

# 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 will show 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, expressed as a probability. Such a conclusion typically has the form “The quantity  $\xi$  inferred from the present experiment has the value  $\alpha \pm \sigma$ .” An experiment never yields the true value of  $\xi$ . Rather, the result is characterised by an interval in which the true value should lie. The true value does not even lie with certainty in that interval. A more precise interpretation of the above interval is “The quantity  $\xi$  is, with a probability  $K = 0.68$ , in the interval  $[\alpha - \sigma, \alpha + \sigma]$ .” Trying to be even more precise, one would say, “We assign a Gaussian distribution to the parameter  $\xi$ . The distribution is centred at  $\alpha$  and has a standard deviation  $\sigma$ . The shortest interval containing  $\xi$  with a 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 length of the error interval because one is free to choose the probability  $K$ .

The present book deals with the generalisation of 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 of 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?

## 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 width of the line that one observes with a spectrograph. The fluctuations have different causes in these two cases but they cannot be avoided. One does not observe the interesting parameter  $\xi$ . Rather, one observes events  $x$  that have a distribution depending on  $\xi$ . Data analysis means to infer  $\xi$  from the event  $x$  – usually on the basis of a distribution  $p$  that depends parametrically on  $\xi$ . This parameter is also called the hypothesis that conditions the distribution of  $x$ . The connection between  $x$  and  $\xi$  is given by  $p(x|\xi)$ : expressed in words, “the distribution  $p$  of  $x$ , given  $\xi$ .” This distribution must depend on  $\xi$  in such a way that different hypotheses entail different distributions of  $x$ , so that one can learn from  $x$  about  $\xi$ .

Inferring  $\xi$  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 is based on one event. Note that even an experiment that produces a huge amount of data yields one event, in the sense that it does not yield all possible data, and its repetition would produce a different event. For this reason, no experiment yields the true value of  $\xi$ , and inference of  $\xi$  is achieved by assigning a distribution to  $\xi$ , although one assumes that all the events are in fact conditioned by one and the same true value of  $\xi$ . Thus the distribution  $P(\xi|x)$  assigned to  $\xi$  is a representation of the limited knowledge about  $\xi$ .

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 a statement? We take it for granted here that the answer is “yes”, and consider the debate as historical.

For the founders of statistical inference, Bayes<sup>1</sup> [9] and Laplace<sup>2</sup> [38, 40], the notion of probability has carried both concepts: the probability attached to a statement<sup>3</sup>  $\xi$  can mean the relative frequency of its occurrence or the state of knowledge about  $\xi$ .

This “or” is not exclusive; it is not an “either or”. It allows statements  $\xi$  that cannot be subjected to a “quality test” that reveals 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 will rain 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 latter statement can

<sup>1</sup> 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.

<sup>2</sup> 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 the spherical harmonics. He formulated and applied Bayes’ theorem independently of him.

<sup>3</sup> We do not distinguish between the quantity  $\xi$  and the statement that the hypothesis has the value  $\xi$ .

only be interpreted as evaluating the available knowledge. A fine description of these different interpretations has been given by Cox [29, 30]; see also Chaps. 13 and 14 of Howson and Urbach [72]. A taste of the above-mentioned debate is given by the polemics in [80].

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

The above distinction is only an apparent one, because the interpretation of probability as a value of the available knowledge is always possible, and is thus the broader interpretation. 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 at in question. However, there could be a statistical quality test which would turn this personal belief into a known frequency of breakdown. But even if the praise by the salesman is objective in this sense, it becomes a personal belief for the interested client in the quality of his/her car – the one car that he/she decides to buy.

Let us try to translate this into the language of measurement. The quantity  $\xi$  has been measured as  $\alpha \pm \sigma$ . Hence, with 68% probability, it is in that interval. Setting aside systematic errors, one can offer a frequency interpretation of this statement: if one were to repeat the measurement, say 100 times, the result would fall into the interval with a frequency of 68%. This is right but does not describe the situation well. If one actually had 100 more measurements, one could 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 [126, 127].

There does not seem to be a clear distinction between those cases which allow the frequency interpretation of probability and those 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 keep in mind that it is the frequency interpretation that leads to mathematically formulated distributions. Some of these distributions 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 a 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 in the graphical representation of a result or its use in related experiments; see Sect. 4.2.1 of “The review of particle physics” [58].

## 1.2 Logic

Since probability can be interpreted as the value of the available knowledge, it can also be considered as the implementation of non-Aristotelian logic into scientific communication [118]. Jaynes has simply termed it “the logic of Science” [82, 76]. It also serves everyday communication, as certain weather forecasts show. In philosophy, it is called the logic of temporal statements [144]: “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  $\xi$  can be considered the value of truth assigned to  $\xi$ . 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 such as “from  $\xi$  follows  $x$ ”. This value of truth is the conditional probability  $p(x|\xi)$ , i.e. the probability of finding  $x$  when  $\xi$  is given. Cox [29] showed in 1946 that consistency between non-Aristotelian and mathematical logic requires the following two rules.

- (i) Let  $\alpha$ ,  $\xi$ ,  $x$  be three statements that possibly imply each other. The values of truth  $\mu$ ,  $p$ , and  $w$  of the implications “ $\xi$  follows from  $\alpha$ ”, “ $x$  follows from  $\xi \wedge \alpha$ ”, and “ $x \wedge \xi$  follows from  $\alpha$ ”, respectively, must be defined such that the product rule

$$w(x \wedge \xi|\alpha) = p(x|\xi \wedge \alpha) \mu(\xi|\alpha) \quad (1.1)$$

holds. Here, the operator  $\wedge$  means the logical “and”. Cox’s result seems obvious to every person who has only a little experience with probabilities. It takes an effort to realise that it is not trivial. Note that probabilistic values of truth need not be positive numbers. There are realms of physics where the positive numbers  $p$  are replaced by complex probability amplitudes  $a$ . Hence, values of truth can be more complicated objects than positive numbers. In any case, however, they must respect the relation (1.1). The symmetry principle introduced in Chap. 6 indeed leads us to consider probability amplitudes in Chap. 8. Although we restrict ourselves

to real  $a$ , probability amplitudes allow richer logical combinations than do the positive numbers  $p$ . Amplitudes can be positive or negative. Thus the probability attached to “ $\xi_1$  or  $\xi_2$ ” can be smaller than the probabilities attached to either one of the statements. One summarises this phenomenon by saying that the statements can interfere, or that the alternative “ $\xi_1$  or  $\xi_2$ ” is coherent. We shall encounter this in Chap. 15.

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

$$\int d\xi P(\xi|x) = 1. \quad (1.2)$$

Here, the integral without indication of the limits of integration extends over the entire domain of definition of  $\xi$ . This rule is necessary in order to assign a probability to a negation. The probability of the assertion “ $\xi$  is not in the interval  $[\xi_<, \xi_>]$ ” is the integral over the complement of the interval  $[\xi_<, \xi_>]$ . The integral over the complement exists, since  $P$  is required to be proper. The assignment of unit probability to the statement that  $\xi$  is somewhere in its domain of definition is a convention. Not only the posterior  $P$  but also all 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  $\xi$  would not be clearly defined. One could multiply it by any non-negative function of  $\xi$  without changing the distribution of  $x$ . Hence, inferring  $\xi$  from the event  $x$  is possible only if (1.2) holds. Nevertheless, in the present book, distributions will be admitted that cannot be normalised – provided that they do not depend on a parameter to be inferred. Such distributions are called improper. One cannot assign a value of truth to a negation that involves 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 Chap. 2.

The joint distribution of the multiple event  $x_1 \wedge x_2 \wedge \dots \wedge x_N$  will be 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 logical operator  $\wedge$  is not written down explicitly in what follows. Instead, the multiple event is denoted by  $x_1, \dots, x_N$  or simply  $x = (x_1, \dots, x_N)$ .

### A.1.1

An immediate consequence of the rules (1.1) and (1.2) is Bayes’ theorem, discussed in Chap. 2. This theorem specifies the posterior probability  $P$  of  $x$ , given  $\xi$ . By the same token, the error interval of  $\xi$  is given. This interval is the smallest interval in which  $\xi$  lies with probability  $K$ . We call it the Bayesian interval  $\mathcal{B}(K)$ . To find the smallest interval, one needs a measure in the space of  $\xi$ . The measure is identified with the prior distribution appearing in Bayes’ theorem.

### 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  $\xi$ , and is called the a priori distribution. All methods of inference described in the present book rely on Bayes' theorem and a unique definition of  $\mu$ .

The definition starts from a symmetry principle. In Chaps. 6, 8, and 11, models  $p(x|\xi)$  are considered that connect the parameter  $\xi$  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 it in Chap. 6, is the prior distribution  $\mu$ . This procedure is inspired by the ideas of Hartigan [66], Stein [136], and Jaynes [75].

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 [83]. 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 [124, 125, 96, 36, 37]. However, we do not go so far as to leave the prior distribution at the disposal of the person or the community analysing given data. This is done in Chap. 14 of Howson and Urbach [72] and in the work by D'Agostini [32, 31, 33]. 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 one to devise a scheme where one parameter  $\xi_1$  is inferred independently of the other parameters  $\xi_2, \dots, \xi_N$  in the sense that  $\xi_1$  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  $\xi_1$  is often linked to and dependent on other parameters that must be included in the model even though they are not interesting. The intensity of a signal depends on the determination of the background, although the interest is focused on the signal.

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 [12], and [120]. Chapter 11 of the present book shows that there are more form-invariant distributions than was 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  $p(x|\xi)$  – i.e. the amplitude  $a_x$  – as a component of a vector that depends parametrically on  $\xi$ . This is the parametric representation of a surface. The measure on the surface is the prior distribution. To understand this, one needs some differential geometry [73, 121, 122], which is explained

in Chap. 9. The differential geometric measure is again given by Jeffreys' rule [83].

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 roots of the probabilities. This, however, becomes obvious from the form-invariant models.

Beyond the observed event  $x$ , information on  $\xi$  is often available that should be incorporated into Bayesian inference and will let one shrink the Bayesian interval. The order of magnitude of  $\xi$  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  $\mu$  to an informed prior  $\mu^{\text{inf}}$ . An informed prior may simply be the posterior of a preceding experiment. It may also be generated by entropy maximisation, given previous information. Jaynes [77, 78] has transferred this method from thermodynamics to the analysis of data. This idea has found much interest and has led to a series of conferences [133, 131, 44, 43, 129, 53, 160, 132, 104, 69, 130, 63, 150] and many publications [27]. 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  $\mu$ . According to Jaynes [75], the method uses  $\mu$ .

## 1.4 Decisions

Bayesian inference chooses from the family of distributions  $p(x|\xi)$  the ones that best reproduce 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 distributions consistent with the available data?

When  $x$  follows a Gaussian distribution, this question is decided by the chi-squared criterion described in Chap. 14. In Chap. 16, generalisations to the histogram and the multinomial model are given. It turns out that to make the decision, one needs a measure in the space of  $\xi$ . We have identified this measure with the prior distribution  $\mu$ .

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 a 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 Chaps. 14 and 16.



<http://www.springer.com/978-3-540-00397-7>

Bayesian Inference

Parameter Estimation and Decisions

Harney, H.L.

2003, XIII, 263 p., Hardcover

ISBN: 978-3-540-00397-7