If the force of mortality is constant over a single-year age interval $(x, x+1)$, say, and is estimated by $\hat{\mu}_x$ in this interval, then $\hat{p}_x = e^{-\hat{\mu}_x}$ is an estimator of the single-year survival probability p_x. This allows us to estimate the survival function recursively for all corresponding ages, using $\hat{\ell}(x+1) = \hat{\ell}(x)\hat{p}_x$ for $x = 0, 1, \ldots$, and the rest of the life table computations follow suit. Life table construction consists in the estimation of the parameters and the tabulation of functions like those above from empirical data. The data can be for age at death for individuals, as in the example indicated above, but they can also be observations of duration until recovery from an illness, of intervals between births, of time until breakdown of some piece of machinery, or of any other positive duration variable.

So far we have argued as if the life table is computed for a group of mutually independent individuals who have all been observed in parallel, essentially a cohort that is followed from a significant common starting point (namely from birth in our mortality example) and which is diminished over time due to *decrements* (*attrition*) caused by the risk in question and also subject to reduction due to censoring (withdrawals). The corresponding table is then called a *cohort life table*. It is more common, however, to estimate a $\{p_x\}$ schedule from data collected for the members of a population during a limited time period and to use the mechanics of life-table construction to produce a *period life table* from the p_x values.

Life table techniques are described in detail in most introductory textbooks in actuarial statistics, ►biostatistics, ►demography, and epidemiology. See, e.g., Chiang (1984), Elandt-Johnson and Johnson (1980), Manton and Stallard (1984), Preston et al. (2001). For the history of the topic, consult Seal (1977), Smith and Keyfitz (1977), and Dupâquier (1996).

About the Author

For biography *see* the entry ►Demography.

Cross References

►Demographic Analysis: A Stochastic Approach
►Demography
►Event History Analysis
►Kaplan-Meier Estimator
►Population Projections
►Statistics: An Overview

References and Further Reading

Chiang CL (1984) The life table and its applications. Krieger, Malabar
Dupâquier J (1996) L'invention de la table de mortalité. Presses universitaires de France, Paris
Elandt-Johnson RC, Johnson NL (1980) Survival models and data analysis. Wiley, New York
Forsén L (1979) The efficiency of selected moment methods in Gompertz–Makeham graduation of mortality. Scand Actuarial J 167–178
Manton KG, Stallard E (1984) Recent trends in mortality analysis. Academic Press, Orlando
Preston SH, Heuveline P, Guillot M (2001) Demography. Measuring and modeling populations. Blackwell, Oxford
Seal H (1977) Studies in history of probability and statistics, 35: multiple decrements of competing risks. Biometrica 63(3):429–439
Smith D, Keyfitz N (1977) Mathematical demography: selected papers. Springer, Heidelberg

Likelihood

Nancy Reid
Professor
University of Toronto, Toronto, ON, Canada

Introduction

The likelihood function in a statistical model is proportional to the density function for the random variable to be observed in the model. Most often in applications of likelihood we have a parametric model $f(y;\theta)$, where the parameter θ is assumed to take values in a subset of $\mathbb{R}^k$, and the variable y is assumed to take values in a subset of $\mathbb{R}^n$: the likelihood function is defined by

$$L(\theta) = L(\theta; y) = cf(y;\theta), \tag{1}$$

where c can depend on y but not on θ. In more general settings where the model is semi-parametric or non-parametric the explicit definition is more difficult, because the density needs to be defined relative to a dominating measure, which may not exist: see Van der Vaart (1996) and Murphy and Van der Vaart (1997). This article will consider only finite-dimensional parametric models.

Within the context of the given parametric model, the likelihood function measures the relative plausibility of various values of θ, for a given observed data point y. Values of the likelihood function are only meaningful relative to each other, and for this reason are sometimes standardized by the maximum value of the likelihood function, although other reference points might be of interest depending on the context.

If our model is $f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, y = 0, 1, \ldots, n$; $\theta \in [0,1]$, then the likelihood function is (any function proportional to)

$$L(\theta; y) = \theta^y(1-\theta)^{n-y}$$

and can be plotted as a function of θ for any fixed value of y. The likelihood function is maximized at $\theta = y/n$. This model might be appropriate for a sampling scheme which recorded the number of successes among n independent trials that result in success or failure, each trial having the same probability of success, θ. Another example is the likelihood function for the mean and variance parameters when sampling from a normal distribution with mean μ and variance σ^2:

$$L(\theta; y) = \exp\{-n \log \sigma - (1/2\sigma^2)\Sigma(y_i - \mu)^2\},$$

where $\theta = (\mu, \sigma^2)$. This could also be plotted as a function of μ and σ^2 for a given sample $y_1, \ldots, y_n$, and it is not difficult to show that this likelihood function only depends on the sample through the sample mean $\bar{y} = n^{-1}\Sigma y_i$ and sample variance $s^2 = (n-1)^{-1}\Sigma(y_i - \bar{y})^2$, or equivalently through Σy_i and Σy_i^2. It is a general property of likelihood functions that they depend on the data only through the minimal sufficient statistic.

Inference

The likelihood function was defined in a seminal paper of Fisher (1922), and has since become the basis for most methods of statistical inference. One version of likelihood inference, suggested by Fisher, is to use some rule such as $L(\hat{\theta})/L(\theta) > k$ to define a range of "likely" or "plausible" values of θ. Many authors, including Royall (1997) and Edwards (1960), have promoted the use of plots of the likelihood function, along with interval estimates of plausible values. This approach is somewhat limited, however, as it requires that θ have dimension 1 or possibly 2, or that a likelihood function can be constructed that only depends on a component of θ that is of interest; see section "►Nuisance Parameters" below.

In general, we would wish to calibrate our inference for θ by referring to the probabilistic properties of the inferential method. One way to do this is to introduce a probability measure on the unknown parameter θ, typically called a prior distribution, and use Bayes' rule for conditional probabilities to conclude

$$\pi(\theta \mid y) = L(\theta; y)\pi(\theta) / \int_\theta L(\theta; y)\pi(\theta)d\theta,$$

where $\pi(\theta)$ is the density for the prior measure, and $\pi(\theta \mid y)$ provides a probabilistic assessment of θ after observing $Y = y$ in the model $f(y; \theta)$. We could then make conclusions of the form, "having observed 5 successes in 20 trials, and assuming $\pi(\theta) = 1$, the posterior probability that $\theta > 0.5$ is 0.013," and so on.

This is a very brief description of Bayesian inference, in which probability statements refer to that generated from the prior through the likelihood to the posterior. A major difficulty with this approach is the choice of prior probability function. In some applications there may be an accumulation of previous data that can be incorporated into a probability distribution, but in general there is not, and some rather *ad hoc* choices are often made. Another difficulty is the meaning to be attached to probability statements about the parameter.

Inference based on the likelihood function can also be calibrated with reference to the probability model $f(y; \theta)$, by examining the distribution of $L(\theta; Y)$ as a random function, or more usually, by examining the distribution of various derived quantities. This is the basis for likelihood inference from a frequentist point of view. In particular, it can be shown that $2\log\{L(\hat{\theta}; Y)/L(\theta; Y)\}$, where $\hat{\theta} = \hat{\theta}(Y)$ is the value of θ at which $L(\theta; Y)$ is maximized, is approximately distributed as a χ^2_k random variable, where k is the dimension of θ. To make this precise requires an asymptotic theory for likelihood, which is based on a central limit theorem (see ►Central Limit Theorems) for the *score function*

$$U(\theta; Y) = \frac{\partial}{\partial\theta} \log L(\theta; Y).$$

If $Y = (Y_1, \ldots, Y_n)$ has independent components, then $U(\theta)$ is a sum of n independent components, which under mild regularity conditions will be asymptotically normal. To obtain the χ^2 result quoted above it is also necessary to investigate the convergence of $\hat{\theta}$ to the true value governing the model $f(y; \theta)$. Showing this convergence, usually in probability, but sometimes almost surely, can be difficult: see Scholz (2006) for a summary of some of the issues that arise.

Assuming that $\hat{\theta}$ is consistent for θ, and that $L(\theta; Y)$ has sufficient regularity, the follow asymptotic results can be established:

$$(\hat{\theta} - \theta)^T i(\theta)(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_k, \tag{2}$$

$$U(\theta)^T i^{-1}(\theta) U(\theta) \xrightarrow{d} \chi^2_k, \tag{3}$$

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_k, \tag{4}$$

where $i(\theta) = E\{-\ell''(\theta; Y); \theta\}$ is the expected Fisher information function, $\ell(\theta) = \log L(\theta)$ is the log-likelihood function, and χ^2_k is the ►chi-square distribution with k degrees of freedom.

These results are all versions of a more general result that the log-likelihood function converges to the quadratic form corresponding to a multivariate normal distribution (see ►Multivariate Normal Distributions), under suitably stated limiting conditions. There is a similar asymptotic result showing that the posterior density is asymptotically

normal, and in fact asymptotically free of the prior distribution, although this result requires that the prior distribution be a proper probability density, i.e., has integral over the parameter space equal to 1.

Nuisance Parameters

In models where the dimension of θ is large, plotting the likelihood function is not possible, and inference based on the multivariate normal distribution for $\hat{\theta}$ or the χ^2_k distribution of the log-likelihood ratio doesn't lead easily to interval estimates for components of θ. However it is possible to use the likelihood function to construct inference for parameters of interest, using various methods that have been proposed to eliminate nuisance parameters.

Suppose in the model $f(y;\theta)$ that $\theta = (\psi, \lambda)$, where ψ is a k_1-dimensional parameter of interest (which will often be 1). The *profile log-likelihood* function of ψ is

$$\ell_P(\psi) = \ell(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the *constrained* maximum likelihood estimate: it maximizes the likelihood function $L(\psi, \lambda)$ when ψ is held fixed. The profile log-likelihood function is also called the concentrated log-likelihood function, especially in econometrics. If the parameter of interest is not expressed explicitly as a subvector of θ, then the constrained maximum likelihood estimate is found using Lagrange multipliers.

It can be verified under suitable smoothness conditions that results similar to those at (2 – 4) hold as well for the profile log-likelihood function: in particular

$$2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\} = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\} \xrightarrow{d} \chi^2_{k_1},$$

This method of eliminating nuisance parameters is not completely satisfactory, especially when there are many nuisance parameters: in particular it doesn't allow for errors in estimation of λ. For example the profile likelihood approach to estimation of σ^2 in the linear regression model (see ▶Linear Regression Models) $y \sim N(X\beta, \sigma^2)$ will lead to the estimator $\hat{\sigma}^2 = \Sigma(y_i - \hat{y}_i)^2/n$, whereas the estimator usually preferred has divisor $n-p$, where p is the dimension of β.

Thus a large literature has developed on improvements to the profile log-likelihood. For Bayesian inference such improvements are "automatically" included in the formulation of the marginal posterior density for ψ:

$$\pi_M(\psi \mid y) \propto \int \pi(\psi, \lambda \mid y) d\lambda,$$

but it is typically quite difficult to specify priors for possibly high-dimensional nuisance parameters. For non-Bayesian inference most modifications to the profile log-likelihood are derived by considering conditional or marginal inference in models that admit factorizations, at least approximately, like the following:

$$f(y;\theta) = f_1(y_1;\psi) f_2(y_2 \mid y_1;\lambda), \quad \text{or}$$
$$f(y;\theta) = f_1(y_1 \mid y_2;\psi) f_2(y_2;\lambda).$$

A discussion of conditional inference and density factorisations is given in Reid (1995). This literature is closely tied to that on higher order asymptotic theory for likelihood. The latter theory builds on saddlepoint and Laplace expansions to derive more accurate versions of (2–4): see, for example, Severini (2000) and Brazzale et al. (2007). The direct likelihood approach of Royall (1997) and others does not generalize very well to the nuisance parameter setting, although Royall and Tsou (2003) present some results in this direction.

Extensions to Likelihood

The likelihood function is such an important aspect of inference based on models that it has been extended to "likelihood-like" functions for more complex data settings. Examples include nonparametric and semi-parametric likelihoods: the most famous semi-parametric likelihood is the proportional hazards model of Cox (1972). But many other extensions have been suggested: to empirical likelihood (Owen 1988), which is a type of nonparametric likelihood supported on the observed sample; to quasi-likelihood (McCullagh 1983) which starts from the score function $U(\theta)$ and works backwards to an inference function; to bootstrap likelihood (Davison et al. 1992); and many modifications of profile likelihood (Barndorff-Nielsen and Cox 1994; Fraser 2003). There is recent interest for multi-dimensional responses Y_i in composite likelihoods, which are products of lower dimensional conditional or marginal distributions (Varin 2008). Reid (2000) concluded a review article on likelihood by stating:

> From either a Bayesian or frequentist perspective, the likelihood function plays an essential role in inference. The maximum likelihood estimator, once regarded on an equal footing among competing point estimators, is now typically the estimator of choice, although some refinement is needed in problems with large numbers of nuisance parameters. The likelihood ratio statistic is the basis for most tests of hypotheses and interval estimates. The emergence of the centrality of the likelihood function for inference, partly due to the large increase in computing power, is one of the central developments in the theory of statistics during the latter half of the twentieth century.

Further Reading

The book by Cox and Hinkley (1974) gives a detailed account of likelihood inference and principles of statistical inference; see also Cox (2006). There are several book-length treatments of likelihood inference, including Edwards (1960), Azzalini (1998), Pawitan (2000), and Severini (2000): this last discusses higher order asymptotic theory in detail, as does Barndorff-Nielsen and Cox (1994), and Brazzale, Davison and Reid (2007). A short review paper is Reid (2000). An excellent overview of consistency results for maximum likelihood estimators is Scholz (2006); see also Lehmann and Casella (1998). Foundational issues surrounding likelihood inference are discussed in Berger and Wolpert (1980).

About the Author

Professor Reid is a Past President of the Statistical Society of Canada (2004–2005). During (1996–1997) she served as the President of the Institute of Mathematical Statistics. Among many awards, she received the Emanuel and Carol Parzen Prize for Statistical Innovation (2008) "for leadership in statistical science, for outstanding research in theoretical statistics and highly accurate inference from the likelihood function, and for influential contributions to statistical methods in biology, environmental science, high energy physics, and complex social surveys." She was awarded the Gold Medal, Statistical Society of Canada (2009) and Florence Nightingale David Award, Committee of Presidents of Statistical Societies (2009). She is Associate Editor of *Statistical Science*, (2008–), *Bernoulli* (2007–) and *Metrika* (2008–).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Chi-Square Test: Analysis of Contingency Tables
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Estimation
►Fiducial Inference
►General Linear Models
►Generalized Linear Models
►Generalized Quasi-Likelihood (GQL) Inferences
►Mixture Models
►Philosophical Foundations of Statistics
►Risk Analysis
►Statistical Evidence
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview
►Statistics: Nelder's view
►Testing Variance Components in Mixed Linear Models
►Uniform Distribution in Statistics

References and Further Reading

Azzalini A (1998) Statistical inference. Chapman and Hall, London
Barndorff-Nielsen OE, Cox DR (1994) Inference and asymptotics. Chapman and Hall, London
Berger JO, Wolpert R (1980) The likelihood principle. Institute of Mathematical Statistics, Hayward
Birnbaum A (1962) On the foundations of statistical inference. Am Stat Assoc 57:269–306
Brazzale AR, Davison AC, Reid N (2007) Applied asymptotics. Cambridge University Press, Cambridge
Cox DR (1972) Regression models and life tables. J R Stat Soc B 34:187–220 (with discussion)
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London
Davison AC, Hinkley DV, Worton B (1992) Bootstrap likelihoods. Biometrika 79:133–130
Edwards AF (1960) Likelihood. Oxford University Press, Oxford
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans R Soc A 222:309–368
Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York
McCullagh P (1983) Quasi-likelihood functions. Ann Stat 11:59–67
Murphy SA, Van der Vaart A (1997) Semiparametric likelihood ratio inference. Ann Stat 25:1471–1509
Owen A (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Pawitan Y (2000) In all likelihood. Oxford University Press, Oxford
Reid N (1995) The roles of conditioning in inference. Stat Sci 10:138–157
Reid N (2000) Likelihood. J Am Stat Assoc 95:1335–1340
Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman and Hall, London
Royall RM, Tsou TS (2003) Interpreting statistical evidence using imperfect models: robust adjusted likelihoodfunctions. J R Stat Soc B 65:391404
Scholz F (2006) Maximum likelihood estimation. In: Encyclopedia of statistical sciences. Wiley, New York, doi: 10.1002/0471667196.ess1571.pub2. Accessed 23 Aug 2009
Severini TA (2000) Likelihood methods in statistics. Oxford University Press, Oxford
Van der Vaart AW (1996) Infinite-dimensional likelihood methods in statistics. http://www.stieltjes.org/archief/biennial9596/frame/node17.html. Accessed 18 Aug 2009
Varin C (2008) On composite marginal likelihood. Adv Stat Anal 92:1–28