.2), one would like to approximate the posterior Q by a chi-squared distribution. Then one could proceed in analogy to Sect. 13.3.

The absolute Shannon information conveyed by the posterior (13.57) is

$$\begin{aligned} S_{\text{binom}} &= \frac{2}{\pi}\int_0^{\pi} \mathrm{d}\xi\; \sin^2\xi \left[\ln\frac{2}{\pi} + \ln\sin^2\xi\right] \\ &= \ln\frac{2}{\pi} + \frac{8}{\pi}\int_0^{\pi/2} \mathrm{d}\xi\; \sin^2\xi\; \ln\sin\xi\,. \end{aligned} \tag{13.58}$$

According to entry 9 in Sect. 4.384 of [8] this gives

$$\begin{aligned} S_{\text{binom}} &= \ln\frac{2}{\pi} + 1 - \ln 4 \\ &= -0.8379\,. \end{aligned} \tag{13.59}$$

With the help of Fig. 13.3 one obtains the effective number of degrees of freedom

$$N_{\text{binom}} = 2.9 \tag{13.60}$$

for the distribution (13.57). The chi-squared distribution $\tilde{\chi}^{\text{sq}}_{N_{\text{binom}}}$ approximates (13.57).

References

1. E. Blanke, H. Driller, W. Glöckle, H. Genz, A. Richter, G. Schrieder, Improved experimental test of detailed balance and time reversibility in the reactions $^{27}\text{Al} + p \leftrightarrow ^{24}$ Mg. Phys. Rev. Lett. **51**, 355–358 (1983)
2. W. von Witsch, A. Richter, P. von Brentano, Test of time-reversal invariance through the reactions $^{24}\text{mg} + \alpha \leftrightarrow ^{27}$ al $+ p$. Phys. Rev. **169**, 923–932 (1968)
3. H.L. Harney, A. Hüpper, A. Richter, Ericson fluctuations, detailed balance and time-reversal invariance. Nucl. Phys. A **518**, 35–57 (1999)
4. P.R. Bevington, D.K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, 3rd edn. (McGraw-Hill, New York, 2002)
5. V. Blobel, E. Lohrmann, *Statistische und numerische Methoden der Datenanalyse* (Teubner, Stuttgart, 1998)
6. B.P. Roe, *Probability and Statistics in Experimental Physics* (Springer, Heidelberg, 1992)
7. E.G. Bilpuch, A.M. Lane, G.E. Mitchell, J.D. Moses, Fine structure of analogue resonances. Phys. Rep. **28**(2), 145–244 (1976)
8. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products* (Academic Press, New York, 2015)

Chapter 14
Summary

We characterise the starting points of the present book in Sect. 14.1. Its main results are summarised in Sect. 14.2. Several questions that one would like to have answered but that remain open, are collected in Sect. 14.3.

14.1 The Starting Points of the Present Book

During the last two and a half centuries much effort has been devoted to the definition of the Bayesian prior distribution. It is the basis of Bayesian inference although it drops out of the posterior distribution when a sufficiently large amount of events is given. As long as the prior remains undefined, one cannot tell how many events are needed to make the posterior independent of the prior. If the prior remains arbitrary, one can generate any posterior.

In the present book, the prior has been defined by combining symmetry arguments with differential geometry. There are models that connect the event variable with the hypothesis parameter by way of a group of transformations of the event. This well-known symmetry has been called form invariance. We have taken the invariant measure of the symmetry group as the prior distribution. It is possible to represent the above transformations of the event as linear transformations of the vector of probability amplitudes. This is a vector with entries equal to the square root of the probability attributed to the event x. The vector parametrically depends on the hypothesis ξ. This amounts to a parameter representation of a surface in the space of the probability amplitudes. When the parameter has n components, the surface is n-dimensional. The geometric measure on the surface is compatible with the invariant measure of the symmetry group. This shows how to proceed if the symmetry is lacking: the prior is defined as the geometric measure on the surface of probability amplitudes. It is given by a functional known as Jeffreys' rule. It exists if the model is proper; that is, it exists for every reasonable model.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1_14

The posterior distribution, by itself, does not allow us to draw the usual conclusions from data. One wants to give an error interval, that is, a suitable area $\mathcal{B}(K)$ in which the parameter ξ is found with the preselected probability K. This allows one to decide, for example, whether some predicted value ξ^{pre} is compatible with the observed events, that is, whether the distribution $p(x|\xi^{\mathrm{pre}})$ is compatible with x. We have defined $\mathcal{B}(K)$ to be the smallest area in which ξ is found with the probability K and have called it the Bayesian area. The notion of the smallest area requires a measure in the space of ξ. We have identified this measure with the prior distribution. This identification is not necessary, however. It is in the spirit of Occam's razor, because by making this identification one avoids introducing yet another object into the theory of inference.

14.2 Results

Bayesian inference, together with the above definition of the prior distribution, yields the most widely known rules of Gaussian error estimation and error propagation. These rules are formulated without reference to a prior distribution. In the present framework, one can show that the prior distribution for the central value of a Gaussian is uniform and therefore drops out of Bayes' theorem. In this way, the present theory of inference conserves the most widespread method of statistics. However, it generalises Gaussian statistics and allows one to infer the hypothesis parameters of any proper distribution.

The chi-squared test, although it is based on a Gaussian distribution of the events, is not exactly retrieved in the present framework. The test is widely used to assess the quality of a fit. It rejects predictions that come too close to observation. In particular, it rejects a fit that reproduces the observed $\boldsymbol{x}$ point by point. The present method yields a chi-squared criterion that is similar to the conventional test. However, the criterion does not reject "overfitting". The complete fit is not less reasonable than any acceptable model. Within the present formalism, there is no reason to reject it. One can do so by Occam's argument: the simplest interpretation of reality is the best interpretation. This requires one to reduce the number of parameters to the point where the model is barely accepted.

The chi-squared test is conventionally applied to judge the quality of a fit even to a histogram. However, the test requires a Gaussian distribution of the data not a Poissonian one. Hence, when applied to a histogram, it fails when the count rates are low. The present framework allows us to work out a criterion that applies to histograms. This criterion is valid for all count rates; for low rates it is a new tool for making decisions.

There is a phenomenon in statistical inference that may cause confusion when one tries to infer one parameter out of several. Two equally natural routes may lead to different results. One can jointly infer all parameters and subsequently integrate over the uninteresting ones. It is also possible to infer the uninteresting parameters, treating the interesting one as silent. Integration over the uninteresting parameters leads to a model that is conditioned by the interesting parameter only. Inferring the interesting parameter from it leads to a posterior that generally differs from the above marginal distribution. In the literature on Bayesian inference, this phenomenon is called the marginalisation paradox. We circumvent the paradox by the assumption that the number N of events is large enough to allow for a Gaussian approximation to the posterior distribution. Every parameter $\xi_k,\ k = 1, \dots, n$, must be transformed to a scale most suited for this approximation. The most suited scales reduce the number N to the smallest possible value at which the approximation works reasonably.

The Gaussian approximation to the posterior needs the existence of the maximum likelihood estimators ξ_k^{ML}. The ML estimators had not been discussed in the first edition of the present book. Here, we require that they exist and we show that they provide the sufficient statistic for the parameter ξ_k.

The so-called Neyman-Scott problem is an argument against the ML estimators. It is intended to show that ξ_k^{ML} does not always converge against the true value of ξ_k when N is taken to infinity. We show that the Neyman-Scott problem disappears in the framework of Bayesian inference.

Some of our results recall principles of quantum mechanics. Statistical models represent one aspect of reality which can be moved to different places or subjected to similarity transformations. Both groups of transformations can be represented as linear transformations acting on a vector of amplitudes. The dynamics of quantum states is represented in just this way: the evolution as a function of time is given by a linear transformation of the initial vector of amplitudes. Amplitudes are not necessarily positive. They can interfere; that is, their sum can increase or decrease the local intensity. The interference patterns of three resonances displayed in Fig. 13.2 provide an example.

14.3 Open Questions

The analogies with quantum mechanics suggest accepting not only real but also complex probability amplitudes in a theory of inference. This would require considering different symmetry groups and generalising the geometric measure. This has not been done. Therefore the extraction of parameters from the interference phenomena in scattering systems is not adequately treated with the present methods. Figure 13.2 shows resonances interfering with each other and with an essentially constant background. This example is described in quantum mechanics with the help of a complex scattering amplitude. It incorporates a fourfold alternative of possible sources of every event: three resonances and the background. The four sources interfere with each other. One can call this a coherent alternative.

It is possible that one has to introduce the concept of density matrices in order to treat incoherent alternatives appropriately. We have hesitated to do so. It would make the mathematical arguments even more demanding than they are already now.

It would destroy the possibility to define uniquely the prior distribution by Jeffreys' rule. Not withstanding some lengthy mathematical arguments, we have come to rather simple and universal rules of inference within the present treatment.

There are experimental results [1–3] which indicate a possibility of circumventing the density matrix: when, for example, a spectral line is observed in the presence of an incoherent and essentially constant background, then the Fourier transformation of the spectrum may allow the separation of the line from the background. The statistical properties of the Fourier coefficients need further investigation.

Within the present treatment, surprisingly, it is not possible to judge the agreement or disagreement of data with a parameter-free distribution. The curve labelled "GOE" in Fig. 14.1 gives an example which is important for chaotic quantum systems. The distribution is known as Wigner's surmise and is given by the expression

$$w(s) = \frac{\pi s}{2} \exp\left(-\frac{\pi s^2}{4}\right) . \tag{14.1}$$

It is not conditioned by any parameter. It describes the distribution of the energy difference s between a given level and its nearest neighbour. It is certainly a great intellectual achievement to find [6] that this quantity has a universal distribution no matter whether the eigenvalues occur in chemistry or in nuclear physics [4] or in electromagnetic resonators [8]. At the same time, the attempt to verify it experimentally leads into the following dilemma. One decides on the compatibility between (14.1) and data with the help of the arguments that lead to the chi-squared criterion. One uses every datapoint s_k to infer a parameter ξ_k of the distribution of s_k and decides

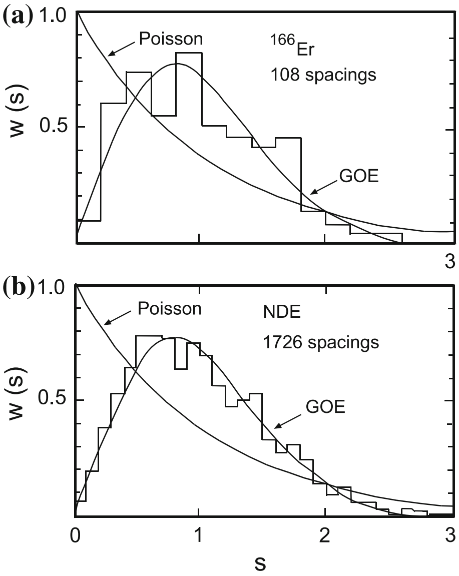


Fig. 14.1 Wigner's surmise: the distribution (14.1) of the nearest-neighbour spacings in a chaotic quantum system [4]. It is labelled "GOE" and is compared to two sets of data taken from **a** neutron scattering on ^{166}Er or **b** the so-called nuclear data ensemble. This figure is found in Ref. [5]

whether the multidimensional distribution of the s_k is compatible with one and the same value $\xi = \xi_k$ for all the k. The lack of any parameter in (14.1) precludes this procedure. One could think of asking whether the multidimensional event formed by the s_k is very improbable from the point of view of the joint distribution of the s_k. However, this requires a measure in the space of s_k. The methods of the present book yield a measure in the space of a parameter, not in the space of the event.

Important work has been done to make plausible that Wigner's surmise does describe the available nuclear data. The authors of [5] have collected a large amount of data, sorted them into a histogram, and compared the histogram to the distribution (14.1). Figure 14.1 seems to show impressive agreement. This figure has become influential because it is one of the justifications of the so-called random matrix theory of chaotic quantum systems [10]. However, the binning of the data is artificial because the events s_k are experimentally determined with a precision much better than the width of the bin into which s_k falls. The result of the comparison will depend on the binning. Arguing strictly, there is no quantified comparison between the parameterless distribution and the data.

Is it possible that we have to consider this dilemma a result rather than a question left open in the present book? If no parameter is specified, one does not know where to look for a deviation between the theoretical and the experimental distributions. If one were to establish that for large s where both the experimental and the predicted distribution have strongly decayed, there is a discrepancy, would that be important enough to reject the surmise? Is the decision impossible if the distribution is parameter-free? For the example at hand, this would mean that one must introduce a model $p(x|\xi)$ that interpolates between the distributions expected for a chaotic system, labelled "GOE" in Fig. 14.1, and a system without chaos, labelled "Poisson". Such an interpolation would have a parameter ξ that measures the deviation from chaos. Without binning the data, one could then decide whether the data are compatible with chaos.

However, we hope to have come close to a solution of the dilemma by showing that the posterior of a histogram of counts is the chi-squared distribution with a number of degrees of freedom uniquely given by the total number of counts in the histogram. This should allow us to devise a chi-squared criterion about the possible agreement between Eq. (14.1) and the data in Fig. 14.1. This criterion would be independent of the binning; that is, one would use the data as they have been obtained and not apply any additional binning. When the level spacings have been measured very precisely, then the "original histogram" may foresee so many bins that most of them show zero counts. Still the posterior of the histogram as a whole would be the above-mentioned chi-squared distribution.

The motto in front of the present book refers to the mystery that probability distributions are ordered by a symmetry although their events do not know of each other. They happen independently.

References

1. B. Dietz, T. Friedrich, H.L. Harney, M. Miski-Oglu, A. Richter, F. Schäfer, H.A. Weidenmüller, Chaotic scattering in the regime of weakly overlapping resonances. Phys. Rev. E **78**, 055204 (2008)
2. B. Dietz, T. Friedrich, H.L. Harney, M. Miski-Oglu, A. Richter, F. Schäfer, J. Verbaarschot, H.A. Weidenmüller, Induced violation of time reversal invariance in the regime of weakly overlapping rersonances. Phys. Rev. Lett. **103**, 064101 (2009)
3. B. Dietz, T. Friedrich, H.L. Harney, M. Miski-Oglu, A. Richter, F. Schäfer, H.A. Weidenmüller, Quantum chaotic scattering in microwave resonators. Phys. Rev. E **81**, 036205 (2010)
4. O. Bohigas, M.-J. Giannoni, Chaotic motion and random matrix theories, in Dehesa et al. [7], pp. 1–99
5. O. Bohigas, R.U. Haq, A. Pandey, Fluctuation properties of nuclear energy levels and widths: comparison of theory with experiment, in Böckhoff [9], pp. 809–813
6. E.P. Wigner, Results and theory of resonance absorption, in C.E. Porter (ed.) *Statistical Theory of Spectra: Fluctuations*, p. 199 (Academic Press, New York, 1957). (The original work is in Oak Ridge National Laboratory Report No. ORNL-2309, 1957)
7. J.S. Dehesa, J.M.G. Gomez, A. Polls (ed.) *Mathematical and Computational Methods in Nuclear Physics: Proceedings of the 6. Granada workshop held in Granada, Spain, Oct. 3–8, 1983*, vol. 209, Lecture Notes in Physics (Springer, Heidelberg, 1984)
8. H.-D. Gräf, H.L. Harney, H. Lengeler, C.H. Lewenkopf, C. Rangacharyulu, A. Richter, P. Schardt, H.A. Weidenmüller, Distribution of eigenmodes in a superconducting stadium billiard with chaotic dynamics. Phys. Rev. Lett. **69**, 1296–99 (1992)
9. K.H. Böckhoff (ed.) *Nuclear Data for Science and Technology. Proceedings of the International Conference Antwerp September 1982* (Dordrecht, Reidel, 1983)
10. Th Guhr, A. Müller-Groeling, H.A. Weidenmüller, Random-matrix theories in quantum physics: common concepts. Phys. Rep. **299**, 189–425 (1998)

Appendix A
Problems and Solutions

In the present appendix, we give the solutions or hints to the solutions of the problems that have been posed within the main text of the book.

A.1 Knowledge and Logic

A.1.1 The Joint Distribution

The joint distribution of $x_1, \dots, x_N$ shall be derived when x_k follows the distribution $p(x_k|\xi)$ for all k.

By (1.1), the distribution of $x_1 \wedge x_2$ is

$$p(x_1 \wedge x_2|\xi) = p(x_1|x_2 \wedge \xi)p(x_2|\xi) \ . \tag{A.1}$$

By assumption, the distribution of x_1 is not conditioned by x_2. Therefore one obtains

$$p(x_1 \wedge x_2|\xi) = p(x_1|\xi)p(x_2|\xi) \ . \tag{A.2}$$

We simplify the notation and replace the logical symbol $\wedge$ by a comma. A proof by induction yields

$$p(x_1, \dots, x_N|\xi) = \prod_{k=1}^{N} p(x_k|\xi) \, . \tag{A.3}$$

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

A.2 Bayes' Theorem

A.2.1 Bayes' Theorem Under Reparameterisations

We convince ourselves that Bayes' theorem behaves properly under reparameterisations.

Let the parameter ξ be expressed by η via (2.7). The value of the integral (2.5) remains unchanged. The posterior distribution (2.6) transforms as the prior; see (2.9), that is, as a density. This is satisfactory.

Let x be continuous. The reparameterisation

$$x' = Tx \tag{A.4}$$

transforms $p(x|\xi)$ in both the numerator and the denominator of (2.6), according to (2.9). The Jacobians drop out. Therefore, in this case $P(\xi|x)$ transforms as a function, not as a density. Again, this is satisfactory.

A.2.2 Transformation to the Uniform Prior

Show that one obtains the uniform prior $\mu_T(\eta) \equiv \text{const}$ if η is an indefinite integral of $\mu(\xi)$.

According to (2.9), one has

$$\mu_T(\eta) = \mu(\xi) \left| \frac{d\eta}{d\xi} \right|^{-1} . \tag{A.5}$$

The claim follows from

$$\left| \frac{d\eta}{d\xi} \right| = \mu(\xi) . \tag{A.6}$$

A.2.3 The Iteration of Bayes' Theorem

Interpret (2.20) as an iteration of Bayes' theorem.

Let $P_{N-1}(\xi|x_1 \dots x_{N-1})$ be given by (2.20). One then has to use $N-1$ events instead of N in this expression. We construct $P(\xi|x_1 \dots x_N)$ by introducing P_{N-1} at the place of μ into Bayes' theorem (2.6). This leads to

$$P(\xi|x_1 \dots x_N) \propto \mu(\xi) p(x_N|\xi) \prod_{k=1}^{N-1} p(x_k|\xi) \tag{A.7}$$

and agrees with (2.20) after normalisation.

A.2.4 The Gaussian Model for Many Events

Convince yourself that (2.23) is the joint distribution of N Gaussian events.

From (2.19) follows

$$p(x_1 \dots x_N|\xi) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\xi)^2\right). \tag{A.8}$$

One rearranges the sum

$$\begin{aligned}\sum_{k=1}^{N}(x_k-\xi)^2 &= \sum_k (x_k^2 - 2\xi x_k + \xi^2)\\ &= N(\xi^2 - 2\xi\langle x\rangle + \langle x^2\rangle)\\ &= N(\xi - \langle x\rangle)^2 + N(\langle x^2\rangle - \langle x\rangle^2)\end{aligned} \tag{A.9}$$

and obtains (2.23).

A.2.5 The Distribution of Leading Digits

Show that

$$q(a) = \log\frac{a+1}{a} \tag{A.10}$$

is the probability for the number ξ to have the leading digit a when ξ is distributed according to (2.31).

One has

$$\begin{aligned}q(a) &\propto \int_a^{a+1} d\xi\, \mu(\xi)\\ &\propto \ln\frac{a+1}{a}\,.\end{aligned} \tag{A.11}$$

Normalising this distribution, one obtains (A.10).

A.3 Probable and Improbable Data

A.3.1 The Size of an Area in Parameter Space

Show that the size

$$V = \int_{\mathcal{I}} d\xi \; \mu(\xi) \tag{A.12}$$

of an area $\mathcal{I}$ does not change under the reparameterisation

$$\eta = T\xi \; . \tag{A.13}$$

Substituting η for ξ in (A.12) yields

$$\begin{aligned} V &= \int_{T\mathcal{I}} d\eta \; \mu(\xi) \frac{\partial \xi}{\partial \eta} \\ &= \int_{T\mathcal{I}} d\eta \; \mu_T(\eta) \; . \end{aligned} \tag{A.14}$$

Here, the transformation (2.9) of a density has been used. We have written $T\mathcal{I}$ for the image of $\mathcal{I}$ in the space of η. Equation (A.14) shows that the volume does not change when it is mapped onto another parameter space. This statement is equivalent to the differential law (2.9) of transformation.

Let the proper distribution $Q(\xi)$ be given, and let $\mathcal{B}(K)$ be a Bayesian area. It remains to show that the transformation $T\mathcal{B}$ of $\mathcal{B}$ is the Bayesian area of the transformed distribution $Q_T(\eta)$. Because every area in the space of ξ has an image in the space of η and because their volumes do not change under the mapping, the volumes can be compared in η as well as in ξ. Therefore the smallest area of a certain set of areas defined in ξ, is mapped onto the smallest area $T\mathcal{I}$ of the corresponding set in η. Hence, the Bayesian interval is independent of the parameterisation.

A.3.2 No Decision Without Risk

The statement that every decision is a risk is now discussed.

Let $Q(\xi)$ be a normalised distribution and $\mathcal{B}(K)$ a Bayesian area and let the value ξ^{pre} of the parameter ξ be predicted. If ξ^{pre} is not in $\mathcal{B}$, one rejects the prediction. However, with probability $1 - K$, the distribution Q allows ξ^{pre} to be outside $\mathcal{B}(K)$. Therefore the decision is wrong with probability $1 - K$.

If one chooses $K = 1$, then $\mathcal{B}$ covers the space of ξ. Every value is then accepted, and there is nothing to decide. One does not obtain any information from the fact that ξ^{pre} is in $\mathcal{B}$ because one knows it anyway.

A.3.3 Normalisation of a Gaussian Distribution

Show that the posterior distribution (3.5) is normalised to unity.

The substitution

$$y = x/\sigma \tag{A.15}$$

yields

$$\begin{aligned}\int_0^\infty d\sigma\ x\sigma^{-2}\exp\left(-\frac{x^2}{2\sigma^2}\right) &= \int_0^\infty dy\ \exp\left(-\frac{y^2}{2}\right)\\ &= \left(\frac{\pi}{2}\right)^{1/2}\end{aligned} \tag{A.16}$$

which proves the statement.

The last version of (A.16) is a consequence of

$$\int_{-\infty}^{\infty} dx\ \ \exp(-x^2) = \sqrt{\pi}\ . \tag{A.17}$$

This equation is implied by the normalisation of the Gaussian (2.16). In order to prove it, one considers the square of the integral (A.17); that is,

$$\begin{aligned}\int_{-\infty}^{\infty} dx\exp(-x^2)\int_{-\infty}^{\infty} dy\exp(-y^2) &= \int_{-\infty}^{\infty} dx\int_{-\infty}^{\infty} dy\ \ \exp\left(-(x^2+y^2)\right)\\ &= \int_0^{2\pi} d\phi\int_0^\infty dr\ \ r\exp(-r^2)\\ &= \pi\int_0^\infty dz\ \ \exp(-z)\\ &= \pi\ .\end{aligned} \tag{A.18}$$

One comes from the first to the second version of this equation by introducing polar coordinates.

A.3.4 The Measure of a Scale-Invariant Model

Find the measure $\mu(\xi)$ of the model

$$p(x|\xi) = \xi^{-1}\ w\left(\frac{x}{\xi}\right)\ , \tag{A.19}$$

where x, ξ are real and positive, and $w(y)$ is any normalised density.

One transforms x via

$$y = \ln x \tag{A.20}$$

and the law (2.9). At the same time, one reparameterises

$$\eta = \ln \xi \tag{A.21}$$

keeping in mind that p transforms as a function with respect to ξ (cf. Chap. 2). This gives

$$\begin{aligned} p_T(y|\eta) &= \frac{x}{\xi}\; w\left(\frac{x}{\xi}\right) \\ &= \exp(\eta - y)\; w\left(\exp(\eta - y)\right)\;, \end{aligned} \tag{A.22}$$

which is a function of the difference $\eta - y$. According to the argument of Sect. 2.3, the measure

$$\mu_T(\eta) \equiv \text{const} \tag{A.23}$$

is uniform with respect to η. With the help of (2.9), one finds

$$\mu(\xi) \propto \xi^{-1}\;. \tag{A.24}$$

A.3.5 A Single Decay Event

The model of radioactive decay is

$$p(t|\tau) = \tau^{-1} \exp\left(-t/\tau\right)\;; \tag{A.25}$$

see Sect. 4.2. Here, t is the time of observation of the event after the experiment has been started at $t = 0$. The parameter τ is the mean lifetime. According to (3.4), the prior is

$$\mu(\tau) \propto \tau^{-1}\,. \tag{A.26}$$

The posterior is

$$P(\tau|t) = \mathcal{N}^{-1}\tau^{-2}\exp(-t/\tau) \tag{A.27}$$

with the normalising factor

$$\begin{aligned} \mathcal{N} &= \int_0^\infty d\tau\,\tau^{-2}\exp(-t/\tau) \\ &= \int_0^\infty d\lambda\,\exp(-t\lambda) \\ &= t^{-1}\,. \end{aligned} \tag{A.28}$$

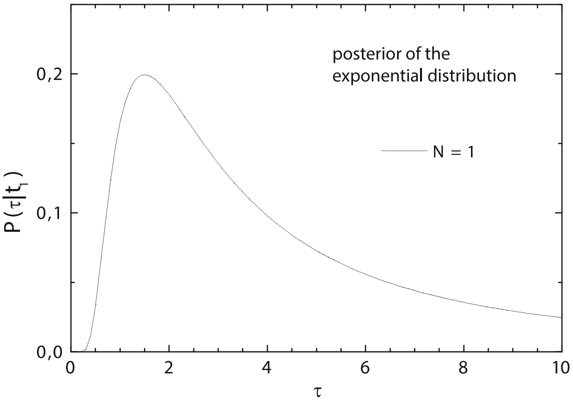


Fig. A.1 Inferring the mean lifetime of a radioactive substance from a single observation. This figure complements the Figs. 2.4 and 2.5 in Chap. 2

Hence, the posterior exists and is

$$P(\tau|t) = t^{-1}\tau^{-2}\exp(-t/\tau)\,. \tag{A.29}$$

It is given in Fig. A.1. One can indeed infer τ from a single observation.

A.3.6 Normalisation of a Posterior of the Gaussian Model

Show that the distribution (3.12) is normalised to unity.

The normalising factor $\mathcal{N}$ of the distribution

$$P(\sigma \mid x) = \mathcal{N}^{-1}\sigma^{-N-1}\exp\left(-\frac{N}{2}\frac{\langle x^2\rangle}{\sigma^2}\right) \tag{A.30}$$

is determined. It is given by

$$\mathcal{N} = \int_0^\infty d\sigma\,\sigma^{-N-1}\exp\left(-\frac{b}{\sigma^2}\right)\,, \tag{A.31}$$

where

$$b = \frac{N}{2}\langle x^2\rangle\,. \tag{A.32}$$

Substituting σ by $\tau = \sigma^{-2}$ one obtains

$$\mathcal{N} = 2^{-1}\int_0^\infty d\tau\,\tau^{N/2-1}\exp(-b\tau)\,. \tag{A.33}$$

By help of the substitution $\tau' = b\tau$, this takes the form

$$\begin{aligned}\mathcal{N} &= 2^{-1} b^{-N/2} \int_0^\infty \mathrm{d}\tau' \, \tau'^{N/2-1} \exp(-\tau') \\ &= 2^{-1} b^{-N/2} \Gamma(N/2)\,. \end{aligned} \tag{A.34}$$

The last line of this equation is due to Eq. (B.23) in Sect. B.4.

A.3.7 The ML Estimator from a Gaussian Likelihood Function

The ML estimator is obtained from the likelihood function $L(\sigma \mid x)$ by requiring

$$\frac{\partial}{\partial \sigma} L(\sigma \mid x) = 0\,. \tag{A.35}$$

With the help of Eq. (3.13) together with (3.4) and (3.12) this leads to the requirement

$$\frac{\partial}{\partial \sigma} \left(-N \ln \sigma - \frac{N \langle x^2 \rangle}{2\sigma^2} \right) = 0\,. \tag{A.36}$$

Its solution is given by

$$(\sigma^{\mathrm{ML}})^2 = \langle x^2 \rangle\,. \tag{A.37}$$

A.3.8 The ML Estimator from a Chi-Squared Model

Show that Eq. (3.18) is the ML estimator of the likelihood function given by the posterior (3.17).

The likelihood function is

$$L(\eta | y) \propto \exp\left(\frac{N}{2} [y - \eta] - e^{y-\eta} \right)\,, \tag{A.38}$$

because the prior distribution of the posterior (3.17) is uniform. This yields

$$\frac{\partial}{\partial \eta} L = -N/2 + e^{y-\eta}\,. \tag{A.39}$$

The ML estimator solves the equation

$$\frac{\partial}{\partial \eta} L = 0\,. \tag{A.40}$$

This gives

$$\exp(\eta^{\mathrm{ML}}) = \frac{2}{N}\, e^{y} \tag{A.41}$$

or

$$\begin{aligned}\eta^{\mathrm{ML}} &= \ln\left(\frac{2}{N}\, e^{y}\right)\\ &= \ln\langle x^2\rangle\,.\end{aligned} \tag{A.42}$$

The second version of this equation is obtained by help of Eq. (3.16).

A.3.9 Contour Lines

Show that contour lines are invariant under reparameterisations.

Both $Q(\xi)$ and $\mu(\xi)$ transform under the reparameterisation (2.8) according to the law (2.9). The Jacobian $|\partial\xi/\partial\eta|$ drops out of the definition (3.23) which therefore holds for the transformed distributions $Q_T(\eta)$, $\mu_T(\eta)$ in the same way.

A.3.10 The Point of Maximum Likelihood

Show that the point $\xi = \xi^{\mathrm{ML}}$, where the likelihood function L of Eq. (3.23) assumes a maximum, does not change under reparameterisations.

Both densities Q and μ are transformed from ξ to

$$\eta = T\xi \tag{A.43}$$

via (2.9). The Jacobian $|\partial\xi/\partial\eta|$ drops out of L. Hence, L transforms as a function; that is,

$$L_T(\eta) = L(\xi)\ . \tag{A.44}$$

The value ξ^{ML} is a solution of

$$\frac{d}{d\xi}L(\xi) = 0\ . \tag{A.45}$$

We consider a neighborhood in the space of ξ where there is only one solution. By virtue of (A.44), the last equation can be written

$$\frac{dL_T(\eta)}{d\eta}\left|\frac{d\eta}{d\xi}\right| = 0\,. \tag{A.46}$$

Because $\eta(\xi)$ is a transformation, the derivative $d\eta/d\xi$ does not vanish. Therefore the point in η, where $dL_T/d\eta$ vanishes, is $T\xi^{\mathrm{ML}}$. This is meant by saying that the place of the maximum does not change under the transformation.

A maximum of the distribution $Q(\xi)$, however, is no longer found at its original place after the transformation to $Q_T(\eta)$.

A.4 Description of Distributions I: Real *x*

A.4.1 The Mean of a Gaussian Distribution

Convince yourself that the mean value $\overline{x}$ of the distribution (4.1) is equal to ξ.

All the odd moments of a Gaussian vanish; that is, one has

$$\int_{-\infty}^{\infty} dx\, x^{2n+1} \exp\left(-\frac{x^2}{2\sigma^2}\right) = 0\,. \tag{A.47}$$

One sees this by the transformation $x \longrightarrow -x$. The substitution $x' = x + \xi$ yields

$$0 = \int dx\, (x - \xi) \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right)\,, \tag{A.48}$$

and this is equivalent to

$$0 = \overline{x} - \xi\,. \tag{A.49}$$

A.4.2 On the Variance

Prove the equivalence of the expressions (4.3), (4.4) for the variance.

The expectation value $\overline{f(x)}$ of a function f is linear with respect to f. Therefore one obtains

$$\begin{aligned}\overline{(y-\overline{y})^2} &= \overline{y^2} - 2\overline{y\overline{y}} + \overline{y}^2 \\ &= \overline{y^2} - \overline{y}^2\,.\end{aligned} \tag{A.50}$$

A.4.3 Moments of a Gaussian

Prove that (4.5) holds for the Gaussian model (4.1).

The function $Z(a)$ of (4.2) has the property

$$(-)^n \frac{d^n}{da^n} Z = \int dx \, (x - \xi)^{2n} \exp\left(-a(x-\xi)^2\right) \, . \tag{A.51}$$

Therefore the variance is

$$\begin{aligned} \mathrm{var}(x) &= - \, (2\pi\sigma^2)^{-1/2} \frac{d}{da} Z(a) \bigg|_{a=(2\sigma^2)^{-1}} \\ &= (2\pi\sigma^2)^{-1/2} \pi^{1/2} 2^{-1} (2\sigma^2)^{3/2} \\ &= \sigma^2 \, . \end{aligned} \tag{A.52}$$

Similarly, the fourth moment is

$$\begin{aligned} \overline{(x-\xi)^4} &= \, (2\pi\sigma^2)^{-1/2} \frac{d^2}{da^2} Z(a) \bigg|_{a=(2\sigma^2)^{-1}} \\ &= (2\pi\sigma^2)^{-1/2} \pi^{1/2} (3/4) (2\sigma^2)^{5/2} \\ &= 3\sigma^4 \, . \end{aligned} \tag{A.53}$$

This proves (4.5).

A.4.4 The Normalisation of a Multidimensional Gaussian

Prove the normalisation (4.17) of the multidimensional Gaussian.

The result (4.17) is easy to understand if C is diagonal. Then (4.16) factorises into n one-dimensional Gaussian distributions. It C is not diagonal, there is an orthogonal transformation O that diagonalises it; that is,

$$OCO^T = C_{\mathrm{diag}} \, . \tag{A.54}$$

The same transformation diagonalises C^{-1}. The normalising integral then takes the form

$$Z = \int dx \; \exp\left(- \left[O(x-\xi)\right]^\dagger (2C_{\mathrm{diag}})^{-1} \left[O(x-\xi)\right]\right) \, . \tag{A.55}$$

One changes the integration variables to

$$x' = Ox \, . \tag{A.56}$$

The Jacobian of this transformation is $|\det O| = 1$. This procedure yields

$$\begin{aligned} Z(C) &= \left((2\pi)^n \det C_{\text{diag}}\right)^{1/2} \\ &= \left((2\pi)^n \det C\right)^{1/2} . \end{aligned} \tag{A.57}$$

A.4.5 The Moments of the Chi-Squared Distribution

Prove the results (4.35, 4.36) for the moments of the chi-squared distribution.

Because the distribution (4.34) is normalised to unity, the equation

$$\int \mathrm{d}T\; T^{N/2-1} \exp(\tau' T) = \Gamma(N/2)\tau'^{-N/2} \tag{A.58}$$

holds. The parameter τ in (4.34) has been replaced by $\tau' = \tau^{-1}$. We differentiate both sides of (A.58) with respect to τ' and obtain

$$\int \mathrm{d}T\; T\; T^{N/2-1} \exp(\tau' T) = \Gamma(N/2)(N/2)\tau'^{-N/2-1} . \tag{A.59}$$

This yields

$$\begin{aligned} \overline{T} &= \int \mathrm{d}T\; T\; \chi^{\text{sq}}(T|\tau') \\ &= \frac{N}{2}\tau'^{-1} . \end{aligned} \tag{A.60}$$

The second moment of the chi-squared distribution is obtained from the second derivative of Eq. (A.58).

A.4.6 Moments of the Exponential Distribution

Calculate the mean value and the variance implied by the exponential distribution.

The exponential model (4.40) leads to the mean value

$$\begin{aligned} \overline{x} &= \xi^{-1} \int_0^\infty dx\, x \exp(-x/\xi) \\ &= \xi^{-1}\xi^2\, \Gamma(2) \\ &= \xi \end{aligned} \tag{A.61}$$

with the help of the properties of the Γ function given in Appendix B.4. Analogously, one finds the second moment

$$\begin{aligned}\overline{x^2} &= \xi^{-1}\int_0^\infty dx\, x^2 \exp(-x/\xi)\\ &= \xi^2\,\Gamma(3)\\ &= 2\,\xi^2\,. \end{aligned}\tag{A.62}$$

This yields the variance

$$\operatorname{var}(x) = \xi^2 \tag{A.63}$$

given in (4.42).

A.5 Description of Distributions II: Natural *x*

A.5.1 The Second Moments of the Multinomial Distribution

Prove that the second moments of the multinomial distribution are given by (5.16).

In the case of $l \neq l'$, one can use (5.4) to write

$$\begin{aligned}\overline{x_l x_{l'}} &= \frac{\partial^2}{\partial\zeta\partial\zeta'}\sum_x \zeta^{x_l}\zeta'^{x_{l'}}\,p(x|\eta)\bigg|_{\zeta,\zeta'=1}\\ &= \frac{\partial^2}{\partial\zeta\partial\zeta'}N!\sum_x \zeta^{x_l}\zeta'^{x_{l'}}\prod_{k=1}^{M}\frac{\eta_k^{x_k}}{x_k!}\bigg|_{\zeta,\zeta'=1}\\ &= \frac{\partial^2}{\partial\zeta\partial\zeta'}\left(\zeta\eta_l + \zeta'\eta_{l'} + \sum_{k\neq l,l'}\eta_k\right)^N\bigg|_{\zeta,\zeta'=1}\\ &= \frac{\partial^2}{\partial\zeta\partial\zeta'}\left(\zeta\eta_l - \eta_l + \zeta'\eta_{l'} - \eta_{l'} + 1\right)^N\bigg|_{\zeta,\zeta'=1}\\ &= \frac{\partial}{\partial\zeta'}N\eta_l(\zeta'\eta_{l'} - \eta_{l'} + 1)^{N-1}\bigg|_{\zeta'=1}\\ &= N(N-1)\eta_l\eta_{l'}\,. \end{aligned}\tag{A.64}$$

For the step from the second to the third line of this equation, the multinomial theorem has been used.

If $l = l'$, one uses (5.6) to find

$$\overline{x_l(x_l-1)} = \frac{\partial^2}{\partial\zeta^2}\sum_x \zeta^{x_l}\,p(x|\eta)\bigg|_{\zeta=1}$$

$$
\begin{aligned}
&= \frac{\partial^2}{\partial\zeta^2} N! \sum_x \zeta^{x_l} \prod_{k=1}^{M} \frac{\eta_k^{x_k}}{x_k!}\Bigg|_{\zeta=1} \\
&= \frac{\partial^2}{\partial\zeta^2} \left(\zeta\eta_l + \sum_{k\neq l}\eta_k\right)^N\Bigg|_{\zeta=1} \\
&= \frac{\partial^2}{\partial\zeta^2} \left(\zeta\eta_l - \eta_l + 1\right)^N\Bigg|_{\zeta=1} \\
&= N(N-1)\eta_l^2 \,. \qquad (A.65)
\end{aligned}
$$

Here, too, the multinomial theorem has been used to obtain the third line of the equation. One thus obtains

$$\overline{x_l^2} = N(N-1)\eta_l^2 + N\eta_l \,. \qquad (A.66)$$

Finally, putting (A.64) and (A.66) together, one finds

$$\overline{x_l x_{l'}} = N(N-1)\eta_l\eta_{l'} + \delta_{ll'}N\eta_l \qquad (A.67)$$

which is (5.16).

A.5.2 A Limit of the Binomial Distribution

Prove that the Poisson distribution (5.19) is the limit of the binomial distribution (5.1) for $N \to \infty$ if $\eta \to \lambda/N$.

In this limit, one obtains

$$
\begin{aligned}
\eta^x(1-\eta)^{N-x} &= \exp\Big(x\ln(\lambda/N) + (N-x)\ln(1-\lambda/N)\Big) \\
& \exp\Big(x\ln\lambda - x\ln N - (N-x)\lambda/N\Big) \\
&\longrightarrow \exp\left(x\ln\lambda - x\ln N - \lambda\right) \\
&\longrightarrow \lambda^x N^{-x}\exp(-\lambda)\,. \qquad (A.68)
\end{aligned}
$$

The binomial coefficient approaches the limit

$$
\begin{aligned}
\binom{N}{x} &= \frac{N!}{x!(N-x)!} \\
&= \frac{1}{x!}\Big(N(N-1)\cdots(N-x+1)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{x!} N^x \left(1 \cdot (1 - 1/N) \cdots (1 - \frac{x-1}{N})\right) \\
&\longrightarrow \frac{N^x}{x!} \, . \qquad \text{(A.69)}
\end{aligned}
$$

Hence, the binomial distribution approaches

$$
p(x|\lambda) = \frac{\lambda^x}{x!} \exp(-\lambda) \qquad \text{(A.70)}
$$

which is the Poisson distribution.

A.6 Form Invariance I

A.6.1 Every Element Can Be Considered the Origin of a Group

The multiplication

$$
G_\rho = G_\xi \, G_\tau \qquad \text{(A.71)}
$$

of all elements G_ξ in $\mathcal{G}$ by a fixed G_τ is a one-to-one mapping of the group onto itself.

From (A.71) follows

$$
G_\xi = G_\rho \, G_\tau^{-1} \, . \qquad \text{(A.72)}
$$

Hence, for every G_ρ, one can find G_ξ. This means that every element of the group is reached by the mapping. It is reached only once because the pair of equations

$$
\begin{aligned}
G_\rho &= G_\xi \, G_\tau \\
&= G_{\xi'} \, G_\tau \qquad \text{(A.73)}
\end{aligned}
$$

entails

$$
G_\xi = G_{\xi'} \, . \qquad \text{(A.74)}
$$

A.6.2 The Domain of Definition of a Group Parameter Is Important

Why do the rotations

$$
G_\phi = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \qquad \text{(A.75)}
$$

with

$$0 \leq \phi < \pi \tag{A.76}$$

not form a group?

In this set of transformations, there is no inverse of G_ϕ when $\phi > 0$.

A.6.3 A Parameter Representation of the Hyperbola

Show that

$$a(\phi) = G_\phi \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{A.77}$$

is a parameter representation of the hyperbola of Fig. 6.2 when

$$\begin{aligned} G_\phi &= \begin{pmatrix} \cosh\phi & \sinh\phi \\ \sinh\phi & \cosh\phi \end{pmatrix} , \\ -\infty\ \phi &< \infty\,. \end{aligned} \tag{A.78}$$

One has

$$\begin{aligned} a(\phi) &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} \cosh\phi \\ \sinh\phi \end{pmatrix} , \end{aligned} \tag{A.79}$$

and one eliminates ϕ by forming the expression $x_1^2 - x_2^2$. This gives

$$x_1^2 - x_2^2 = 1\,. \tag{A.80}$$

This is the equation of the hyperbola. Actually, this equation allows for two branches placed symmetrically to the x_2-axis. The parameter representation (A.77) produces only the right-hand branch.

A.6.4 Multiplication Functions for the Symmetry Groups of the Circle and the Hyperbola

Show that for both groups, (A.75) and (A.78), the multiplication function is

$$\Phi(\phi', \phi) = \phi + \phi'\ . \tag{A.81}$$

For the rotations (A.75), the trigonometric formulae

$$\begin{aligned} \cos(\phi + \phi') &= \cos\phi \, \cos\phi' - \sin\phi \, \sin\phi' \,, \\ \sin(\phi + \phi') &= \sin\phi \, \cos\phi' + \cos\phi \, \sin\phi' \end{aligned} \tag{A.82}$$

yield

$$\begin{aligned} G_\phi \, G_{\phi'} &= \begin{pmatrix} \cos(\phi+\phi') & -\sin(\phi+\phi') \\ \sin(\phi+\phi') & \cos(\phi+\phi') \end{pmatrix} \\ &= G_{\phi+\phi'} \ . \end{aligned} \tag{A.83}$$

This proves (A.81) for the example of the rotations.

For the hyperbolic transformation, the formulae

$$\begin{aligned} \cosh(\phi + \phi') &= \cosh\phi \, \cosh\phi' + \sinh\phi \, \sinh\phi' \\ \sinh(\phi + \phi') &= \cosh\phi \, \sinh\phi' + \sinh\phi \, \cosh\phi' \end{aligned} \tag{A.84}$$

yield

$$G_\phi \, G_{\phi'} = \begin{pmatrix} \cosh(\phi+\phi') & \sinh(\phi+\phi') \\ \sinh(\phi+\phi') & \cosh(\phi+\phi') \end{pmatrix} \ . \tag{A.85}$$

This proves (A.81) for the group of hyperbolic transformations.

A.6.5 The Group of Dilations

Show that the dilations

$$G_\sigma \, x = \sigma x \, , \qquad 0 < \sigma < \infty \, , \tag{A.86}$$

form an Abelian group and that

$$\Phi(\sigma', \sigma) = \sigma \, \sigma' \tag{A.87}$$

is its multiplication function.

The multiplication function is obvious. It shows that the multiplication of G_σ with $G_{\sigma'}$ images the usual multiplication of the positive real numbers. Therefore the set of transformations G_σ endowed with the multiplication rule (A.87) is isomorphic to the positive real numbers endowed with their usual multiplication. By checking the axioms in Sect. 6.1 one easily verifies that these numbers form a group.

A.6.6 The Combination of Translations and Dilations

Consider the group of transformations

$$G_{\xi,\sigma}\,x = \xi + \sigma x \tag{A.88}$$

with

$$-\infty < \xi < \infty \quad \text{and} \quad 0 < \sigma < \infty\,; \tag{A.89}$$

Work out the multiplication function and show that translations in general do not commute with dilations.

The product of transformations yields

$$\begin{aligned} G_{\Phi(\xi',\sigma';\xi,\sigma)} &= G_{\xi,\sigma}\,G_{\xi',\sigma'}\,x \\ &= G_{\xi,\sigma}\,(\xi' + \sigma' x) \\ &= G_{0,\sigma}\,\Big(\xi' + \sigma'(\xi + x)\Big) \\ &= \xi' + \sigma'\xi + \sigma\sigma' x\ . \end{aligned} \tag{A.90}$$

From this, one sees that

$$\Phi(\xi', \sigma'; \xi, \sigma) = (\xi' + \xi\sigma', \sigma\sigma')\ . \tag{A.91}$$

The translation $(\xi', \sigma' = 1)$ followed by the dilation $(\xi = 0, \sigma)$ yields

$$\Phi(\xi', 1; 0, \sigma) = (\xi', \sigma)\ . \tag{A.92}$$

Reversing the operations leads to

$$\Phi(0, \sigma; \xi', 1) = (\sigma, \xi')\ . \tag{A.93}$$

The two results are generally different.

A.6.7 Reversing the Order of Translation and Dilation

Show that in Eq. (6.14) reversing the order of translation and dilation changes both Eq. (6.14) and the multiplication function (6.19).

After the interchange, Eq. (6.14) reads

$$\begin{aligned} G_{\sigma,\xi}\,x &= G_{\xi,1} G_{0,\sigma}\,x \\ &= \sigma(\xi + x)\,. \end{aligned} \tag{A.94}$$

The product $G_{\sigma,\xi}\, G_{\sigma',\xi'}$ applied to x yields

$$\begin{aligned} G_{\sigma,\xi}\, G_{\sigma',\xi'}\, x &= G_{\sigma,\xi}\left(\sigma'(\xi'+x)\right) \\ &= G_{\xi,1} G_{0,\sigma}\left(\sigma'(\xi'+x)\right) \\ &= G_{\xi,1}\left(\sigma'(\xi'+\sigma x)\right) \\ &= \sigma'\left(\xi'+\sigma(x+\xi)\right) \\ &= \sigma'\xi'+\sigma\sigma'\xi+\sigma\sigma' x \\ &= \sigma\sigma'\left(\frac{\xi'}{\sigma}+\xi+x\right). \end{aligned} \tag{A.95}$$

From this follows the multiplication function

$$\Phi(\sigma',\xi';\ \sigma,\xi) = \left(\sigma\sigma',\ \frac{\xi'}{\sigma}+\xi\right). \tag{A.96}$$

A.6.8 A Transformation of the Group Parameter

Show that the mapping

$$\tilde{G}_\tau\, \xi = \Phi(\xi;\tau) \tag{A.97}$$

is a transformation of the domain in which ξ is defined.

This is another version of problem Sect. A.6.1. However, the fixed and the running transformations are multiplied in reverse order. Still the solution is completely analogous to that of problem Sect. A.6.1.

A.6.9 A Group of Transformations of the Group Parameter

Show that the transformations $\tilde{G}_\tau$ form a group $\tilde{\mathcal{G}}$ when τ runs over all values of the group parameter of $\mathcal{G}$; show also that $\tilde{\mathcal{G}}$ and $\mathcal{G}$ are isomorphic.

We check the four axioms laid down in Sect. 6.1: The problem Sect. A.6.8 shows that $\tilde{G}_\tau$ is a transformation of ξ. Therefore the inverse $\tilde{G}_\tau^{-1}$ exists. Let ϵ be the value of ξ that labels the unit element $G_\epsilon = \mathbf{1}$ of the group $\mathcal{G}$. Because

$$\Phi(\xi;\epsilon) = \xi\ , \tag{A.98}$$

the group $\tilde{\mathcal{G}}$ also contains a unit element and its label is ϵ. The multiplication of the elements of $\mathcal{G}$ is associative, therefore the multiplication function has the property

$$\Phi(\Phi(\xi;\tau');\tau) = \Phi(\xi;\Phi(\tau';\tau))\ . \tag{A.99}$$

This shows that together with $\tilde{G}_\tau$ and $\tilde{G}_{\tau'}$ the product $\tilde{G}_\tau \tilde{G}_{\tau'}$ is in $\tilde{\mathcal{G}}$ and that the product is labelled $\Phi(\tau'; \tau)$. This means that $\mathcal{G}$ and $\tilde{\mathcal{G}}$ have the same multiplication function, whence they are isomorphic. The associativity of the multiplication in $\tilde{\mathcal{G}}$ is a consequence of the isomorphism.

A.6.10 The Model p Is Normalised when the Common Form w Is Normalised

Show that

$$p(x|\xi) = w(G_\xi^{-1} x) \left| \frac{\partial G_\xi^{-1} x}{\partial x} \right| \tag{A.100}$$

is normalised to unity for all ξ when w is normalised.

By a change of variables one comes from the first to the second one of the equations

$$\begin{aligned} \int dx\; p(x|\xi) &= \int w(G_\xi^{-1} x) \left| \frac{\partial G_\xi^{-1} x}{\partial x} \right| dx \\ &= \int w(G_\xi^{-1} x)\, dG_\xi^{-1} x \; . \end{aligned} \tag{A.101}$$

One integrates over the full range of x because G_ξ^{-1} is a transformation of the domain in which x is defined. By renaming the integration variable in the second version, one comes to the result

$$\begin{aligned} \int dx\; p(x|\xi) &= \int dx\; w(x) \\ &= 1 \; . \end{aligned} \tag{A.102}$$

A.6.11 Two Expressions Yielding the Measure μ

It is proven that the expressions (6.42) and (6.43) yield the same result.

We reformulate the second derivative of ln p in

$$\begin{aligned} \sum_x p(x|\xi) \frac{\partial^2}{\partial \xi^2} \ln p(x|\xi) &= \sum_x p(x|\xi) \frac{\partial}{\partial \xi} \frac{\frac{\partial}{\partial \xi} p}{p} \\ &= \sum_x \left(\frac{\partial^2}{\partial \xi^2} p(x|\xi) - p(x|\xi) \left(\frac{\frac{\partial}{\partial \xi} p}{p} \right)^2 \right) \end{aligned}$$

$$= \frac{\partial^2}{\partial \xi^2} \sum_x p(x|\xi)$$
$$- \sum_x p(x|\xi) \left(\frac{\partial}{\partial \xi} \ln p(x|\xi) \right)^2 . \qquad \text{(A.103)}$$

The derivative of the first sum on the r.h.s. of the last line vanishes because the sum equals unity for every ξ. The remaining term establishes the identity which was to be proven.

A.6.12 Form Invariance of the Posterior Distribution

Show that the posterior distribution of a form-invariant model is form invariant, too.

The symmetries of p and μ and m imply that

$$\begin{aligned} P(\xi|x) &= \frac{p(x|\xi)\mu(\xi)}{m(x)} \\ &= \frac{p(G_\rho x|G_\rho \xi)\mu(\xi)}{m(x)} \left| \frac{\partial G_\rho x}{\partial x} \right| \\ &= \frac{p(G_\rho x|G_\rho \xi)\mu(\xi)}{m(G_\rho x)} \\ &= \frac{p(G_\rho x|G_\rho \xi)\mu(G_\rho \xi)}{m(G_\rho x)} \left| \frac{\partial G_\rho \xi}{\partial \xi} \right| \\ &= P(G_\rho \xi|G_\rho x) \left| \frac{\partial G_\rho \xi}{\partial \xi} \right| . \end{aligned} \qquad \text{(A.104)}$$

Form invariance of p brings one from the first to the second line of this equation when $G_\rho \in \mathcal{G}$. The invariance of m leads to the third and the invariance of μ to the fourth line. By the definition of P, one arrives at the last line. It is analogous to (6.23); we therefore call it the form invariance of P.

A.6.13 Invariance of the Shannon Information

Show that the value of S in Eq. (6.56) does not change when ξ is substituted by

$$\eta = T\xi . \qquad \text{(A.105)}$$

Both the distribution P and the measure μ transform as densities; that is,

$$\begin{aligned} P(\xi|x) &= P(T^{-1}\eta|x)\left|\frac{\partial\eta}{\partial\xi}\right| \\ \mu(\xi) &= \mu(T^{-1}\eta)\left|\frac{\partial\eta}{\partial\xi}\right| . \end{aligned} \tag{A.106}$$

Compare Sect. 2.2. It follows

$$\begin{aligned} S &= \int \mathrm{d}\xi\, P(\xi|x) \ln\frac{P(\xi|x)}{\mu(\xi)} \\ &= \int \mathrm{d}\xi\, P(T^{-1}\eta|x)\left|\frac{\partial\eta}{\partial\xi}\right| \ln\frac{P(T^{-1}\eta|x)|\partial\eta/\partial\xi|}{\mu(T^{-1}\eta)|\partial\eta/\partial\xi|} \\ &= \int \mathrm{d}\xi\, P(T^{-1}\eta|x) \ln\frac{P(T^{-1}\eta|x)|}{\mu(T^{-1}\eta)} \end{aligned} \tag{A.107}$$

which says that the substitution of the integration variable boils down to renaming it. This does not affect the value of the integral.

A.7 Examples of Invariant Measures

A.7.1 The Invariant Measure of the Group of Translation-Dilation

Show that the group (7.13)—the combination of translation and dilation—has the invariant measure

$$\mu(\xi,\sigma) \propto \sigma^{-1} . \tag{A.108}$$

The multiplication function (7.15) is

$$\Phi(\xi',\sigma';\xi,\sigma) = (\xi'+\xi\sigma';\sigma\sigma') . \tag{A.109}$$

The Jacobian matrix of the derivatives of Φ with respect to its primed variables is

$$\frac{\partial\Phi(\xi',\sigma';\xi,\sigma)}{\partial(\xi',\sigma')} = \begin{pmatrix} 1 & 0 \\ \xi & \sigma \end{pmatrix} . \tag{A.110}$$

By (6.33), it yields the invariant measure

$$\begin{aligned} \mu(\xi,\sigma) &= \mu(0,1)\begin{vmatrix} 1 & 0 \\ \xi & \sigma \end{vmatrix}^{-1} \\ &= \mu(0,1)\,\sigma^{-1} . \end{aligned} \tag{A.111}$$

A.7.2 Groups of Finite Volume

The group of transformations G_ϕ has a finite volume (7.25). Show that the volume V of the space of ϕ does not depend on the parameterisation. When the integral over ϕ does not exist, show that this fact is again independent of the parameterisation.

Let us reparameterise via the transformation

$$\eta = T\phi \; . \tag{A.112}$$

The transformed measure is

$$\mu_T(\eta) = \mu(\phi) \left| \frac{\partial \phi}{\partial \eta} \right| \; . \tag{A.113}$$

Therefore the volume

$$V = \int \mu(\phi)\, \mathrm{d}\phi \tag{A.114}$$

of the space of ϕ can be rewritten

$$V = \int \mu_T(\eta)\, \mathrm{d}\eta \tag{A.115}$$

by the change (A.112) of integration variables. Thus V has one and the same value in all parameterisations. By consequence: when the integral over ϕ does not exist, then the integral over η will not exist either.

A.7.3 The Inverse of a Triangular Matrix

Show that the inverse of the matrix

$$G_{\alpha,\gamma,\beta} = \begin{pmatrix} \alpha\,, & \gamma \\ 0\,, & \beta \end{pmatrix} \tag{A.116}$$

has the index $(\alpha^{-1}, -\gamma(\alpha\beta)^{-1}, \beta^{-1})$.

By Eq. (7.38) the index of the unit element is

$$\epsilon = (1, 0, 1)\,. \tag{A.117}$$

We obtain the inverse of $G_{\alpha,\gamma,\beta}$ from the multiplication function (7.40). This inverse has the index $(\alpha', \gamma', \beta')$ which solves the equation

$$(1, 0, 1) = (\alpha\alpha', \alpha\gamma' + \gamma\beta', \beta\beta')\,. \tag{A.118}$$

For α' and β' the solutions obviously are

$$\begin{aligned}\alpha' &= \alpha^{-1}\,,\\ \beta' &= \beta^{-1}\,.\end{aligned} \tag{A.119}$$

For γ' we solve $\alpha\gamma' + \gamma\beta' = 0$ and obtain

$$\gamma' = -\gamma(\alpha\beta)^{-1}\,. \tag{A.120}$$

Thus Eq. (7.39) is correct.

A.7.4 The Invariant Measure of a Group of Triangular Matrices

Show that the invariant measure of the group of matrices (A.116) is given by the determinant in Eq. (7.41).

We use Eq. (6.33) to find the invariant measure. The matrix in (7.41) contains the partial derivatives with respect to the primed quantities in the multiplication function (7.40). The derivatives with respect to α' are written in the first column of the matrix; the derivatives with respect to γ' and β' follow in the next two columns. This yields the matrix as well as its inverse determinant that are given in Eq. (7.41).

A.8 A Linear Representation of Form Invariance

A.8.1 Transforming a Space of Square-Integrable Functions

Show that

$$\mathbf{T}f = f(Tx)\left|\frac{\partial Tx}{\partial x}\right|^{1/2} \tag{A.121}$$

is a square-integrable function.

In the context of Sect. 8.2, the function f is square integrable and T is a transformation of x. By making a change of the integration variable from x to Tx one sees that $\mathbf{T}f$ is square integrable.

A.8.2 An Integral Kernel

Verify that

$$\mathbf{T}_{xx'} = \delta(x' - T^{-1}x)\left|\frac{\partial T^{-1}x}{\partial x}\right|^{1/2} \tag{A.122}$$

is the integral kernel of the operator $\mathbf{T}$.

This is a consequence of the properties of the δ distribution. We have

$$\begin{aligned}\int dx'\,\mathbf{T}_{xx'}f(x') &= \int \mathrm{d}x'\,\delta(x'-T^{-1}x)\left|\frac{\partial T^{-1}x}{\partial x}\right|^{1/2} f(x') \\ &= f(T^{-1}x)\left|\frac{\partial T^{-1}x}{\partial x}\right|^{1/2} \\ &= \mathbf{T}f \end{aligned} \tag{A.123}$$

for every element f of the function space.

A.9 Beyond Form Invariance: The Geometric Prior

A.9.1 Jeffreys' Rule Transforms as a Density

Show that Eq. (9.1) transforms as a density.

According to Eq. (9.4) we can write Jeffreys's rule in the form

$$\mu(\xi) \propto \det\left(\left(\frac{\partial}{\partial_\xi}\right)\left(\frac{\partial}{\partial_{\xi'}}\right)^\dagger a^\dagger(\xi)a(\xi')\Big|_{\xi=\xi'}\right)^{1/2} \tag{A.124}$$

and we express ξ via a transformation

$$\eta = T\xi \tag{A.125}$$

by the parameter(s) η. This gives

$$\begin{aligned}\mu(\xi) &= \det\left(\left(\frac{\partial}{\partial_{T^{-1}\eta}}\right)\left(\frac{\partial}{\partial_{T^{-1}\eta'}}\right)^\dagger a^\dagger(T^{-1}\eta)a(T^{-1}\eta')\Big|_{\eta=\eta'}\right)^{1/2} \\ &= \det\left(\left(\frac{\partial}{\partial_\eta}\right)\left(\frac{\partial}{\partial_{\eta'}}\right)^\dagger a^\dagger(T^{-1}\eta)a(T^{-1}\eta')\Big|_{\eta=\eta'}\right)^{1/2}\left|\frac{\partial\eta}{\partial\xi}\right| \\ &= \mu_T(\eta)\left|\frac{\partial\eta}{\partial\xi}\right| \ . \end{aligned} \tag{A.126}$$

We proceed from the first to the second line of this equation with the help of the rule (9.9) and by factorising the determinant into a product of two determinants. The third line is obtained when we observe that the first factor of the second line is the result of Jeffreys' rule when the model p is considered to be conditioned by η. As a result, Jeffreys' rule transforms in agreement with (2.9).

A.9.2 The Fisher Matrix Is Positive Definite

Show that the eigenvalues of the matrix F of Eq. (9.4) with the elements

$$F_{\nu,\nu'} = 4\int \mathrm{d}x\, \left(\frac{\partial}{\partial\xi_\nu} a_x(\xi)\right)\left(\frac{\partial}{\partial\xi_{\nu'}} a_x(\xi)\right) \tag{A.127}$$

are not negative.

The matrix is real and symmetric. Hence, there is an orthogonal transformation O that diagonalises it according to

$$F[\mathrm{diag}] = O^\dagger F O\ . \tag{A.128}$$

This means that the elements of $F[\mathrm{diag}]$ can be written

$$\left(O^\dagger F O\right)_{\kappa,\kappa'} = 4\int \mathrm{d}x \sum_{\nu=1}^{n} O_{\nu,\kappa}\Big(\frac{\partial}{\partial\xi_\nu} a_x(\xi)\Big) \sum_{\nu'=1}^{n} O_{\nu',\kappa'} \frac{\partial}{\partial\xi_{\nu'}} a_x(\xi)\,. \tag{A.129}$$

One cannot diagonalise F with the same transformation O at every ξ ; the transformation O depends on ξ. However, the differentiations acting on $a(\xi)$ do not act on O. For a given ξ the elements in the diagonal of $F[\mathrm{diag}]$ are the eigenvalues of F. We obtain

$$\begin{aligned} F[\mathrm{diag}]_{\kappa,\kappa} &= 4\int \mathrm{d}x \Big(\sum_{\nu=1}^{n} O_{\nu,\kappa}\frac{\partial}{\partial\xi_\nu} a_x(\xi)\Big)^2 \\ &> 0\,. \end{aligned} \tag{A.130}$$

The integrand in this equation is not only nonnegative, it is positive because we require all eigenvalues of F to be nonzero.

A.9.3 The Measure on the Sphere

Prove that the geometric measure on the surface of an M-dimensional sphere with unit radius is given by (9.20).

Let us set

$$\omega_M = \left(1 - \sum_{k=1}^{M-1} \omega_k^2\right)^{1/2}\ . \tag{A.131}$$

From (9.19), we obtain the matrix of tangential vectors

$$\frac{\partial a}{\partial \omega} = \begin{pmatrix} 1 & 0 & 0 & \dots & -\omega_1/\omega_M \\ 0 & 1 & 0 & & -\omega_2/\omega_M \\ \vdots & & \ddots & & \\ 0 & & & 1 & -\omega_{M-1}/\omega_M \end{pmatrix} . \tag{A.132}$$

This matrix has $M - 1$ rows and M columns.

The desired measure is

$$\mu(\omega) = \det\left(\frac{\partial a}{\partial \omega} \left(\frac{\partial a}{\partial \omega} \right)^{\dagger} \right)^{1/2} . \tag{A.133}$$

We find the product of matrices to be

$$\frac{\partial a}{\partial \omega} \left(\frac{\partial a}{\partial \omega} \right)^{\dagger} = \mathbf{1}_{M-1} + \hat{\omega}(\hat{\omega})^{\dagger} . \tag{A.134}$$

This is the sum of a unit matrix and a dyadic product. The vector $\hat{\omega}$ is

$$\hat{\omega} = \omega_M^{-1} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_{M-1} \end{pmatrix} \tag{A.135}$$

According to Sect. D.2 the determinant of (A.134) is

$$\begin{aligned} \det\left(\frac{\partial a}{\partial \omega} \left(\frac{\partial a}{\partial \omega} \right)^{\dagger} \right) &= 1 + (\hat{\omega})^{\dagger} \hat{\omega} \\ &= \left(1 - \sum_{k=1}^{M-1} \omega_k^2 \right)^{-1} . \end{aligned} \tag{A.136}$$

This proves (9.20).

A.9.4 Another Form of the Measure on the Sphere

Show that (9.22) is the measure on the sphere in terms of the parameters η introduced by (9.21).

The transformation (9.21) implies the Jacobian determinant

$$\left| \frac{\partial \omega}{\partial \eta} \right| = \prod_{k=1}^{M-1} \left(\frac{1}{2} \eta_k^{-1/2} \right) . \tag{A.137}$$

Therefore the transformation brings the measure (9.22) into the form

$$\begin{aligned}\mu_T(\eta) &= \mu(\omega)\prod_{k=1}^{M-1}\left(\frac{1}{2}\eta_k^{-1/2}\right)\\ &= 2^{1-M}\left(1-\sum_{k=1}^{M-1}\eta_k\right)^{-1/2}\prod_{k=1}^{M-1}\eta_k^{-1/2}\\ &= 2^{1-M}\eta_M^{-1/2}\prod_{k=1}^{M-1}\eta_k^{-1/2}\\ &= 2\prod_{k=1}^{M}\left(\frac{1}{2}\eta_k^{-1/2}\right),\end{aligned} \tag{A.138}$$

where η_M is defined in (9.23).

A.10 Inferring the Mean or the Standard Deviation

A.10.1 Calculation of a Fisher Matrix

Verify Eq. (10.22).

We have to calculate the expectation values of the second derivatives given in Eq. (10.21). The expression $e^{-2\eta}$ is independent of x. Therefore

$$\begin{aligned}-\overline{\frac{\partial^2}{\partial\xi^2}\ln q} &= \int \mathrm{d}x\, q(x|\xi,\eta)e^{-2\eta}\\ &= e^{-2\eta}\,.\end{aligned} \tag{A.139}$$

Here, q is given by Eq. (10.1). The expectation value of $x-\xi$ vanishes and

$$-\overline{\frac{\partial^2}{\partial\xi\partial\eta}\ln q} = 0\,. \tag{A.140}$$

The expectation value of $(x-\xi)^2$ equals $\sigma^2 = e^{2\eta}$ and

$$-\overline{\frac{\partial^2}{\partial\eta^2}\ln q} = 2\,. \tag{A.141}$$

This yields the Fisher matrix (10.22).

A.10.2 The Expectation Value of an ML Estimator

The expectation value is calculated of the ML estimator $(\sigma^{\mathrm{ML}})^2$ given in Eq. 10.30.

The expectation value

$$\begin{aligned}\overline{(\sigma^{\mathrm{ML}})^2} &= \frac{1}{N}\sum_{i=1}^{N}\left(\overline{\langle x^2\rangle_i} - \overline{\langle x\rangle_i^2}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\overline{x_{i,j}^2} - \overline{\left(\frac{1}{n}\sum_{j=1}^{n}x_{i,j}\right)^2}\right)\end{aligned} \tag{A.142}$$

is considered. In the second line of this equation, the expectation values—indicated by overlines—are taken with respect to the distribution (10.23). This distribution entails Eq. (10.30) from which we take

$$\overline{x_{i,j}^2} = \sigma^2 + \xi_i^2\,. \tag{A.143}$$

It thus follows

$$\begin{aligned}\overline{(\sigma^{\mathrm{ML}})^2} &= \frac{1}{N}\sum_i\left(n^{-1}\sum_j(\sigma^2+\xi^2) - n^{-2}\sum_{j,j'}\overline{x_{i,j}x_{i,j'}}\right)\\ &= \frac{1}{N}\sum_i\left(\sigma^2+\xi_i^2 - n^{-2}\sum_{j\neq j'}\overline{x_{i,j}x_{i,j'}} - n^{-2}\sum_j\overline{x_{i,j}^2}\right)\\ &= \frac{1}{N}\sum_i\left(\sigma^2+\xi_i^2 - n^{-2}n(n-1)\xi_i^2 - -n^{-2}\sum_j(\sigma^2+\xi_i^2)\right)\\ &= \frac{1}{N}\sum_i\left(\sigma^2+\xi_i^2 - n^{-1}(n-1)\xi_i^2 - n^{-1}(\sigma^2-\xi^2)\right)\\ &= \frac{1}{N}\sum_i\sigma^2(1-n^{-1})\\ &= \sigma^2(1-n^{-1})\,.\end{aligned} \tag{A.144}$$

This proves Eq. (10.30).

A.11 Form Invariance II: Natural *x*

A.11.1 The Identity of Two Expressions

Show that the expression (11.7) agrees with (11.8).

This means to prove the equation

$$\exp(i\xi\mathbf{g}) = \begin{pmatrix} \cos\xi & \sin\xi \\ -\sin\xi & \cos\xi \end{pmatrix} . \tag{A.145}$$

Some powers of $i\xi\mathbf{g}$ are

$$\begin{aligned} i\xi\mathbf{g} &= \xi \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} , \\ (i\xi\mathbf{g})^2 &= \xi^2 \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} , \\ (i\xi\mathbf{g})^3 &= \xi^3 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} , \\ (i\xi\mathbf{g})^4 &= \xi^4 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ (i\xi\mathbf{g})^5 &= \xi^5 i\mathbf{g} , \\ (i\xi\mathbf{g})^6 &= \xi^6 (i\mathbf{g})^2 . \end{aligned} \tag{A.146}$$

One recognises the rule

$$(i\mathbf{g})^k = (i\mathbf{g})^{(k \text{ modulo } 4)} . \tag{A.147}$$

This leads to

$$\exp(i\xi\mathbf{g}) = \sum_{k=0}^{\infty} \frac{\xi^k}{k!} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}^k \tag{A.148}$$

$$\begin{aligned} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \xi \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \frac{\xi^2}{2!} \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \\ &\quad + \frac{\xi^3}{3!} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + \frac{\xi^4}{4!} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \dots \\ &= \begin{pmatrix} \sum_k (-)^k \frac{xi^{2k}}{(2k)!} & \sum_k (-)^{2k+1} \frac{\xi^{2k+1}}{(2k+1)!} \\ -\sum_k (-)^{2k+1} \frac{xi^{2k+1}}{(2k)!} & \sum_k (-)^k \frac{xi^{2k}}{(2k)!} \end{pmatrix} \\ &= \begin{pmatrix} \cos\xi & \sin\xi \\ -\sin\xi & \cos\xi \end{pmatrix} , \end{aligned} \tag{A.149}$$

where the sums over k extend from 0 to ∞. One obtains the result to be proven.

A.11.2 Form Invariance of the Binomial Model

Show that the binomial model is form invariant.

If the amplitude vector $a(\xi)$ is obtained from an initial one $a(\epsilon)$ via a linear transformation

$$a(\xi) = G_\xi \, a(\epsilon) \tag{A.150}$$

and the set of transformations G_ξ forms a group, then the model is form invariant. The binomial model fulfils this condition, because all vectors (11.5) are obtained from (11.6) via the transformations (11.7) exactly once if one takes care of the periodicity of ξ. The unit element is among the G_ξ. Equation (11.8) shows that the product $G_\xi G_{\xi'}$ of any two G_ξ, $G_{\xi'}$ is among the G_ξ. This equation also shows that the product is associative.

A.11.3 The Multiplication Function of a Group of Matrices

Show that the multiplication function of Φ of the group of matrices (11.7) is

$$\Phi(\xi; \xi') = \xi' + \xi \, . \tag{A.151}$$

According to Eq. (11.8) the product of two matrices is

$$\begin{aligned} G_\xi G_{\xi'} &= \exp(i\xi\mathbf{g}) \exp(i\xi'\mathbf{g}) \\ &= \exp(i(\xi + \xi')\mathbf{g}) \, . \end{aligned} \tag{A.152}$$

This proves (A.151).

A.11.4 An ML Estimator for the Binomial Model

Show that the ML estimator from the event n obtained by the binomial model (11.14) is given by

$$\sin^2(\xi^{\mathrm{ML}}(n)) = \frac{n}{N} \, . \tag{A.153}$$

The ML equation is given in (11.18). It entails

$$\begin{aligned} 0 &= n\cos^2\xi - (N - n)\sin^2\xi \\ &= n(1 - \sin^2\xi) - (N - n)\sin^2\xi \\ &= n - N\sin^2\xi \end{aligned} \tag{A.154}$$

from which (A.153) immediately follows.

A.11.5 A Prior Distribution for the Poisson Model

Verify Eq. (11.24).

The derivative of $a_n(\lambda)$ is

$$\begin{aligned}\frac{\partial}{\partial\lambda}a_n(\lambda) &= \left(\frac{n}{2\sqrt{n!}} - \frac{\lambda^{n/2}}{2\sqrt{n!}}\right)\exp(-\lambda/2)\\ &= \frac{1}{2}a_n(\lambda)(n\lambda^{-1}-1)\,.\end{aligned} \tag{A.155}$$

This gives

$$\left(\frac{\partial}{\partial\lambda}a_n(\lambda)\right)^2 = \frac{1}{4}p(n|\lambda)\left(\frac{n}{\lambda}-1\right)^2 \tag{A.156}$$

and

$$(\mu_T(\lambda))^2 = \sum_{n=0}^{\infty} p(n|\lambda)\left(\frac{n}{\lambda}-1\right)^2 \tag{A.157}$$

which is Eq. (11.24).

A.11.6 A Limiting Case of the Poisson Model

Verify Eq. (11.30).

In Eq. (11.28) the value of $\xi = 0$ entails $p(n|\xi) = 0$ for every $n \neq 0$. If $n = 0$ one obtains

$$p(0|0) = 1 \tag{A.158}$$

because $0^0 = 1$ as well as $0! = 1$.

A.12 Item Response Theory

A.12.1 Expectation Values Given by the Binomial Model

Find the expectation values $\overline{x^2}$ and $\overline{x(1-x)}$ and $\overline{(1-x)^2}$ from the binomial model (12.1).

The expectation value of a function $f(x)$ is

$$\begin{aligned}\overline{f(x)} &= \sum_{x=0}^{1} q(x|\theta_p, \sigma_i)\, f(x) \\ &= \sum_{x=0}^{1} [R(\theta_p, \sigma_i)]^x [1 - R(\theta_p, \sigma_i)]^{1-x}\, f(x)\,. \end{aligned} \tag{A.159}$$

The expression $x^2\, q$ equals zero for $x = 0$; it equals $R(\theta_p, \sigma_i)$ for $x = 1$. Hence, one obtains

$$\overline{x^2} = R(\theta_p, \sigma_i)\,. \tag{A.160}$$

The product $x(1-x)\, q$ vanishes for $x = 0$ as well as for $x = 1$. This gives

$$\overline{x(1-x)} = 0\,. \tag{A.161}$$

The product $(1-x)^2\, q$ equals R for $x = 0$ and vanishes for $x = 1$. Thus one gets

$$\overline{(1-x)^2} = R(\theta_p,\, \sigma_i)\,. \tag{A.162}$$

A.13 On the Art of Fitting

A.13.1 A Maximum Likelihood Estimator

Show that for the model (13.7) the ML estimator of the parameter η is given by (13.10).

The common scale of y and η in the model (13.7) has a uniform measure. Therefore the distribution $\tilde{\chi}_N^{\mathrm{sq}}$ is a likelihood function. See the definition of likelihood in Eq. (3.13). We determine the place of its maximum by solving the ML equation

$$0 = \frac{\partial}{\partial \eta} \ln \tilde{\chi}_N^{\mathrm{sq}}(y|\eta)\,. \tag{A.163}$$

This leads to the equation

$$0 = -\frac{N}{2} + e^{y-\eta} \tag{A.164}$$

which is solved by

$$\eta^{\mathrm{ML}} = y - \ln(N/2)\,. \tag{A.165}$$

A.13.2 Gaussian Approximation to a Chi-Squared Model

Show that the Gaussian approximation to the distribution $\tilde{\chi}_N^{\text{sq}}(y|\eta')$ in Eq. (13.12) is given by Eq. (13.15).

According to Eq. (13.13) the maximum of the likelihood lies at $\eta' = y$. Therefore the Gaussian approximation is a function of the difference $y - \eta'$. The inverse variance of the Gaussian equals the Fisher function F. Equation (9.18) shows that $F/4$ equals the square of the geometric measure (13.8). This gives

$$F = \frac{N}{2} \tag{A.166}$$

in agreement with (13.14). Therefore the variance σ^2 of the Gaussian (4.1) is set equal to $2/N$ and we obtain the approximation

$$\begin{aligned}
\tilde{\chi}_N^{\text{sq}}(y|\eta') &\approx \frac{1}{\sqrt{2\pi}}\left(\frac{N}{2}\right)^{1/2} \exp\left(-\frac{N}{2}\,\frac{(y-\eta')^2}{2}\right) \\
&\approx \left(\frac{N}{4\pi}\right)^{1/2} \exp\left(-\frac{N}{4}(y-\eta')^2\right).
\end{aligned} \tag{A.167}$$

Appendix B
Description of Distributions I: Real *x*

B.1 The Correlation Matrix

We show that the matrix C that appears in the multidimensional Gaussian model (4.16) has the elements

$$C_{\nu\nu'} = \overline{(x-\xi)_\nu (x-\xi)_{\nu'}}\,. \tag{B.1}$$

Here, the expectation value is taken with respect to the distribution (4.16), (4.17).

In order to prove Eq. (B.1) we consider the function

$$\mathrm{L}(\boldsymbol{\xi}) = \ln p(\boldsymbol{x}|\boldsymbol{\xi})\,. \tag{B.2}$$

Its derivative with respect to ξ_ν is

$$\begin{aligned} \frac{\partial}{\partial \xi_\nu} L(\boldsymbol{\xi}) &= \sum_{\nu'} (C^{-1})_{\nu\nu'} (x_{\nu'} - \xi_{\nu'}) \\ &= \left(C^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\right)_\nu \,. \end{aligned} \tag{B.3}$$

If we define the n-dimensional vector of derivatives

$$\left(\vec{\partial}\, L\right) = \begin{pmatrix} \partial/\partial\xi_1\, L \\ \vdots \\ \partial/\partial\xi_n\, L \end{pmatrix}, \tag{B.4}$$

Equation (B.3) can be written as

$$C\vec{\partial}\, L = \boldsymbol{x} - \boldsymbol{\xi} \tag{B.5}$$

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

and (B.1) takes the form

$$C_{\nu,\nu'} = \int \mathrm{d}^n x \; p(\boldsymbol{x}|\boldsymbol{\xi}) \; \Big((C\vec{\partial}L)(C\vec{\partial}L)^{\dagger} \Big)_{\nu,\nu'} \; . \tag{B.6}$$

Here,

$$\begin{aligned} (C\vec{\partial}L)^{\dagger} &= (\vec{\partial}L)^{\dagger}C^{\dagger} \\ &= (\vec{\partial}L)^{\dagger}C \end{aligned} \tag{B.7}$$

is the transpose of the vector $C\vec{\partial}L$ because C is symmetric. One arrives at

$$C_{\nu,\nu'} = \int \mathrm{d}^n x \; p(\boldsymbol{x}|\boldsymbol{\xi}) \; \Big(C\vec{\partial}L(\vec{\partial}L)^{\dagger}C \Big)_{\nu,\nu'} \; . \tag{B.8}$$

Now we use the identity

$$\int \mathrm{d}^n x \; p(\boldsymbol{x}|\boldsymbol{\xi}) \left(\frac{\partial}{\partial \xi_\nu} \ln p \right) \left(\frac{\partial}{\partial \xi_{\nu'}} \ln p \right) = - \int \mathrm{d}^n x \; p(\boldsymbol{x}|\boldsymbol{\xi}) \frac{\partial^2}{\partial \xi_\nu \partial \xi_{\nu'}} \ln p \, . \tag{B.9}$$

To derive it, one uses the fact that $p(x|\xi)$ is normalised to unity for every value of ξ. This can be rewritten as

$$\int \mathrm{d}^n x \; p(\mathbf{x}|\boldsymbol{\xi}) \; \Big(\vec{\partial}L(\vec{\partial}L)^{\dagger} \Big)_{\nu,\nu'} = - \int \mathrm{d}^n x \; p(\boldsymbol{x}|\boldsymbol{\xi}) \; \big(\partial^2 L \big)_{\nu,\nu'} \tag{B.10}$$

when $\partial^2 L$ means the matrix of second derivatives of L. This brings (B.8) into the form

$$C_{\nu\nu'} = - \int \mathrm{d}^n x \; p(\boldsymbol{x}|\boldsymbol{\xi}) \; \big(C(\partial^2 L)C \big)_{\nu\nu'} \; . \tag{B.11}$$

The matrix of second derivatives of the logarithm of the Gaussian distribution (4.16) equals the matrix $-C^{-1}$. The matrix C is independent of $\boldsymbol{x}$ and $\boldsymbol{\xi}$ so that the r.h.s of (B.11) is equal to $C_{\nu,\nu'}$. This was to be proven.

B.2 Projecting the Multidimensional Gaussian

We prove the rule formulated at the end of Sect. 4.1.2 which says that the integration over one of the event variables x_ν of the Gaussian model (4.16) yields a Gaussian model whose correlation matrix K is obtained by omitting the νth row and the νth column of the original correlation matrix C.

The numerical value of the vector $\boldsymbol{\xi}$ in Eq. (4.16) is immaterial for what follows. We set it equal to zero and consider the Gaussian model

$$p(\boldsymbol{x}|0) = ((2\pi)^n \det C)^{-1/2} \exp\left(-\boldsymbol{x}^\dagger (2C)^{-1}\boldsymbol{x}\right) . \tag{B.12}$$

In order to integrate over the variable x_n the expression $\boldsymbol{x}^\dagger(2C)^{-1}\boldsymbol{x}$ is rewritten as follows

$$\begin{aligned}
\boldsymbol{x}^\dagger(2C)^{-1}\boldsymbol{x} &= \sum_{\nu,\nu'=1}^{n} x_\nu (C^{-1})_{\nu,\nu'} x_{\nu'} \\
&= \sum_{\nu,\nu'=1}^{n-1} x_\nu (C^{-1})_{\nu,\nu'} x_{\nu'} \\
&\quad + 2x_n \sum_{\nu=1}^{n-1} \left(C^{-1}\right)_{n,\nu} x_\nu + x_n^2 \left(C^{-1}\right)_{n,n} \\
&= \sum_{\nu,\nu'=1}^{n-1} x_\nu \left(C^{-1}\right)_{\nu,\nu'} x_{\nu'} \\
&\quad + \left(C^{-1}\right))_{n,n} \left(x_n + \frac{\sum_{\nu=1}^{n-1} (C^{-1})_{n,\nu} x_\nu}{(C^{-1})_{n,n}} \right)^2 \\
&\quad - \frac{\left(\sum_{\nu=1}^{n-1} (C^{-1})_{n,\nu} x_\nu \, .\right)^2}{(C^{-1})_{n,n}} \, .
\end{aligned} \tag{B.13}$$

Integrating the exponential function of this expression over x_n removes the term containing x_n in the third version of this equation. This is the only term to depend on x_n and the shift of

$$\frac{\sum_{\nu=1}^{n-1} (C^{-1})_{n,\nu} x_\nu}{(C^{-1})_{n,n}}$$

is immaterial. Thus the integration leads to

$$p^\downarrow \propto \exp\left(- \sum_{\nu,\nu'=1}^{n-1} x_\nu (K^{-1})_{\nu,\nu'} x_{\nu'} \right) , \tag{B.14}$$

where the matrix K^{-1} has the elements

$$(K^{-1})_{\nu,\nu'} = (C^{-1})_{\nu,\nu'} - \frac{(C^{-1})_{n,\nu} (C^{-1})_{n,\nu'}}{C^{-1}_{n,n}} \, . \tag{B.15}$$

The eigenvalues of K are all positive; otherwise one could not integrate expression (B.13) over $x_1, \ldots, x_{n-1}$. This integral exists, however, because (B.12) is a proper distribution. This shows that $p^{\downarrow}$ is a multidimensional Gaussian distribution.

The expectation value of $x_\nu x_{\nu'}$ for $\nu, \nu' < n$ is the same with respect to both $p(\boldsymbol{x}|0)$ and $p^{\downarrow}$; that is,

$$\int \mathrm{d}x_1 \ldots \mathrm{d}x_n \; x_\nu x_{\nu'} \, p(\boldsymbol{x}|0) = \int \mathrm{d}x_1 \ldots \mathrm{d}x_{n-1} \; x_\nu x_{\nu'} \, p^{\downarrow} \, . \tag{B.16}$$

According to Sect. B.1 this entails

$$C_{\nu,\nu'} = K_{\nu,\nu'} \quad \text{for } \nu, \nu' = 1, \ldots, n-1 \, , \tag{B.17}$$

which was to be shown.

B.3 Calculation of a Jacobian

The Jacobian J of the transformation from $(r_1, \ldots, r_N)$ to $(T, r_1, \ldots, r_{N-1})$ defined by Eqs. (4.25) and (4.26), and also used in (12.7, 12.8), is calculated.

The Jacobian is the determinant

$$\begin{aligned} J &= \left| \frac{\partial(r_1, \ldots, r_n)}{\partial(T, t_1, \ldots t_{N-1})} \right| \\ &= \begin{vmatrix} t_1 & t_2 & t_3 & \ldots & t_N \\ T & 0 & 0 & \ldots & -T \\ 0 & T & 0 & \ldots & -T \\ 0 & & \ddots & & \\ 0 & \ldots & 0 & T & -T \end{vmatrix} \, . \end{aligned} \tag{B.18}$$

Here, in the first line one finds the derivatives of the r_k with respect to T ; the second line contains the derivatives of the r_k with respect to t_1 ; in the third line there are the derivatives of the r_k with respect to t_2 ; and so on. In the last line, the derivatives with respect to t_{N-1} are given. By extracting the factor T from $N-1$ lines, one finds

$$J = T^{N-1} \begin{vmatrix} t_1 & t_2 & t_3 & \ldots & t_N \\ 1 & 0 & 0 & \ldots & -1 \\ 0 & 1 & 0 & \ldots & -1 \\ \vdots & & \ddots & & \\ 0 & \ldots & 0 & 1 & -1 \end{vmatrix} \, . \tag{B.19}$$

The determinant in (B.19) is equal to unity. One shows this by expanding it with respect to the last column; that is,

$$\begin{aligned}
J\,T^{-N+1} &= t_N + \begin{vmatrix} t_1 & t_2 & t_3 & \dots & t_{N-1} \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & \\ \vdots & & & \ddots & \\ 0 & \dots & & 0 & 1 \end{vmatrix} \\
&\quad - \begin{vmatrix} t_1 & t_2 & t_3 & \dots & t_{N-1} \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ 0 & & & 0 & 1 \end{vmatrix} \\
&\quad + \dots \\
&\quad + (-)^{N-1} \begin{vmatrix} t_1 & t_2 & t_3 & \dots & t_{N-1} \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & & \ddots & \\ 0 & & & 1 & 0 \end{vmatrix} . \qquad \text{(B.20)}
\end{aligned}$$

The first determinant on the r.h.s. has the value of t_1. The following ones are found by expanding them with respect to the second, third, and so on, column. This yields

$$\begin{aligned}
J\,T^{-N+1} &= t_N + \sum_{k=1}^{N-1} t_k \\
&= 1 \qquad \text{(B.21)}
\end{aligned}$$

and proves (4.27).

B.4 Properties of the Γ Function

The Γ function is the analytical continuation of the factorials,

$$\Gamma(n) = (n-1)!\,. \qquad \text{(B.22)}$$

It has the integral representation

$$\Gamma(z) = \int_0^\infty dt\, t^{z-1} \exp(-t)\,; \qquad \text{(B.23)}$$

given by Euler[1]; compare Sect. 8.310 of [2]. It satisfies the functional equation

$$\Gamma(z+1) = z\Gamma(z)\,. \tag{B.24}$$

Special values are

$$\begin{aligned}\Gamma(1/2) &= \sqrt{\pi}\,,\\ \Gamma(1) &= 1\,,\\ \Gamma(2) &= 1\,.\end{aligned} \tag{B.25}$$

The logarithm of the Γ function can be calculated with the help of a MacLaurin formula based on the Euler procedure of differences; see pp. 269 and 396 of [3]. We choose the version

$$\begin{aligned}\ln\Gamma(z) \approx{}& z\ln z - z - \frac{1}{2}\ln z + \ln\sqrt{2\pi}\\ &+\sum_{k=1}^{n-1}\frac{B_{2k}}{2k(2k-1)z^{2k-1}} + R_n(z)\,.\end{aligned} \tag{B.26}$$

for positive $\Re z$ which is given in Sect. 8.344 [2]. In this expression, R_n is the remainder of the series. For real z, it satisfies the inequality

$$|R_n(z)| < \frac{|B_{2n}|}{2n(2n-1)z^{2n-1}}\,. \tag{B.27}$$

The B_{2k} are the Bernoulli numbers

$$\begin{aligned}B_2 &= \frac{1}{6}\\ B_4 &= -\frac{1}{30}\,,\end{aligned} \tag{B.28}$$

see Sect. 9.71 of [2].

The present Eqs. (B.26), (B.27) are confirmed by a few examples in Table B.1, where the equations have been used with $n = 2$; that is,

$$\ln\Gamma(z) \approx z\ln z - z - \frac{1}{2}\ln z + \ln\sqrt{2\pi} + \frac{1}{12z} + R_2(z)\,. \tag{B.29}$$

[1] Leonhard Euler, 1707–1783, Swiss mathematician, associate of the St. Petersburg Academy of Science (since 1727), member of the Berlin Academy (since 1741). He achieved for modern geometry what Euclid's *Elements* had done for ancient geometry [1]; but he formulated it no longer in terms of points and lines but rather via arithmetical and analytical relations. He discovered the exponential function and the natural logarithm. He found the identity $e^{i\theta} = \cos\theta + i\sin\theta$.

Table B.1 An Euler–MacLaurin Series

z	$\Gamma(z)$	$\ln\Gamma(z)$	Eq. (B.29)	Δ	$1/(360z^3)$
1	1	0	2.273×10^{-3}	-2.273×10^{-3}	2.78×10^{-3}
1.5	0.886227	−0.120782	−0.120036	-7.44×10^{-4}	8.23×10^{-4}
2	1	0	3.27×10^{-4}	-3.27×10^{-4}	3.47×10^{-4}
3	2	0.693147	0.693248	-1.00×10^{-4}	1.03×10^{-4}
4	6	1.791760	1.791802	-4.28×10^{-5}	4.34×10^{-5}
1/2	1.772454	0.572365	0.252272	0.320	0.022

and

$$R_2(z) < \frac{1}{360z^3}\,. \tag{B.30}$$

The fifth column of Table B.1 gives the difference Δ between the correct value of $\ln\Gamma(z)$ and the approximation (B.29) without the remainder R_2. The last column shows the upper limit (B.30) of Δ. For the values with $z \geq 1$ given in the table, the approximation (B.26) is seen to be valid.

B.5 The Beta Function

The Beta function is defined as

$$B(x,\ y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\ ; \tag{B.31}$$

compare Sect. 8.384 of [2]. The integral representation

$$B(\mu,\nu-\mu) = \beta^{\mu}\int_0^{\infty}\mathrm{d}x\ \frac{x^{\mu-1}}{(1+\beta x)^{\nu}}\ , \tag{B.32}$$

given in Sect. 3.194 of [2], yields the normalisations of (4.43) and (4.49).

For $\mu = 1/2$ and $\nu = 1$, the substitution $\beta X = t^2\gamma^{-2}$ yields the integral

$$\begin{aligned}\int_{-\infty}^{\infty} dt\,\left(1+\frac{t^2}{\gamma^2}\right)^{-1} &= \gamma\ B(1/2,1/2)\\ &= \gamma\pi\end{aligned} \tag{B.33}$$

which normalises the Cauchy distribution (4.45).

References

1. Encyclopædia Britannica (2005)
2. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products* (Academic Press, New York, 2015)
3. I.N. Bronstein, K.A. Semendjajew, *Taschenbuch der Mathematik*, 25th edn. (B.G. Teubner, Stuttgart, 1991)

Appendix C
Form Invariance I

C.1 The Invariant Measure of a Group

We show that expression (6.33) is an invariant distribution; that is, it possesses the symmetry (6.30).

The proof essentially relies on the fact that the multiplication of group elements is associative. We rewrite $\mu(G_\rho \xi)$ in a series of steps starting from Eq. (6.33),

$$\begin{aligned}\mu(G_\rho \xi) &= \mu(\epsilon) \left| \frac{\partial \Phi(\tau; G_\rho \xi)}{\partial \tau} \right|_{\tau=\epsilon}^{-1} \\ &= \mu(\epsilon) \left| \frac{\partial \Phi\Big(\tau; \Phi(\xi; \rho)\Big)}{\partial \tau} \right|^{-1} \\ &= \mu(\epsilon) \left| \frac{\partial \Phi\Big(\Phi(\tau; \xi); \rho\Big)}{\partial \tau} \right|^{-1} .\end{aligned} \tag{C.1}$$

In all the lines of this equation, it is understood that after the differentiation with respect to τ, one sets $\tau = \epsilon$. The definition of $G_\rho \xi$ leads from the first to the second line. The associativity of the multiplication of group elements brings one to the third line. More steps are needed. We write

$$\varphi = \Phi(\tau; \xi) \tag{C.2}$$

to obtain

$$\mu(G_\rho \xi) = \mu(\epsilon) \left| \frac{\partial \Phi(\varphi; \rho)}{\partial \varphi} \right|_{\varphi=\Phi(\tau;\xi)}^{-1} \left| \frac{\partial \Phi(\tau; \xi)}{\partial \tau} \right|^{-1}$$

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

$$
\begin{aligned}
&= \mu(\epsilon) \left| \frac{\partial \Phi(\xi; \rho)}{\partial \xi} \right|^{-1} \frac{\mu(\xi)}{\mu(\epsilon)} \\
&= \left| \frac{\partial G_\rho \xi}{\partial \xi} \right|^{-1} \mu(\xi) \ . \qquad \text{(C.3)}
\end{aligned}
$$

By differentiating an implicit function one obtains the first line of (C.3) from the last line of (C.1). With $\tau \to \epsilon$ the second line of (C.3) is reached. This leads to the last line which expresses the claimed symmetry of $\mu(\xi)$.

C.2 On the Existence of the Measure $m(x)$ in the Space of the Events

To show that the measure $m(x)$, defined by Eq. (2.5), exists, takes a somewhat lengthy argument. It is presented here.

The transformation T of (6.49) maps the space of events x onto the space of parameters ξ^{ML} and further coordinates x^ϵ,

$$
x \longrightarrow Tx = \Big(\xi^{\text{ML}}(x), x^\epsilon(x)\Big) , \qquad \text{(C.4)}
$$

such that a transformation $G_\rho \in \mathcal{G}$, when applied to Tx, acts only on the value of ξ^{ML}, not on x^ϵ,

$$
G_\rho(\xi^{\text{ML}}, x^\epsilon) = (G_\rho \xi^{\text{ML}}, x^\epsilon) \, . \qquad \text{(C.5)}
$$

Because

$$
G_\rho \, \xi^{\text{ML}}(x) = \xi^{\text{ML}}(G_\rho x) \, , \qquad \text{(C.6)}
$$

the transformations T and G_ρ commute with each other.

We start from the definition (2.5) of the measure $m(x)$ and rewrite it with the help of the form invariance of the model p,

$$
\begin{aligned}
m(x) &= \int \mathrm{d}\xi \; p(x|\xi)\mu(\xi) \\
&= \int \mathrm{d}\xi \; p(G_\xi^{-1} x|\epsilon) \left| \frac{\partial G_\xi^{-1} x}{\partial x} \right| \mu(\xi) \, . \qquad \text{(C.7)}
\end{aligned}
$$

Inasmuch as $m(x)$ is a density, it transforms according to

$$
m(x)\, \mathrm{d}x = m_T(Tx)\, \mathrm{d}Tx \ , \qquad \text{(C.8)}
$$

whence, in terms of the coordinates Tx, Eq. (C.7) reads

$$\begin{aligned} m(x)\,\mathrm{d}x &= \left[\int \mathrm{d}\xi\, p(G_\xi^{-1}Tx|\epsilon)\left|\frac{\partial G_\xi^{-1}Tx}{\partial Tx}\right|\mu(\xi)\right]\mathrm{d}Tx \\ &= \left[\int \mathrm{d}\xi\, p(G_\xi^{-1}\xi^{\mathrm{ML}}, x^\epsilon|\epsilon)\left|\frac{\partial\Big(G_\xi^{-1}\xi^{\mathrm{ML}}, x^\epsilon\Big)}{\partial(\xi^{\mathrm{ML}}, x^\epsilon)}\right|\mu(\xi)\right]\mathrm{d}Tx\,. \end{aligned} \tag{C.9}$$

We write

$$G_\xi^{-1}\,\xi^{\mathrm{ML}} = \Phi(\xi^{\mathrm{ML}};\overline{\xi}) \tag{C.10}$$

and use the fact that the Jacobian matrix within Eq. (C.9) is blockwise diagonal; that is,

$$\frac{\partial\Big(G_\xi^{-1}\xi^{\mathrm{ML}}, x^\epsilon\Big)}{\partial(\xi^{\mathrm{ML}}, x^\epsilon)} = \begin{pmatrix} \frac{\partial\Phi(\xi^{\mathrm{ML}};\overline{\xi})}{\partial\xi^{\mathrm{ML}}} & , 0 \\ 0 & \mathbf{1} \end{pmatrix}. \tag{C.11}$$

In the upper left part, there is an n-dimensional Jacobian matrix; in the lower right part, there is an $(N-n)$-dimensional unit matrix, if the x space is N-dimensional and the ξ space is n-dimensional. The structure of the matrix (C.11) turns the last line of Eq. (C.9) into

$$m(x)\mathrm{d}x = \left[\int \mathrm{d}\xi\, p\Big(\Phi(\xi^{\mathrm{ML}};\overline{\xi}), x^\epsilon|\epsilon\Big)\left|\frac{\partial\Phi(\xi^{\mathrm{ML}};\overline{\xi})}{\partial\xi^{\mathrm{ML}}}\right|\mu(\xi)\right]\mathrm{d}Tx\,. \tag{C.12}$$

Without loss of generality, we can choose the parameterisation of the group $\mathcal{G}$ such that the event x has led to the ML estimator $\xi^{\mathrm{ML}}(x) = \epsilon$. Then we obtain

$$m(x)\mathrm{d}x = \left[\int \mathrm{d}\xi\, p(\overline{\xi}, x^\epsilon|\epsilon)\left|\frac{\partial\Phi(\xi^{\mathrm{ML}};\overline{\xi})}{\partial\xi^{\mathrm{ML}}}\right|_{\xi^{\mathrm{ML}}=\epsilon}\right]\mathrm{d}Tx\,. \tag{C.13}$$

The Jacobian determinant in this expression is the (inverse of) the invariant measure (6.30) albeit given as a function of $\overline{\xi}$, not ξ. The mapping $\xi \to \overline{\xi}$ will not be among the transformations that leave (6.30) invariant. Therefore the (inverse of) the Jacobian in (C.13) is proportional to $\tilde{\mu}(\overline{\xi})$, where μ and $\tilde{\mu}$ are related via

$$\mu(\xi)\mathrm{d}\xi = \tilde{\mu}(\overline{\xi})\mathrm{d}\overline{\xi}\,. \tag{C.14}$$

This brings (C.13) into the form

$$m(x)\mathrm{d}x = \left[\int \mathrm{d}\xi\, p(\overline{\xi}, x^\epsilon|\epsilon)\frac{\mu(\epsilon)\mu(\xi)}{\tilde{\mu}(\overline{\xi})}\right]\mathrm{d}Tx$$

$$= \left[\mu(\epsilon)\int \mathrm{d}\bar{\xi}\; p(\bar{\xi}, x^{\epsilon}|\epsilon)\right] \mathrm{d}Tx$$

$$= \left[\mu(\epsilon)\int \mathrm{d}\xi\; p(\xi, x^{\epsilon}|\epsilon)\right] \mathrm{d}Tx\,. \tag{C.15}$$

The second line of this equation is an application of (C.14). The last line simply renames the integration variable.

The result says that the transformation of $m(x)$ from the variables x to the variables Tx leads to the integration of $p(x|\xi)$ over a surface within the x space. Because p is normalised to unity, its integral over the entire space of x exists. Therefore the integral, that defines $m(x)$, exists.

Appendix D
Beyond Form Invariance: The Geometric Prior

D.1 The Definition of the Fisher Matrix

The definition of the Fisher matrix in Eq. (9.2) can be rewritten in the following way.

$$\begin{aligned} F_{\nu,\nu'} &= -\int dx\ p(x|\xi)\frac{\partial^2}{\partial_\nu\partial_{\nu'}}\ln p(x|\xi) \\ &= \int dx\ p(x|\xi)\Big(\frac{\partial}{\partial\xi_\nu}\ln p(x|\xi)\Big)\frac{\partial}{\partial\xi_{\nu'}}\ln p(x|\xi) - \int dx\ \frac{\partial^2}{\partial\xi_\nu\partial\xi_{\nu'}}p(x|\xi) \\ &= \int dx\ p(x|\xi)\Big(\frac{\partial}{\partial\xi_\nu}\ln p(x|\xi)\Big)\frac{\partial}{\partial\xi_{\nu'}}\ln p(x|\xi)\,. \end{aligned} \tag{D.1}$$

The second line of this equation is reached by calculating the second derivatives of $\ln p$. The third line is a consequence of the fact that $\int \mathrm{d}x\ p(x|\xi)$ equals unity for every ξ.

This result can be expressed by the amplitudes

$$a_x(\xi) = \sqrt{p(x|\xi)} \tag{D.2}$$

because

$$\frac{\partial}{\partial\xi_\nu}a_x(\xi) = \frac{1}{2}\frac{\partial}{\partial\xi_\nu}\ln p(x|\xi)\,. \tag{D.3}$$

We obtain

$$F_{\nu,\nu'} = 4\int dx\ \Big(\frac{\partial}{\partial\xi_\nu}a_x(\xi)\Big)\frac{\partial}{\partial\xi_{\nu'}}a_x(\xi)\,. \tag{D.4}$$

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

D.2 Evaluation of a Determinant

The determinant of a matrix, which is the sum

$$F = \mathbf{1} + ff^{\dagger}\,, \tag{D.5}$$

of the unit operator and the dyadic product $ff^{\dagger}$, is now calculated.

One uses the identity

$$\det F = \exp(\mathrm{tr}\ln F) \tag{D.6}$$

valid for any matrix F. One can verify this equation by diagonalising F. Hence, the determinant of the matrix (D.5) is given by

$$\begin{aligned}\det F &= \exp\big(\mathrm{tr}\ln(1 + ff^{\dagger})\big)\\ &= \exp\left(\mathrm{tr}\sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}(ff^{\dagger})^{l}\right)\\ &= \exp\left(\sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}(f^{\dagger}f)^{l}\right)\\ &= \exp\big(\ln(1 + f^{\dagger}f)\big)\\ &= 1 + f^{\dagger}f\end{aligned} \tag{D.7}$$

The second line of this equation is obtained by the Taylor expansion of the matrix $\ln(1 + ff^{\dagger})$. The third line results from the identity

$$\mathrm{tr}(ff^{\dagger})^{l} = (f^{\dagger}f)^{l}\,, \qquad \text{for } l \geq 1\,. \tag{D.8}$$

The fourth line evaluates the sum over l. The procedure holds for $|f^{\dagger}f| < 1$. In an analogous way, we find the result

$$\det(F^{-1}) = 1 - f^{\dagger}f\,. \tag{D.9}$$

D.3 Evaluation of a Fisher Matrix

The Fisher matrix (9.8) is calculated for the model with the amplitudes (9.25). The elements of F are given by

$$\frac{1}{4}F_{\nu\nu'} = \frac{\partial}{\partial\omega_{\nu}}\frac{\partial}{\partial\omega'_{\nu'}}\int \mathrm{d}x\,\Big(\sum_{\rho=1}^{n}\omega_{\rho}c_{x}(\rho)\Big)\Big(\sum_{\rho'=1}^{n}\omega_{\rho'}c_{x}(\rho')\Big)\bigg|_{\omega=\omega'}$$

$$= \frac{\partial}{\partial\omega_\nu}\frac{\partial}{\partial\omega'_{\nu'}}\sum_{\rho=1}^{n}\omega_\rho\omega'_\rho\bigg|_{\omega=\omega'} . \tag{D.10}$$

The matrix F is $(n-1)$-dimensional; that is, $\nu, \nu' = 1, \dots, n-1$. The second line of this equation is due to the orthonormality of the basis functions $c(\rho)$ for $\rho = 1, \dots, n$. The coefficients ω_ρ are related to each other via the normalisation (9.26). We define

$$\omega_n = \Big(1 - \sum_{\rho=1}^{n-1}\omega^2\Big)^{1/2} . \tag{D.11}$$

This allows us to rewrite the elements of the Fisher matrix

$$\begin{aligned}
\frac{1}{4}F_{\nu\nu'} &= \frac{\partial}{\partial\omega_\nu}\frac{\partial}{\partial\omega'_{\nu'}}\left(\sum_{\rho=1}^{n-1}\omega_\rho\omega'_\rho + \sqrt{1-\sum_\rho\omega_\rho^2}\sqrt{1-\sum_{\rho'}{\omega'_{\rho'}}^2}\right)\Bigg|_{\omega=\omega'} \\
&= \frac{\partial}{\partial\omega_\nu}\left(\omega_\rho - \frac{(1-\sum_\rho\omega_\rho^2)^{1/2}}{(1-\sum_{\rho'}\omega'^2_{\rho'})^{1/2}}\,\omega'_{\nu'}\right)\Bigg|_{\omega=\omega'} \\
&= \delta_{\nu\nu'} + (1-\sum_\rho\omega_\rho^2)^{-1/2}\,(1-\sum_{\rho'}\omega'^2_{\rho'})^{-1/2}\,\omega_{\nu'}\omega'_\nu\Bigg|_{\omega=\omega'} \\
&= \delta_{\nu\nu'} + \frac{\omega_\nu\omega_{\nu'}}{\omega_n^2} .
\end{aligned} \tag{D.12}$$

By the rule derived in Sect. D.2, the determinant of the Fisher matrix is

$$\begin{aligned}
\det\Big(\frac{1}{4}F\Big) &= 1 + \sum_{\nu=1}^{n-1}\frac{\omega_\nu^2}{\omega_n^2} \\
&= \omega_n^{-2}\left(\omega_n^2 + \sum_{\nu=1}^{n-1}\omega_\nu^2\right) \\
&= \omega_n^{-2} ,
\end{aligned} \tag{D.13}$$

whence we obtain the prior distribution (9.28). It is a geometric measure.

D.4 The Fisher Matrix of the Multinomial Model

The Fisher matrix (9.2) of the multinomial model (9.32) is now found.

Let us write the Fisher matrix in the form of the third line of (D.16). This gives

$$\begin{aligned}
\frac{1}{4}F_{\nu,\nu'} &= \frac{1}{4}\sum_{x_1\ldots x_M} p(x|\omega)\Big(\frac{\partial}{\partial\omega_\nu}\ln p(x|\omega)\Big)\frac{\partial}{\partial\omega_{\nu'}}\ln p(x|\omega) \\
&= \frac{1}{4}\sum_{x_1\ldots x_M} p(x|\omega)\left(\frac{\partial}{\partial\omega_\nu}\sum_{k=1}^{M}\ln\beta_k^{2x_k}\right)\frac{\partial}{\partial\omega_{\nu'}}\sum_{k'=1}^{M}\ln\beta_{k'}^{2x_{k'}} \\
&= \sum_{x_1\ldots x_M} p(x|\omega)\left(\sum_{k=1}^{M} x_k\frac{\partial}{\partial\omega_\nu}\ln\beta_k\right)\sum_{k'=1}^{M} x_{k'}\frac{\partial}{\partial\omega_{\nu'}}\ln\beta_{k'} \\
&= \sum_{k,k'=1}^{M}\overline{x_k x_{k'}}\left(\frac{\partial}{\partial\omega_\nu}\ln\beta_k\right)\frac{\partial}{\partial\omega_{\nu'}}\ln\beta_{k'}\ . \qquad (D.14)
\end{aligned}$$

Here, $\overline{x_k x_{k'}}$ is the expectation value

$$\overline{x_k x_{k'}} = \sum_{x_1\ldots x_M} p(x|\omega)\, x_k x_{k'} \qquad (D.15)$$

treated in Eq. (5.16). According to Eq. (9.31) the β_k are given by

$$\beta_k = \sum_{\nu=1}^{n}\omega_\nu c_k(\nu).$$

For (9.2) to be valid we assume that none of the β_k vanishes. From (9.26) we take

$$\omega_n = \Big(1-\sum_{\rho=1}^{n-1}\omega_\rho^2\Big)^{1/2}. \qquad (D.16)$$

This yields the derivative

$$\frac{\partial}{\partial\omega_\nu}\ln\beta_k = \beta_k^{-1}\Big(c_k(\nu)-\frac{\omega_\nu}{\omega_n}c_k(n)\Big). \qquad (D.17)$$

In the present context, the result (5.16) of Chap. 5 reads

$$\overline{x_k x_{k'}} = N(N-1)\beta_k^2\beta_{k'}^2 + N\beta_k^2\delta_{kk'}\ . \qquad (D.18)$$

Thus the last line of (D.14) can be written

$$\begin{aligned}
\frac{1}{4}F_{\nu,\nu'} = \sum_{k,k'=1}^{M}&\Big(N(N-1)\beta_k^2\beta_{k'}^2+\delta_{kk'}N\beta_k^2\Big) \\
&\times(\beta_k\beta_{k'})^{-1}\Big(c_k(\nu)-\frac{\omega_\nu}{\omega_n}c_k(n)\Big)\Big(c_{k'}(\nu')-\frac{\omega_{\nu'}}{\omega_n}c_{k'}(n)\Big)
\end{aligned}$$

$$
\begin{aligned}
&= N(N-1)\left(\sum_{k=1}^{M} \beta_k\Big(c_k(\nu) - \frac{\omega_\nu}{\omega_n} c_k(n)\Big)\right)\sum_{k'=1}^{M} \beta_{k'}\Big(c_{k'}(\nu') - \frac{\omega_{\nu'}}{\omega_n} c_{k'}(n)\Big) \\
&\quad + N\sum_{k=1}^{M}\Big(c_k(\nu) - \frac{\omega_\nu}{\omega_n} c_k(n)\Big)\Big(c_k(\nu') - \frac{\omega_{\nu'}}{\omega_n} c_k(n)\Big) . \qquad \text{(D.19)}
\end{aligned}
$$

From the orthogonality of the $c(\nu)$ follows that the first term on the r.h.s. of the last line of this equation vanishes. For this we consider

$$
\begin{aligned}
\sum_{k=1}^{M} \beta_k\Big(c_k(\nu) - \frac{\omega_\nu}{\omega_n} c_k(n)\Big) &= \sum_{k=1}^{M}\left(\sum_{\rho=1}^{n} \omega_\rho c_k(\rho)\right)\Big(c_k(\nu) - \frac{\omega_\nu}{\omega_n} c_k(n)\Big) \\
&= \sum_{\rho=1}^{n} \omega_\rho \sum_{k=1}^{M}\Big(c_k(\rho)c_k(\nu) - \frac{\omega_\nu}{\omega_n} c_k(\rho)c_k(n)\Big) \\
&= \sum_{\rho=1}^{n} \omega_\rho\Big(\delta_{\rho\nu} - \frac{\omega_\nu}{\omega_n}\delta_{\rho n}\Big) \\
&= \omega_\nu - \omega_\nu \\
&= 0 \, . \qquad \text{(D.20)}
\end{aligned}
$$

Therefore the Fisher matrix becomes

$$
\begin{aligned}
\frac{1}{4} F_{\nu,\nu'} &= N\sum_{k=1}^{M}\Big(c_k(\nu) - \frac{\omega_\nu}{\omega_n} c_k(n)\Big)\Big(c_k(\nu') - \frac{\omega_{\nu'}}{\omega_n} c_k(n)\Big) \\
&= N\Big(\delta_{\nu\nu'} + \frac{\omega_\nu \omega_{\nu'}}{\omega_n^2}\Big) . \qquad \text{(D.21)}
\end{aligned}
$$

The second line of this equation is obtained because $\nu, \nu' = 1, \ldots, n-1$ which entails that $c(\nu)$ and $c(\nu')$ are both orthogonal to $c(n)$. The second line was to be proven.

Appendix E
Inferring Mean or Standard Deviation

E.1 Normalising the Posterior Distribution of ξ, σ

The normalisation of the distribution (10.5) is verified; that is, the integral

$$m(x_1 \dots x_N) = \int_{-\infty}^{\infty} d\xi \int_0^{\infty} d\sigma\, \sigma^{-N-2} \exp\left(-\frac{N}{2\sigma^2}\left[(\xi - \langle x \rangle)^2 + V\right]\right) \qquad \text{(E.1)}$$

must be calculated. The integral over ξ is given by the normalisation of the simple Gaussian of Problem Sect. A.3.3. The integral over σ is obtained with the help of the substitution

$$\begin{aligned} \lambda &= \sigma^{-2}\,, \\ d\lambda &= -2\sigma^{-3} d\sigma\,. \end{aligned} \qquad \text{(E.2)}$$

One finds

$$\begin{aligned} m(x) &= (2\pi/N)^{1/2} \int_0^{\infty} d\sigma\, \sigma^{-N-1} \exp\left(-\frac{NV}{2\sigma^2}\right) \\ &= (2\pi/N)^{1/2}\, 2^{-1} \int_0^{\infty} d\lambda\, \lambda^{(N-2)/2} \exp\left(-\frac{NV}{2}\lambda\right)\,. \end{aligned} \qquad \text{(E.3)}$$

By Appendix B.4, this integral yields a Γ function, namely

$$m(x) = \pi^{1/2} 2^{(N-1)/2} N^{-(N+1)/2}\, V^{-N/2}\, \Gamma(N/2)\,. \qquad \text{(E.4)}$$

From this result, (10.5) immediately follows.

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

E.2 Rewriting a Product of Gaussian Models

We show that the first line of Eq. (10.24) agrees with the second line.

For this the double sum over the squared differences $(x_{i,j} - \xi_i)^2$ is rewritten as follows

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{n}(x_{i,j} - \xi_i)^2 &= \sum_{i=1}^{N}\sum_{j=1}^{n}(x_{i,j}^2 - 2\xi_i x_{i,j} + \xi_i^2)\\ &= \sum_i\sum_j x_{i,j}^2 + 2(\sum_i \xi_i)(\sum_j x_{i,j}) + \sum_i n\xi_i^2 \,. \qquad \text{(E.5)}\end{aligned}$$

Making use of the notation defined by Eq. (10.27) this turns into

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{n}(x_{i,j} - \xi_i)^2 &= n\sum_i \Big(\langle x^2\rangle_i - 2\xi\langle x\rangle_i + \xi^2\Big)\\ &= n\sum_i \Big(\langle x^2\rangle_i + (\xi_i - \langle x\rangle_i)^2 - \langle x\rangle_i^2\Big)\,. \qquad \text{(E.6)}\end{aligned}$$

with the definition of V_i in Eq. (10.26) we obtain the second line of Eq. (10.24).

Appendix F
Form Invariance II: Natural *x*

It is shown in which way the generator **g** of (11.33) is related to creation and destruction operators. For this we introduce the operator

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{4} & 0 \\ 0 & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix} \tag{F.1}$$

Its elements are

$$\mathbf{A}_{x,x'} = \sqrt{x'}\delta_{x+1,x'}\,, \qquad x = 0, 1, 2, \dots\,. \tag{F.2}$$

Thus the operator **g** of (11.33) can be expressed as

$$\mathbf{g} = i(\mathbf{A}^{\dagger} - \mathbf{A})\,, \tag{F.3}$$

where $\mathbf{A}^{\dagger}$ is the transpose of **A** and has the elements

$$\mathbf{A}^{\dagger}_{x,x'} = \sqrt{x}\delta_{x,x'+1}\,. \tag{F.4}$$

In order to interpret these operators we look at their products

$$\begin{aligned} (\mathbf{A}\mathbf{A}^{\dagger})_{xx'} &= \sum_{x''} \mathbf{A}_{xx''}\mathbf{A}^{\dagger}_{x''x'} \\ &= \sum_{x''} \sqrt{x''}\delta_{x+1,x''}\sqrt{x''}\delta_{x'',x'+1} \\ &= (x+1)\delta_{x,x'} \end{aligned} \tag{F.5}$$

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

and similarly

$$(\mathbf{A}^\dagger\mathbf{A})_{x,x'} = x\delta_{x,x'}\,. \tag{F.6}$$

The difference of the last two equations is the commutator

$$\begin{aligned}[\mathbf{A}\mathbf{A}^\dagger, \mathbf{A}^\dagger\mathbf{A}] &= \mathbf{A}\mathbf{A}^\dagger - \mathbf{A}^\dagger\mathbf{A} \\ &= \delta_{x,x'}\,, \qquad x, x' = 0, 1, 2, \ldots \end{aligned} \tag{F.7}$$

or simply

$$[\mathbf{A}\mathbf{A}^\dagger, \mathbf{A}^\dagger\mathbf{A}] = \mathbf{1}\,. \tag{F.8}$$

This means that the operators $\mathbf{A}$ and $\mathbf{A}^\dagger$ are destruction and creation operators, respectively. Such operators are a basic tool of quantum field theory. In the present context they destroy and create events of the Poisson distribution.

Note that the diagonal operator

$$\mathbf{A}^\dagger\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & 0 & 2 & 0 & \\ & 0 & 0 & 3 & \ddots \\ & \vdots & \vdots & \ddots & \ddots \end{pmatrix}\,. \tag{F.9}$$

shows in its diagonal elements all possible events of the Poisson model. It is called the number operator.

Appendix G
Item Response Theory

G.1 ML Estimators from Guttman Data

We want to prove that the ML estimators for the Guttman scheme of Table 12.1 are given by

$$\begin{aligned} \theta_p^{\mathrm{ML}} &= \left(t_p - \frac{N}{2}\right)\Delta\,, \qquad p = 1, \ldots, N\,, \\ \sigma_i^{\mathrm{ML}} &= \left(\frac{N+1}{2} - s_i\right)\Delta\,, \qquad i = 1, \ldots, N\,, \end{aligned} \tag{G.1}$$

where the Guttman scheme yields the scores

$$\begin{aligned} t_p &= N + 1 - p\,, \\ s_i &= i\,. \end{aligned} \tag{G.2}$$

The value of the step width

$$\Delta = \frac{\pi}{2N} \tag{G.3}$$

is given in Eq. (12.14).

The data matrix $\mathbf{x}$ of Table 12.1 is characterised by

$$\begin{aligned} x_{p,i} &= 1 \quad \text{for} \quad i \geq p\,, \\ &= 0 \quad \text{for} \quad i < p\,. \end{aligned} \tag{G.4}$$

This turns the ML equations (12.11) into

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

$$0 = \sum_{i=p}^{N} \cot\Big(\pi/4 + \theta_p - \sigma_i\Big) - \sum_{i=1}^{p-1} \tan\Big(\pi/ + \theta_p - \sigma_i\Big)\,, \quad p = 1, \ldots, N\,;$$
$$0 = \sum_{p=1}^{i} \cot\Big(\theta_p - \sigma_i\Big) - \sum_{p=i+1}^{N} \tan\Big(\pi/4 + \theta_p - \sigma_i\Big)\,, \quad i = 1, \ldots, N \qquad \text{(G.5)}$$

From Eqs. (G.1), (G.2) and (G.3) follows

$$\theta_p - \sigma_i = (i + 1/2 - p)\frac{\pi}{2N}\,. \qquad \text{(G.6)}$$

The first of the ML equations (G.5) then reads

$$0 = \sum_{i=p}^{N} \cot\Big(\pi/4 + \theta_p - \sigma_i\Big) - \sum_{i=1}^{p-1} \tan\Big(\pi/ + \theta_p - \sigma_i\Big)\,. \qquad \text{(G.7)}$$

We use the identity

$$\cot\alpha = -\tan(\alpha - \pi/2) \qquad \text{(G.8)}$$

to express the cot function in (G.7) by a tan function. Together with Eq. (G.6) this turns Eq. (G.7) into

$$0 = \sum_{i=p}^{N} \tan\left(-\pi/4 + (i + 1/2 - p)\frac{\pi}{2N}\right) + \sum_{i=1}^{p-1} \tan\left(\pi/4 + (i + 1/2 - p)\frac{\pi}{2N}\right)\,. \qquad \text{(G.9)}$$

Substituting in the first sum on the r.h.s. the index of summation by

$$i' = i - p - N/2 \qquad \text{(G.10)}$$

and

$$i'' = i - p + 1/2 \qquad \text{(G.11)}$$

we obtain

$$\begin{aligned} 0 &= \sum_{i'=-N/2}^{-p+N/2} \tan\left((i' + 1/2)\frac{\pi}{2N}\right) + \sum_{i''=1-p+N/2}^{-1+N/2} \left((i'' + 1/2)\frac{\pi}{2N}\right) \\ &= \sum_{i'=-N/2}^{N/2-1} \tan\left((i' + 1/2)\frac{\pi}{2N}\right) \end{aligned} \qquad \text{(G.12)}$$

The substitution

$$i = i' + 1/2 \qquad \text{(G.13)}$$

turns this into

$$0 = \sum_{i=-N/2+1/2}^{N/2-1/2} \tan\left(i\frac{\pi}{2N}\right) . \tag{G.14}$$

In the Guttman scheme of Sect. 12.3.1 the number N is even; therefore the summation index i takes half-integer values in steps of unity. Any two terms in the sum that have indices of opposite sign cancel each other. Therefore the sum on the r.h.s. of Eq. (G.14) is indeed equal to zero.

This proves that the ML estimators (12.12) satisfy the first of the ML equations (12.11). In an analogous way one proves that they satisfy the second ML equation too.

G.2 The Fisher Matrix of the Trigonometric Model

It is shown that the Fisher matrix of the trigonometric model (12.2)–(12.4) has the structure described in Eqs. (12.15)–(12.19).

The element p, p' in the upper left block a of the Fisher matrix is defined by

$$\begin{aligned} F_{p,p'} &= \sum_{k,k'=1}^{N_P} \sum_{l,l'=1}^{N_I} \overline{\left(\frac{\partial}{\partial\theta_p} \ln q(x_{k,l}|R_{k,l})\right) \frac{\partial}{\partial\theta_{p'}} \ln q(x_{k',l'}|R_{k',l'})} \\ p, p' &= 1, \dots, N_P \, . \end{aligned} \tag{G.15}$$

Here, $R_{k,l}$ stands for the IRF given in Eq. (12.7)

$$\begin{aligned} R_{k,l} &= R(\theta_k - \sigma_l) \\ &= \sin^2(\pi/4 + \theta_k - \sigma_l) \, . \end{aligned} \tag{G.16}$$

The overline means the expectation value with respect to the event variables $\boldsymbol{x}$. The variable $x_{k,l}$ is statistically independent of $x_{k',l'}$ for $(k, l) \neq (k', l')$. In this case the expectation value of the product of derivatives in (G.15) equals the product of the expectation values of the derivatives. However, one has

$$\begin{aligned} \overline{\frac{\partial}{\partial\theta_p} \ln q(x_{p,l}|R_{p,l})} &= \frac{\partial}{\partial\theta_p} \sum_{x=0}^{1} q(x|R_{p,l}) \\ &= 0 \, . \end{aligned} \tag{G.17}$$

This derivative vanishes because the distribution q is normalised to unity for every value of its parameter. Hence, Eq. (G.15) simplifies to

$$F_{p,p'} = \sum_{k=1}^{N_P}\sum_{l=1}^{N_I} \overline{\left(\frac{\partial}{\partial\theta_p}\ln q(x_{k,l}|R_{k,l})\right)\frac{\partial}{\partial\theta_{p'}}\ln q(x_{k,l}|R_{k,l})}\,. \tag{G.18}$$

For the partial derivatives in this equation to be different from zero, the indices p and p' must be equal to k. Thus the last equation simplifies further to

$$\begin{aligned} F_{p,p'} &= \delta_{p,p'}\sum_{l=1}^{N_I}\overline{\left(\frac{\partial}{\partial\theta_p}\ln q(x_{p,l}|R_{p,l})\right)^2} \\ &= \delta_{p,p'}\sum_{l=1}^{N_I}\left(\frac{\partial R_{p,l}}{\partial\theta_p}\right)^2\overline{\left(\frac{x_{p,l}}{R_{p,l}}-\frac{1-x_{p,l}}{1-R_{p,l}}\right)^2}\,. \end{aligned} \tag{G.19}$$

In Sect. A.12.1 the expectation values

$$\begin{aligned} \overline{x_{p,l}} &= R_{p,l}\,, \\ \overline{x_{p,l}^2} &= R_{p,l}\,, \\ \overline{x_{p,l}(1-x_{p,l})} &= 0\,, \\ \overline{(1-x_{p,l})^2} &= 1-R_{p,l} \end{aligned} \tag{G.20}$$

are derived. Their use turns $F_{p,p'}$ into

$$\begin{aligned} F_{p,p'} &= \delta_{p,p'}\sum_{l=1}^{N_I}\left(\frac{\partial R_{p,l}}{\partial\theta_p}\right)^2\frac{1}{R_{p,l}(1-R_{p,l})} \\ &= 4\delta_{p,p'}\,N_I\,; \\ p,\,p' &= 1,\ldots,N_P\,. \end{aligned} \tag{G.21}$$

Following the analogous way one obtains for the lower right block b of F the result

$$F_{N_P+i,N_P+i'} = 4\delta_{i,i'}\,N_P\,, \qquad i,i' = 1,\ldots,(N_I-1)\,. \tag{G.22}$$

The elements of the upper right block s of F are

$$\begin{aligned} F_{p,N_P+i} &= \sum_{k=1}^{N_P}\sum_{l=1}^{N_I}\overline{\left(\frac{\partial}{\partial\theta_p}\ln q(x_{k,l}|R_{k,l})\right)\frac{\partial}{\partial\sigma_i}\ln q(x_{k,l}|R_{k,l})}\,, \\ p &= 1,\ldots,N_P\,, \\ i &= 1,\ldots,(N_I-1)\,. \end{aligned} \tag{G.23}$$

Here, the argument that led from Eqs. (G.15) to (G.18) has been used. Because $R_{k,l}$ depends on θ_k and σ_l, and only on these parameters, the last equation simplifies to

$$\begin{aligned}
F_{p,N_P+i} &= \overline{\left(\frac{\partial}{\partial\theta_p}\ln q(x_{p,i}|R_{p,i})\right)\frac{\partial}{\partial\sigma_i}\ln q(x_{p,i}|R_{p,i})} \\
&= \frac{\partial R_{p,i}}{\partial\theta_p}\frac{\partial R_{p,i}}{\partial\sigma_i}\overline{\left(\frac{x_{p,i}}{R_{p,i}}-\frac{1-x_{p,i}}{1-R_{p,i}}\right)^2}\,.
\end{aligned} \tag{G.24}$$

By way of Eq. (G.20) this becomes

$$\begin{aligned}
F_{p,N_P+i} &= \frac{\partial R_{p,i}}{\partial\theta_p}\frac{\partial R_{p,i}}{\partial\sigma_i}\frac{1}{R_{p,i}(1-R_{p,i})} \\
&= -4\,, \\
p &= 1,\ldots,N_P\,, \\
i &= 1,\ldots,N_I\,.
\end{aligned} \tag{G.25}$$

Thus the elements of the Fisher matrix are as given by Eqs. (12.15)–(12.19).

G.3 On the Inverse of the Fisher Matrix of the Trigonometric Model

Aiming at the Gaussian approximation (12.28) to the posterior of the trigonometric model, we look for the inverse $F^{-1} = C$ of the Fisher matrix F. The latter one is given by Eqs. (12.15–12.19). This inverse equals the correlation matrix C of the desired Gaussian. Actually, we restrict ourselves to calculate the diagonal elements of C. Via Eq. (12.29) they give access to the Gaussian errors of the parameters. Equation (12.30) shows that the determinants of the matrices $F^{(k,k)}$ and F must be calculated; the determinant of F is already known by Eq. (12.26). So the present section is devoted to the determinant of $F^{(k,k)}$.

The matrix $F^{(k,k)}$ is obtained by omitting the kth row and the kth column from F. This produces a matrix of the same structure as F. Let the index $p = 1,\ldots,N_P$ refer to the competence parameters of the persons. Then one has

$$F^{(p,p)} = 4\left(\begin{array}{cccccc}
 & & & -1, & \ldots, & -1 \\
 & N_I\mathbf{1}_{N_P-1} & & \vdots & & \vdots \\
 & & & -1, & \ldots, & -1 \\
-1, & \ldots, & -1, & & & \\
\vdots & & \vdots & & N_P\mathbf{1}_{N_I-1} & \\
-1 & \ldots, & -1, & & &
\end{array}\right)\,. \tag{G.26}$$

The essential difference to the matrix (12.15) occurs in the upper left block: It is proportional to the unit matrix in only $N_P - 1$ dimensions. To calculate the determinant of (G.26) one uses the same procedure as described in Sect. G.2 for the determinant of F and one finds

$$\det\left(F^{(p,p)}\right) = 4^{N_P+N_I-2}\, N_I^{N_P-2}\, N_P^{N_I-2}\, (N_I + N_P - 1)\,. \tag{G.27}$$

According to Eq. (12.30) this leads to the result (12.32).

The errors of the item parameters $i = 1, \ldots, N_I$ are given by the diagonal elements C_{N_P+i,N_P+i} of the correlation matrix. The matrix $F^{(N_P+i,N_P+i)}$ has the structure

$$F^{(N_P+i,N_P+i)} = 4 \left(\begin{array}{cccccc} & & & -1, & \ldots, & -1 \\ & N_I\mathbf{1}_{N_P} & & \vdots & & \vdots \\ & & & -1, & \ldots, & -1 \\ -1, & \ldots, & -1, & & & \\ \vdots & & \vdots & & N_P\mathbf{1}_{N_I-2} & \\ -1 & \ldots, & -1, & & & \end{array} \right)\,. \tag{G.28}$$

The essential difference to the matrix (12.15) is seen in the lower right block which is proportional to the unit matrix in only $N_I - 2$ dimensions. The determinant of (G.28) is

$$\det\left(F^{(N_P+i,N_P+i)}\right) = 2 \times 4^{N_P+N_I-2}\, N_P^{N_I-2} N_I^{N_P-1}\,. \tag{G.29}$$

Via Eq. (12.30) one obtains the result given in Eq. (12.36).

Appendix H
On the Art of Fitting

H.1 The Geometric Measure on the Scale of a Chi-Squared Distribution

According to the second line of Eq. (9.18) the geometric measure on the scale of η is

$$\mu_g(\eta) = \frac{1}{2}\left[F(\eta)\right]^{1/2}, \tag{H.1}$$

where F is the Fisher function given in Eq. (9.2), which means

$$F(\eta) = -\int \mathrm{d}y\, \tilde{\chi}_N^{\mathrm{sq}}(y|\eta)\, \frac{\partial^2}{\partial\eta^2} \ln \tilde{\chi}_N^{\mathrm{sq}}(y|\eta)\,. \tag{H.2}$$

Here, $\tilde{\chi}_N^{\mathrm{sq}}(y|\eta)$ is the model (13.7). This yields

$$\begin{aligned} F(\eta) &= -\frac{1}{\Gamma(N/2)}\int \mathrm{d}y\, \exp\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right) \frac{\partial^2}{\partial\eta^2}\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right) \\ &= \frac{1}{\Gamma(N/2)}\int \mathrm{d}y\, \exp\left(\frac{N}{2}[y-\eta] - e^{y-\eta}\right) e^{y-\eta} \\ &= \frac{1}{\Gamma(N/2)}\int_{-\infty}^{\infty} \mathrm{d}y\, \exp\left((\frac{N}{2}+1)[y-\eta] - e^{y-\eta}\right) \\ &= \frac{\Gamma(N/2+1)}{\Gamma(N/2)} \\ &= \frac{N}{2}\,. \end{aligned} \tag{H.3}$$

For the step from the third to the fourth line of this equation, one uses the fact that $\tilde{\chi}_N^{\mathrm{sq}}$ is normalised to unity. The last line follows from Eq. (B.24) in Appendix B. By

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1

Eq. (H.1) one obtains

$$\mu_g(\eta) = \left(\frac{N}{8}\right)^{1/2} . \tag{H.4}$$

H.2 Convoluting Chi-Squared Distributions

Let a set of positive numbers t_k , $k = 1, \ldots, N$, be given so that each one follows a chi-squared distribution (4.34). The distribution of t_k,

$$q_k(t_k) = \frac{1}{\Gamma(f_k/2)} t_k^{f_k/2-1} \exp(-t_k) , \tag{H.5}$$

shall have the number f_k of degrees of freedom. The f_k are positive; they need not be integer. For $k \neq k'$ the number f_k may be different from $f_{k'}$. We show that the quantity

$$T = \sum_{k=1}^{N} t_k \tag{H.6}$$

follows a chi-squared distribution with

$$f^{\text{tot}} = \sum_{k=1}^{N} f_k \tag{H.7}$$

degrees of freedom. This is a consequence of the convolution theorem. We explain the notion of "convolution" and state the theorem. It describes the structure of the Fourier[2] transform of a convolution. From this follows the distribution of T; see Sect. H.4. The Fourier transformation is defined in Sect. H.3.

A convolution $q_1 \circ q_2$ of the functions q_1 , q_2, defined and integrable over the real axis, is

$$\begin{aligned} q_1 \circ q_2\,(x) &= \int \mathrm{d}t_2\, q_1(x - t_2) q_2(t_2) \\ &= \int \mathrm{d}t_1 \mathrm{d}t_2\, \delta(x - t_1 - t_2) q_1(t_1) q_2(t_2) , \end{aligned} \tag{H.8}$$

where $\delta(x)$ is Dirac's δ distribution. Integrating $q_1 \circ q_2$ over x from 0 to ∞ yields unity because the distributions q_k are normalised to unity. The N-fold convolution is

[2] Joseph Fourier, 1768–1830, French mathematician and physicist, member of the Académie des Sciences. He studied the transport of heat in solids. In this context he discovered the possibility to expand distributions into the series which nowadays carries his name.

$$q_1 \circ q_2 \circ \cdots \circ q_N\,(T) = \int \mathrm{d}t_1 \ldots \mathrm{d}t_N\, \delta(T - \sum_{k=1}^{N} t_k) \prod_{k=1}^{N} q_k(t_k)\,. \tag{H.9}$$

This distribution is again normalised to unity.

The Fourier transform of q_k is called

$$F_k(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}t_k\, q_k(t_k) e^{i\xi t_k}\,. \tag{H.10}$$

Here, we have set $q_k(t_k) = 0$ for negative values of t_k in order formally to obtain the integration from $-\infty$ to ∞ required by the definition of the Fourier transform. For all further transforms in the present section this is also done. Note that

$$F_k(0) = \frac{1}{\sqrt{2\pi}} \tag{H.11}$$

because the distribution q_k is normalised to unity.

Let $F_{1\circ 2}$ be the Fourier transform of $q_1 \circ q_2$. Then the convolution theorem says that

$$F_{1\circ 2}(\xi) \propto F_1(\xi) F_2(\xi)\,, \tag{H.12}$$

that is, the Fourier transform of the convolution $q_1 \circ q_2$ is proportional to the product of the transforms of q_1 and q_2. Again the proportionality constant is such that

$$F_{1\circ 2}(0) = \frac{1}{\sqrt{2\pi}} \tag{H.13}$$

because $q_1 \circ q_2$ is normalised to unity.

This can be generalised to the statement: the Fourier transform $F_{1\circ\cdots\circ N}(\xi)$ of the N-fold convolution (H.9) is proportional to the product of the N Fourier transforms F_k ; that is,

$$F_{1\circ\cdots\circ N}(\xi) \propto \prod_{k=1}^{N} F_k(\xi)\,. \tag{H.14}$$

The proportionality constant is again such that

$$F_{1\circ\cdots\circ N}(0) = \frac{1}{\sqrt{2\pi}}\,. \tag{H.15}$$

Equation (H.32) yields the Fourier transform

$$F_k(\xi) = \frac{1}{\sqrt{2\pi}} \frac{1}{(1 - i\xi)^{f_k/2}} \tag{H.16}$$

of the distribution q_k in Eq. (H.5). By Eq. (H.12) the Fourier transform of $q_1 \circ q_2$ is

$$F_{1\circ 2}(\xi) = \frac{1}{\sqrt{2\pi}} \frac{1}{(1 - i\xi)^{(f_1+f_2)/2}} \,. \tag{H.17}$$

Generally, the Fourier transform of the N-fold convolution $q_1 \circ q_2 \circ \cdots \circ q_N$ is

$$F_{1\circ 2\circ \ldots N}(\xi) = \frac{1}{\sqrt{2\pi}} \frac{1}{(1 - i\xi)^{f^{\text{tot}}/2}} \,, \tag{H.18}$$

where

$$f^{\text{tot}} = \sum_{k=1}^{N} f_k \,. \tag{H.19}$$

Inverting the Fourier transformation that has given (H.18), one finds

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}\xi \, F_{1\circ 2 \cdots \circ N}(\xi) e^{-iT\xi} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}\xi \, \frac{1}{(1 - i\xi)^{f^{\text{tot}}/2}} \, e^{-iT\xi} \\ &= \begin{cases} \frac{1}{\Gamma(f^{\text{tot}}/2)} T^{f^{\text{tot}}/2-1} \, e^{-T} \,, \\ 0 \quad \text{for } T < 0 \,, \end{cases} \end{aligned} \tag{H.20}$$

See Eqs. (H.33) and (H.30). This shows that the quantity T of Eq. (H.6) follows a chi-squared distribution with f^{tot} degrees of freedom which was to be shown.

H.3 Definitions of Fourier Transforms

Let $f(x)$ be a real function which can be integrated over the whole real axis; that is, the integral

$$\int_{-\infty}^{\infty} \mathrm{d}x \; f(x)$$

exists. We do not require the function $f(x)$ to be regular at all x. The Fourier transform

$$F(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x \; f(x) e^{i\xi x} \tag{H.21}$$

exists for all real ξ. This definition of F follows Sect. 17.31 of [1].

The Fourier transform is an expansion of f in terms of the orthogonal functions

$$\frac{1}{\sqrt{2\pi}} \, e^{i\xi x} .$$

They are orthogonal in the sense that

$$\frac{1}{2\pi}\int \mathrm{d}x\, e^{i(\xi-\xi')x} \;=\; \delta(\xi-\xi')\,. \tag{H.22}$$

The inversion of the Fourier transform is given by

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}\xi\, F(\xi) e^{-i\xi x}\,. \tag{H.23}$$

The so-called Fourier sine and Fourier cosine transforms are

$$F_s(\xi) = \left(\frac{2}{\pi}\right)^{1/2}\int_0^{\infty} \mathrm{d}x\, f(x)\sin(\xi x) \tag{H.24}$$

and

$$F_c(\xi) = \left(\frac{2}{\pi}\right)^{1/2}\int_0^{\infty} \mathrm{d}x\, f(x)\cos(\xi x)\,, \tag{H.25}$$

See Sect. 17.31 of [1]. The symmetric and antisymmetric parts of $f(x)$ - in the sense of a reflection at the origin - are picked up by the cosine and sine transformations. The symmetric part is

$$f^S(x) = \frac{1}{2}\Big(f(x) + f(-x)\Big) \tag{H.26}$$

whereas

$$f^A(x) = \frac{1}{2}\Big(f(x) - f(-x)\Big) \tag{H.27}$$

is the antisymmetric part. This leads to

$$\begin{aligned} F(\xi) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}x\, \Big[f^S(x) + f^A(x)\Big]\Big[\cos(\xi x) + i\sin(\xi x)\Big] \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}x\, \Big[f^S(x)\cos(\xi x) + i f^A(x)\sin(\xi x)\Big] \end{aligned} \tag{H.28}$$

because the integrals over products of a symmetric function with an antisymmetric one vanish. It follows

$$F(\xi) = \left(\frac{2}{\pi}\right)^{1/2}\int_0^{\infty} \mathrm{d}x\, \Big[f^S(x)\cos(\xi x) + i f^A(x)\sin(\xi x)\Big]\,. \tag{H.29}$$

H.4 The Fourier Transform of the Chi-Squared Distribution

We are interested in the Fourier transform $F(\xi)$ (see H.21) of a chi-squared distribution (see H.5). This means that the function $f(x)$ in (H.21) is

$$f(x) = \begin{cases} \frac{1}{\Gamma(\nu)}\, x^{\nu-1} e^{-x} & \text{for } x > 0\,, \\ 0 \quad \text{for } x < 0\,. \end{cases} \tag{H.30}$$

Then f^S and f^A in Eqs. (H.25, H.26) both equal $f(x)/2$ for $x > 0$. In this case Eq. (H.28) yields

$$F(\xi) = \frac{1}{2}\Big[F_c(\xi) + i\,F_s(\xi)\Big]\,. \tag{H.31}$$

According to entry 16 in Sect. 17.33 as well as entry 7 in Sect. 17.34 of [1] the Fourier transform (H.31) is

$$\begin{aligned} F(\xi) &= \frac{1}{\sqrt{2\pi}} \frac{1}{(1+\xi^2)^{\nu/2}}\Big[\cos(\nu \tan^{-1}\xi) + i\,\sin(\nu \tan^{-1}\xi)\Big] \\ &= \frac{1}{\sqrt{2\pi}}(1+\xi^2)^{-\nu/2}\exp(i\nu \tan^{-1}\xi) \\ &= \frac{1}{\sqrt{2\pi}}\Big[\sqrt{1+\xi^2}\exp(-i\nu \tan^{-1}\xi)\Big]^{-\nu} \\ &= \frac{1}{\sqrt{2\pi}}\,\frac{1}{(1-i\xi)^\nu}\,. \end{aligned} \tag{H.32}$$

The inversion (H.23) of this Fourier transformation yields the function in Eq. (H.30),

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\xi\, \frac{e^{-i\xi x}}{(1-i\xi)^\nu} = f(x)\,, \tag{H.33}$$

from which we started the transformations.

Reference

1. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products* (Academic Press, New York, 2015)

Index

H.L. Harney, *Bayesian Inference*, DOI 10.1007/978-3-319-41644-1