# “Likelihood”

*A.W.F. Edwards*
Gonville and Caius College, Cambridge, U.K.

CIMAT, Guanajuato, Mexico, March 1999

*This article is a preliminary version of what will be published in the International Encyclopedia of the Social and Behavioral Sciences, and was written while Professor Edwards was visiting CIMAT, March 8-19, 1999.

A statistical model for phenomena in the sciences or social sciences is a mathematical construct which associates a probability with each of the possible outcomes. If the data are discrete, such as the numbers of people falling into various classes, the model will be a discrete probability distribution, but if the data consist of measurements or other numbers which may take any values in a continuum, the model will be a continuous probability distribution. When two different models, or perhaps two variants of the same model differing only in the value of some adjustable parameter(s), are to be compared as explanations for the same observed outcome, the probability of obtaining this particular outcome can be calculated for each and is then known as likelihood for the model or parameter value(s) given the data.

Probabilities and likelihoods are easily (and frequently) confused, and it is for this reason that in 1921 R.A.Fisher introduced the new word:

*“What we can find from a sample is the likelihood of any particular*

> *value of [the parameter], if we define the likelihood as a quantity proportional to the probability that, from a population having that particular value, the [observed sample] should be obtained. So defined, probability and likelihood are quantities of an entirely different nature."*

The first difference to be noted is that the variable quantity in a likelihood statement is the hypothesis (a word which conveniently covers both the case of a model and of particular parameter values in a single model), the outcome being that actually observed, in contrast to a probability statement, which refers to a variety of outcomes, the hypothesis being assumed and fixed. Thus a manufacturer of dice may reasonably assert that the outcomes 1, 2, 3, 4, 5, 6 of a throw each have probability 1/6 on the hypothesis that his dice are well-balanced, whilst an inspector of casinos charged with testing a particular die will wish to compute the likelihoods for various hypotheses about these probabilities on the basis of data from actual tosses.

The second difference arises directly from the first. If all the outcomes of a statistical model are considered their total probability will be 1 since one of them must occur and they are mutually exclusive; but since in general hypotheses are not exhaustive — one can usually think of another one — it is not to be expected that the sum of likelihoods has any particular meaning, and indeed there is no addition law for likelihoods corresponding to the addition law for probabilities. It follows that only relative likelihoods are informative, which is the reason for Fisher's use of the word "proportional" in his original definition.

The most important application of likelihood is in parametric statistical models. Consider the simplest binomial example, such as that of the distribution of the number of boys $r$ in families of size $n$ (an example which has played an important role in the development of statistical theory since the early eighteenth century). The probability of getting exactly $r$ boys will be given by the binomial distribution indexed by a parameter $p$, the probability of a male birth. Denote this probability of $r$ boys by $P(r|p)$, $n$ being assumed fixed and of no statistical interest. Then we write

$$L(p||r) \propto P(r|p)$$

for the likelihood of $p$ given the particular value $r$, the double vertical line $||$ being used to indicate that the likelihood of $p$ is not conditional on $r$ in the technical probability sense. In this binomial example $L(p||r)$ is a continuous function of the parameter $p$ and is known as the likelihood function. When only two hypotheses are compared, such as two particular values of $p$ in the present example, the ratio of their likelihoods is known as the likelihood ratio.

The value of $p$ which maximises $L(p||r)$ for an observed $r$ is known as the maximum-likelihood estimate of $p$ and is denoted by $\widehat{p}$; expressed in general form as a function of $r$ it is known as the maximum-likelihood estimator. Since the pioneering work of Fisher in the 1920s it has been known that maximum-likelihood estimators possess certain desirable properties under repeated-sampling (consistency and asymptotic efficiency, and in an important class of models sufficiency and full efficiency), and for this reason they have come to occupy a central position in repeated-sampling (or "frequentist") theories of statistical inference.

However, partly as a reaction to the unsatisfactory features which repeated-sampling theories display when used as theories of evidence, coupled with a reluctance to embrace the full-blown Bayesian theory of statistical inference, likelihood is increasingly seen as a fundamental concept enabling hypotheses and parameter values to be compared directly.

The basic notion, championed by Fisher as early as 1912 whilst still an undergraduate at Cambridge but now known to have been occasionally suggested by other writers even earlier, is that the likelihood ratio for two hypotheses or parameter values is to be interpreted as the degree to which the data support the one hypothesis against the other. Thus a likelihood ratio of 1 corresponds to indifference between the hypotheses on the basis of the evidence in the data, whilst the maximum-likelihood value of a parameter is regarded as the best-supported value, other values being ranked by their lesser likelihoods accordingly. This was formalised as the Law of Likelihood by Ian Hacking in 1965. Fisher's final advocacy of the direct use of likelihood will be found in his last book Statistical Methods and Scientific Inference (1956).

Such an approach, unsupported by any appeal to repeated-sampling criteria, is ultimately dependent on the primitive notion that the best hypothesis or parameter-value on the evidence of the data is the one which would explain what has in fact been observed with the highest probability. The strong intuitive appeal of this can be captured by recognizing that it is the value which would lead, on repeated sampling, to a precise repeat of the data with the least expected delay. In this sense it offers the best statistical explanation of the data.

In addition to specifying that relative likelihoods measure degrees of support, the likelihood approach requires us to accept that the likelihood function or ratio contains all the information we can extract from the data about the hypotheses in question on the assumption of the specified statistical model — the so-called Likelihood Principle. It is important to include the qualification requiring the specification of the model, first because the adoption of a different model might prove necessary later and secondly because in some cases the structure of the model enables inferences to be made in terms of fiducial probability which, though dependent on the likelihood, are stronger, possessing repeated-sampling properties which enable confidence intervals to be constructed.

Though it would be odd to accept the Law of Likelihood and not the Likelihood Principle, Bayesians necessarily accept the Principle but not the Law, for although the likelihood is an intrinsic component of Bayes's Theorem, Bayesians deny that a likelihood function or ratio has any meaning in isolation. For those who accept both the Law and the Principle it is convenient to express the two together as:

The Likelihood Axiom: Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other (Edwards, 1972).

The likelihood approach has many advantages apart from its intuitive appeal. It is straightforward to apply because the likelihood function is usually simple to obtain analytically or easy to compute and display. It leads directly to the important theoretical concept of sufficiency according to which the function of the data which is the argument of the likelihood function itself carries the information. This reduction of the data is often a simple statistic such as the sample mean. Moreover, the approach illuminates many of the controversies surrounding repeated-sampling theories of inference, especially those concerned with ancillarity and conditioning. Birnbaum (1962) argued that it was possible to derive the Likelihood Principle from the concepts of sufficiency and conditionality, but to most people the Principle itself seems the more primitive concept and the fact that it leads to notions of sufficiency and conditioning seems an added reason for accepting it.

Likelihoods are multiplicative over independent data sets referring to the same hypotheses or parameters, facilitating the combination of information. For this reason log-likelihood is often preferred because information is then combined by addition. In the field of genetics, where likelihood theory is widely applied, the log-likelihood with the logarithms taken to the base 10 is known as a LOD, but for general use natural logarithms to the base $\mathrm{e}$ are to be preferred, in which case log-likelihood is sometimes called support. Most importantly, the likelihood approach is compatible with Bayesian statistical inference in the sense that the posterior Bayes distribution for a parameter is, by Bayes's Theorem, found by multiplying the prior distribution by the likelihood function. Thus when, in accordance with Bayesian principles, a parameter can itself be given a probability distribution (and this assumption is the Achilles' heel of Bayesian inference) all the information the data contain about the parameter is transmitted via the likelihood function in accordance with the Likelihood Principle. It is indeed difficult to see why the medium through which such information is conveyed should depend on the purely external question of whether the parameter may be considered to have a probability distribution, and this is another powerful argument in favour of the Principle itself.

In the case of a single parameter the likelihood function or the log-likelihood function may easily be drawn, and if it is unimodal limits may be assigned to the parameter, analogous to the confidence limits of repeated-sampling theory. Calling the log-likelihood the support, $m$-unit support limits are the two parameter values astride the maximum at which the support is $m$ units less than at the maximum. For the simplest case of estimating the mean of a normal distribution of known variance the 2-unit support limits correspond closely to the $95\%$ confidence limits which are at $\pm 1.96$ standard errors. In this normal case the support function is quadratic and may therefore be characterized completely by the maximum-likelihood estimate and the curvature at the maximum (the reciprocal of the radius of curvature) which is defined as the observed information. Comparable definitions apply in multiparameter cases, leading to the concept of an $m$-unit support region and an observed information matrix. In cases in which the support function is not even approximately quadratic the above approach may still be applied if a suitable transformation of the parameter space can be found.

The representation of a support function for more than two parameters naturally encounters the usual difficulties associated with the visualisation of high-dimensioned spaces, and a variety of methods have been suggested to circumvent the problem. It will often be the case that information is sought about some subset of the parameters, the others being considered to be nuisance parameters of no particular interest. In fortunate cases it may be possible to restructure the model so that the nuisance parameters are eliminated, and in all cases in which the support function is quadratic (or approximately so) the dimensions corresponding to the nuisance parameters can simply be ignored.

Several other approaches are in use to eliminate nuisance parameters. Marginal likelihoods rely on finding some function of the data which does not depend on them; notable examples involve the normal distribution, where a marginal likelihood for the variance can be found from the distribution of the sample variance which is independent of the mean, and a marginal likelihood for the mean can similarly be found using the t-distribution. Profile likelihoods, also called maximum relative likelihoods, are found by replacing the nuisance parameters by their maximum-likelihood estimates at each value of the parameters of interest. It is easy to visualise from the case of two parameters why this is called a profile likelihood.

Naturally, a solution can always be found by strengthening the model through adopting particular values for the nuisance parameters, just as a Bayesian solution using integrated likelihoods can always be found by adopting a prior distribution for them and integrating them out, but such assumptions do not command wide assent. When a marginal likelihood solution has been found it may correspond to a Bayesian integrated likelihood for some choice of prior, and such priors are called neutral priors to distinguish them from so-called uninformative priors for which no comparable justification exists. However, in the last analysis there is no logical reason why nuisance parameters should be other than a nuisance, and procedures for mitigating the nuisance must be regarded as expedients.

All the common repeated-sampling tests of significance have their analogues in likelihood theory, and in the case of the normal model it may seem that only the terminology has changed. At first sight an exception seems to be the $\chi^2$ goodness-of-fit test, where no alternative hypothesis is implied. However, this is deceptive, and a careful analysis shows that there is an implied alternative hypothesis which allows the variances of the underlying normal model to depart from their multinomial values. In this way the paradox of small values of $\chi^2$ being interpreted as meaning that the model is “too good” is exposed, for in reality they mean that the model is not good enough and that one with a more appropriate variance structure will have a higher likelihood. Likelihood ratio tests are based on the distribution under repeated-sampling of the likelihood ratio and are therefore not part of likelihood theory.

When likelihood arguments are applied to models with continuous sample spaces it may be necessary to take into account the approximation involved in representing data, which are necessarily discrete, by a continuous model. Neglect of this can lead to the existence of singularities in the likelihood function or other artifacts which a more careful analysis will obviate.

It is often argued that in comparing two models by means of a likelihood ratio, allowance should be made for any difference in the number of parameters by establishing a "rate of exchange" between an additional parameter and the increase in log-likelihood expected. The attractive phrase "Occam's bonus" has been suggested for such an allowance (J.H.Edwards, 1969). However, the proposal seems only to have a place in a repeated-sampling view of statistical inference, where a bonus such as that suggested by Akaike's information criterion is sometimes canvassed.

The major application of likelihood theory so far has been in human genetics, where log-likelihood functions are regularly drawn for recombination fractions (linkage values) (see Ott, 1991), but even there a reluctance to abandon significance-testing altogether has led to a mixed approach. Other examples, especially from medical fields, will be found in the books cited below.

Although historically the development of a likelihood approach to statistical inference was almost entirely due to R.A.Fisher, it is interesting to recall that the Neyman-Pearson approach to hypothesis testing derives ultimately from a remark of "Student's" (W.S.Gossett) in a letter to E.S.Pearson in 1926 that

> *"if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability . . . you will be very much more inclined to consider that the original hypothesis is not true",*

a direct likelihood statement (quoted in McMullen and Pearson, 1939). Indeed, it has been remarked that

> *"Just as support [log-likelihood] is Bayesian inference without the priors, so it turns out to be Neyman-Pearson inference without the 'errors' "*

(Edwards, 1972).

The literature on likelihood is gradually growing as an increasing number of statisticians become concerned at the inappropriate use of significance levels, confidence intervals and other repeated-sampling criteria to represent evidence. The movement is most advanced in biostatistics as may be seen from books such as Clayton and Hills (1993) and Royall (1997), but general texts such as Lindsey (1995) exist as well. Amongst older books Cox and Hinkley (1974) contains much that is relevant to likelihood, whilst Edwards (1972, 1992) was the first book to advocate a purely likelihood approach, and is rich in relevant quotations from Fisher's writings. The history of likelihood is treated by Edwards (1974; reprinted in Edwards 1992).

---

## Bibliography

Birnbaum, A. (1962) On the foundations of statistical inference. *J. Amer. Statist. Ass.* 57, 269-326.

Clayton, D.G. and Hills, M. (1993) *Statistical Models in Epidemiology*. Oxford University Press.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. London: Chapman & Hall.

Edwards, A.W.F. (1972) *Likelihood.* Cambridge University Press.

Edwards, A.W.F. (1974) The history of likelihood. *Int. Statist. Rev.* 42, 9-15.

Edwards, A.W.F. (1992) *Likelihood.* Baltimore: Johns Hopkins University Press.

Edwards, J.H. (1969) In: Computer Applications in Genetics, ed. N.E.Morton. Honolulu: University of Hawaii Press.

Fisher, R.A. (1912) On an absolute criterion for fitting frequency curves. *Mess. Math.* 41, 155-60.

Fisher, R.A. (1921) On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* 1 pt 4, 3-32.

Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd.

Hacking, I. (1965) *Logic of Statistical Inference*. Cambridge University Press.

Lindsey, J.K. (1995) *Introductory Statistics: A Modelling Approach*. Oxford: Clarendon Press.

McMullen, L. and Pearson, E.S. (1939) William Sealy Gosset, 1876-1937. *Biometrika* 30, 205-50.

Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.