Statistics

Elements of Likelihood Theory*

Advanced School of Economics
Università "Ca' Foscari" di Venezia

N. Sartori

sartori@unive.it

*A large part of these slides was kindly made available by Anthony Davison © 2007

What is Statistics?

Definition 1 *Statistics is a mathematical science pertaining to collection, analysis, interpretation and presentation of data. It is applicable to a wide variety of academic disciplines from the physical and social sciences to the humanities, as well as to business, government, and industry. (Wikipedia)*

- **Collection**:
  - surveys (social science, public policy-questionnaires, sampling, national statistics institutes, private companies)
  - design of experiments (science, medicine, agriculture, industry-clinical trials, optimisation of production, quality control)
- **Analysis**: methods of data handling (mathematics, computing,. . .)
- **Interpretation**: interaction with subject matter, explanation, prediction
- **Presentation**: construction of graphs, tables, etc. to make understanding easy (psychology)

Key idea

Variation is represented using probability distributions: regard data as outcome of random experiment.

$\Rightarrow$

The data might have been different, so the conclusion might have been different.

$\Rightarrow$

A statistical conclusion provides an explicit statement of uncertainty.

Example 2 1000 *people from a population are asked about their voting intentions.* 300 *say they will vote for party A.*

- *Not statistical:* 30% *of the population intend to vote A.*
- *Statistical:* 30% *(*$\pm x\%$*) of the population intend to vote A.*

Statistics, Mathematics and Probability

Statistics is a **mathematical science**, not a branch of mathematics!

- Mathematics concerns (correct) deduction:

$$\text{Axioms(general)} \Rightarrow \text{Consequences(specific)}$$

- Statistics concerns (correct) induction:

$$\text{Data(specific)} \Rightarrow \text{World(general)}$$

Probability provides reservoir of models and calculus to manage them – the reasoning must be correct!
Hence, we do need to know how to play with probability!

Aim of the (short) course

- Refresh/introduce probability (3 Lectures. . . in separate slides)
- Introduce likelihood theory for inference in parametric statistical models (7 Lectures).

Why likelihood theory?

- Likelihood approach to inference provides one of the most widely-used paradigm for inference based on (semi-)parametric models.
- Likelihood is also a central concept for Bayesian inference (although we are not concerned with it here. . .)

An example

Motivating example

- A company buys plastic O-rings from a supplier in lots of 5000.
- These O-rings are used in the production of a water pump.
- It is necessary to evaluate the quality of the lot in order to minimize the possibility of using a defective O-ring in the production process.
- It is not *convenient* to check all 5000 O-rings; hence only a random sample of size 50 is examined.
- There might be many reasons for *inconvenience* of examining the entire population, such as costs (monetary/ethical), time, accuracy, *virtual* population, damaging sampling,. . . .
- Therefore, we want to use the information in the *sample* of size 50 to draw conclusions on the whole *population* of size 5000.

Statistics and Probability

- Inference (from *sample* to *population*) is subject to uncertainty.
- The degree of agreement between a (**random**) sample and the population is mathematically measured using probability.
- We denote by θ the true proportion of defective O-rings in the population.
- Let us assume that we observed $y = 4$ defective O-rings among the 50 in the sample.
- Then, denoting by x the number in the lot it is natural to think that

$$4 \ : \ 50 = \ x \ : \ 5000 \quad \Rightarrow \theta = 4/50 \, .$$

Some issues

- Are there other reasonable choices for θ?
- How accurate is our evaluation of θ? What is an interval of plausible values?
- Had we observed 8 defective O-rings in a sample of 100, we would have found the same evaluation for θ. It seems reasonable that there should be more information in the larger sample. But the ratios $4/50$ and $8/100$ are not reflecting this. . . .
- If the supplier guarantees a percentage of defective O-rings not greater than 5%, are there elements for rejecting the lot? (we only observed a random sample. . .).
- The sample proportion in this particular case is a very sensible choice for evaluating θ. But is there a **general method** that could be applied in any situation, even much more complex than the present example?

Sample and population

- We need to define the probabilistic relation between the number of defective O-rings in the population (or equivalently θ) and the number y observed in the sample.
- The number y is the result of a **random experiment**.
- Hence we may think in terms of a random variable Y and that y (sometimes y_{obs}) is one of its possible values.
- The distribution of Y depends on the characteristics of the population under study (in particular θ) and partly on the sampling scheme.
- In the present context, *with some approximation*, we may think that *a priori* $Y \sim Bin(50, \theta)$

$$\Pr(Y = y) = \binom{50}{y} \theta^y (1-\theta)^{50-y}, \quad 0 \leq \theta \leq 1 \, . \tag{1}$$

Sample and population

- In principle, $\theta = 0, 1/5000, 2/5000, \ldots, 1$.
- But, given that 5000 is rather large, we may consider any real number in the interval $[0, 1]$.
- We assume the observed value y has been generated by the **true**, **unknown**, value of θ.
- In this setting, inference on θ is not seen as inference on a certain characteristic of the population under study, but on the parameter value θ that identifies a specific distribution of Y in the class (1).

Likelihood and estimates

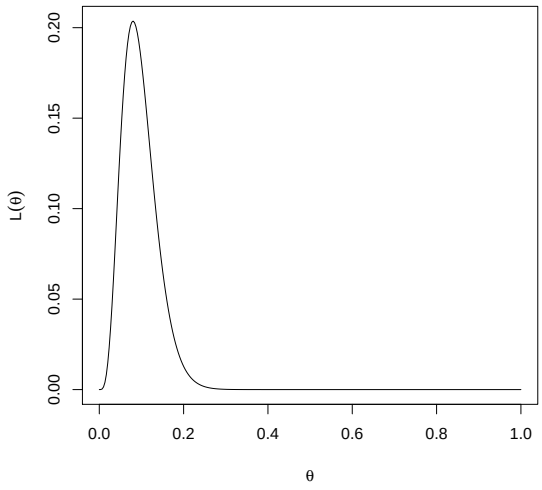
- Once observed the value $y = 4$, then (1) is a function of θ only

$$L(\theta) = \binom{50}{4} \theta^4 (1-\theta)^{46}, \quad 0 \leq \theta \leq 1\,.$$

- For any given value of θ, this is the *a priori* probability of observing the value of y that has eventually been observed.
- This can be interpreted as the degree of "agreement" between the value of θ and the observed value of y.
- For these reasons, $L(\theta)$ is called the **likelihood for** θ.
- For instance, considering three possible value of θ, $(0.05, 0.10, 0.50)$, we have

$$L(0.05) = 0.136, \quad L(0.10) = 0.181, \quad L(0.50) = 2.05 \times 10^{-10}\,.$$

Likelihood for θ



Maximum likelihood estimate

- At this point, if we had to choose a single value of θ, it is quite natural to choose the one with higher likelihood value.
- It is straightforward to see (simple mathematics...) that the maximum in this case is $\widehat{\theta} = 4/50$, which is usually called **maximum likelihood estimate**.
- Note the generality of the procedure.
- More generally, considering a sample of size n with y defective elements, the likelihood is

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

  and has maximum likelihood estimate $\widehat{\theta} = y/n$.
- Note that, even though each single value of $L(\theta)$ is a probability, $L(\theta)$ as a function of θ **is not** a probability function.

Interval estimation and hypothesis testing

- An estimate generally cannot be equal to the true parameter value.
- Hence we need to provide an estimate **and** an interval of plausible values for the parameter.
- Determining such an interval lead to what is usually called **interval estimation**.
- Alternatively, we might need to check whether a particular value of θ is not "too far" from $\widehat{\theta}$.
- This leads to what is known as **hypothesis testing**.
- Interval estimation and hypothesis testing are closely linked together.
- The likelihood function gives a natural way to address such problems.

Interval estimation

- A sensible way to determine an interval of values supported (by the observed data) is to choose all values of θ with likelihood higher than a certain threshold.
- Let us consider the **relative likelihood**

$$RL(\theta) = \frac{L(\theta)}{L(\widehat{\theta})},$$

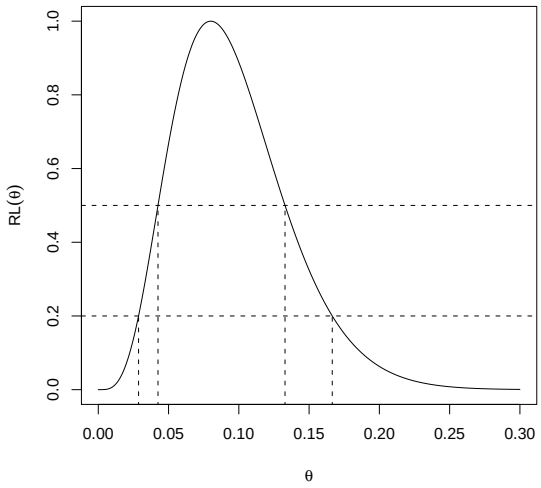
  which has values in $[0, 1]$.
- For instance we might choose the values of θ such that

$$RL(\theta) > 1/2$$

  or

$$RL(\theta) > 1/5\,.$$

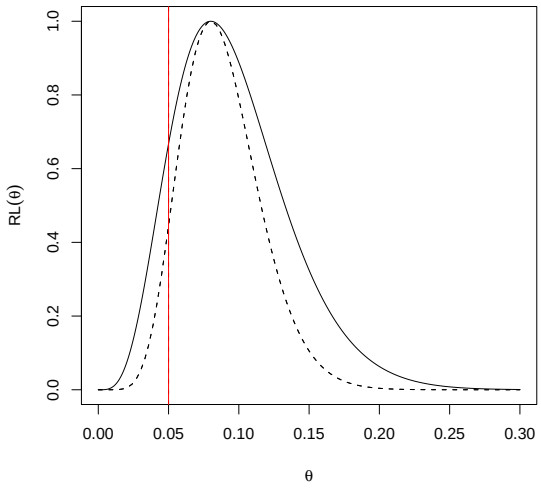
Relative likelihood



More information

- What would have happened if we had $y = 8$ in a sample of size $n = 100$.
- Of course, we'd have had the same estimate $\widehat{\theta} = 8/100 = 8\%$.
- But the likelihood would have been **more concentrated** than with $n = 50$ and the analogous intervals would have been shorter.
- This reflects the fact that a larger samples carries more information.

Two likelihoods



Hypothesis testing

- The supplier guarantees a percentage not larger than 5%.
- $\theta = 0.05$ is in the intervals...
- In particular, with $n = 50$

$$RL(0.05) = \frac{L(0.05)}{L(0.08)} = 0.668\,,$$

  and, with $n = 100$

$$RL(0.05) = \frac{L(0.05)}{L(0.08)} = 0.446\,.$$

- There seems to be no evidence in the data to reject the possibility that 0.05 be the actual true value of θ.

Repeated sampling

- We consider our estimates (and other quantities derived from statistical techniques) as random variables.
- For instance in our example $\widehat{\theta} = y/n$, where y is a realization of random variable. Therefore also $\widehat{\theta}$ is a random variable.
- In **repeated sampling** we judge properties of $\widehat{\theta}$ by studying its distribution.
- The name comes from the fact that we assume, at least in principle, that our sample is one of the possible ones that we could have observed, if we could have replicated the experiment many times under the same circumstances.

Example

- In our case we have that $Y \sim Bin(n, \theta)$.
- Therefore

$$\mathrm{E}(\widehat{\theta}) = \mathrm{E}\left\{\frac{Y}{n}\right\} = \frac{\mathrm{E}(Y)}{n} = \theta$$

- Hence $\widehat{\theta}$ is an **unbiased** estimator for θ, in the sense that at least on average it is right.
- And

$$\mathrm{var}(\widehat{\theta}) = \mathrm{var}\left\{\frac{Y}{n}\right\} = \frac{\mathrm{var}(Y)}{n^2} = \frac{\theta(1-\theta)}{n}\,.$$

- Therefore $\mathrm{var}(\widehat{\theta})$ **gets smaller as** n **gets bigger** (which is good)
- We might evaluate this variance by substituting θ with its estimate $\widehat{\theta}$.

Statistical model

- Represent data y as realisation of a random variable $Y \sim f(y;\theta)$, where parameter θ belongs to some parameter space $\Omega_\theta \subset \mathbb{R}^p$ (**parametric model**)
- f denotes probability mass function (discrete case), or probability density function (continuous case)—note these have different dimensions
- Ignore measurability issues—**all** data are discrete, so density (continuous) is just approximation to mass function (discrete) in situations where granularity is unimportant
- In general discussion use continuous/discrete case as most convenient
- We could also define nonparametric and semiparametric models, though we won't be dealing with such models here.

Parameterizations

- When we specify a statistical model, several equivalent parameterization could be chosen
- Let us assume that $Y \sim f(\cdot;\theta)$ with $\theta \in \Omega_\theta$
- If h is a one-to-one function from Ω_θ to Ω_ψ the statistical model could be written as

$$\begin{aligned} Y &\sim f(\cdot;\psi), \quad \psi = h(\theta), \quad \theta \in \Omega_\theta \\ Y &\sim f(\cdot;\psi), \quad \psi \in \Omega_\psi\,, \end{aligned}$$

  where

$$\Omega_\psi = \{\psi : \psi = h(\theta), \theta \in \Omega_\theta\}\,.$$

- We would like inferential results to be **independent on the chosen parameterization**
- Sometimes some parameterizations are more convenient than others (physical interpretation, computational aspects,...).

Alternative parameterizations

Example 3 (Exponential distribution) *Let Y have an exponential distribution with parameter θ, that is $Y \sim f(\cdot;\theta)$, where*

$$f(y;\theta) = \theta e^{-\theta y} \quad y, \theta \in \mathbb{R}^+$$

The parameter θ represent the **hazard rate** *which is*

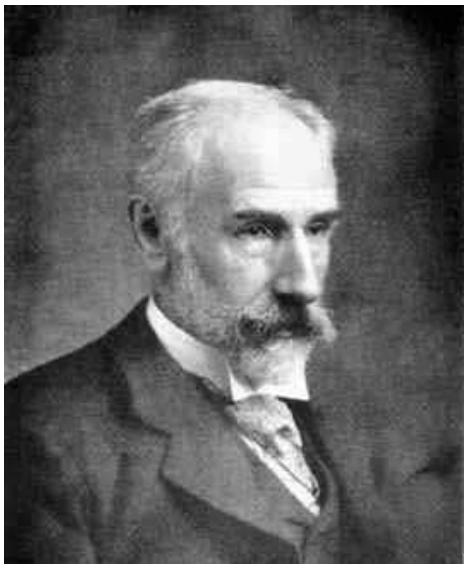
$$\theta = \Pr(Y = y + dy | Y \geq y) = \frac{f(y;\theta)}{\{1 - F(y;\theta)\}} dy$$

An alternative parameterization would be $\psi = 1/\theta$ which represents the **expected value** *of the distribution. In such parameterization we have*

$$f(y;\psi) = \psi^{-1} e^{-y/\psi} \quad y, \psi \in \mathbb{R}^+\,.$$

Two giants

Left: Francis Ysidro Edgeworth (1845–1926), who first computed correlations for the multivariate normal model, and who almost invented likelihood
Right: Ronald Alymer Fisher (1890–1962), who computed the density of the sample correlation coefficient, invented likelihood and developed its key properties



Likelihood

Definition 4 *Let y be a data set, assumed to be the realisation of a random variable $Y \sim f(y;\theta)$, where $\theta \in \Omega_\theta$ and is unknown. Then the* **likelihood** *(for θ based on y) and the corresponding* **log likelihood** *are*

$$L(\theta) = L(\theta; y) = f_Y(y;\theta), \quad \ell(\theta) = \log L(\theta), \quad \theta \in \Omega_\theta.$$

The **maximum likelihood estimate** *(MLE) $\widehat{\theta}$ satisfies $\ell(\widehat{\theta}) \geq \ell(\theta)$, for all $\theta \in \Omega_\theta$. The* **relative likelihood** *is $RL(\theta) = L(\theta)/L(\widehat{\theta})$.*
Often $\widehat{\theta}$ is unique and in many cases it satisfies the **score (or likelihood) equation**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

which is interpreted as a vector equation of dimension $p \times 1$ if θ is a $p \times 1$ vector.
The **observed information** *and* **expected (Fisher) information** *are defined as*

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad I(\theta) = \mathrm{E}\left\{J(\theta)\right\};$$

these are $p \times p$ matrices if θ has dimension p.

Properties

Likelihood is a central concept in parametric statistical inference. It gives general recipes and collections of tools for inference on θ: frequentist/repeated sampling and Bayesian.
Some elementary properties:

- **Invariance to data transformation**: if the mapping $Y \mapsto X = g(Y)$ is invertible, then the likelihood is unchanged, i.e. $L(\theta; y) = L(\theta; x)$.
- **Invariance to parameter transformation**: if $\theta \mapsto \psi = \psi(\theta)$, then $f^*(y; \psi) = f^*(y; \psi(\theta)) = f(y; \theta)$, say, so the corresponding likelihoods satisfy $L^*(\psi) = L(\theta)$. In particular $\widehat{\psi} = \psi(\widehat{\theta})$.
- **Combination of datasets**: if X, Y are independent data sets both from distributions that depend on θ, then $L(\theta; x, y) = L(\theta; y)L(\theta; x)$.
- **Shape**: we regard the graph of $\theta \mapsto L(\theta; x)$ as a summary of the information in x about θ. This could be symmetric around a unique maximum (if we're lucky!) or it can be multimodal, flat, or asymmetric, telling us how well the parameter is determined. So it gives a shape to our beliefs about θ.

Sufficiency

Definition 5 (Statistic) *A function $T(\cdot)$ from the sample space $\mathcal{Y}$ to $\mathbb{R}^r$, for some integer r, such that $T(y)$ does not depend on θ, is called a* **statistic**, *and the value $t = T(y)$ corresponding to the observed value y is called the sample value of the statistic.*

Definition 6 (Sufficient statistic) *For a given statistical model a statistic $T(y)$ is said to be sufficient for θ if it takes the same value at two points in the sample space only if these two points have equivalent likelihoods, i.e. for all $y, z \in \mathcal{Y}$*

$$T(y) = T(z) \Rightarrow L(\theta; y) \propto L(\theta; z) \text{ for all } \theta \in \Omega_\theta.$$

Theorem 7 (Neyman's factorization) *For the statistical model $T \sim f(\cdot; \theta)$, the statistic $T(\cdot)$ is sufficient for θ if and only if $f(y; \theta)$ can be written in the form*

$$f(y; \theta) = h(y)g(T(y); \theta)$$

for some functions $h(\cdot)$ and $g(\cdot; \theta)$.

Some comments on MLE

- The MLE may not exist
- The MLE need not to be unique
- The likelihood function has to be maximized in the space Ω_θ specified by the statistical model, not over the set of the mathematically admissible values of θ
- Often $\widehat{\theta}$ cannot be written explicitly as a function of the sample values, i.e. in general the MLE estimator has no closed-form expression
- In the above-mentioned case, the MLE has to be obtained numerically, for the observed value y. In real applications, this aspects is very relevant, and often give rise to interesting questions of numerical methods, or **computational statistics**, as it is now called.
- If $T(y)$ is a sufficient statistic, the MLE (as the likelihood) is a function of $T(y)$.

Examples

Example 8 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, find the maximum likelihood estimate of $\theta = (\mu, \sigma^2)^{\mathrm{T}}$, show it is unique, and find the observed and expected information.*

Example 9 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(\theta)$, show that the maximum likelihood estimate of θ does not satisfy the score equation.*

Example 10 *Suppose $Y_1, \ldots, Y_n$ follows an autoregressive process of order one, that is, $Y_j \mid Y_1, \ldots, Y_{j-1} \sim \mathcal{N}(\mu + \rho(y_{j-1} - \mu), \sigma^2)$, for $j = 2, \ldots, n$, where $|\rho| < 1$ and $Y_1 \sim \mathcal{N}(\mu, \sigma^2/(1-\rho^2))$. Find the likelihood for $\theta = (\mu, \sigma^2, \rho)^{\mathrm{T}}$.*

Pure likelihood inference

Likelihood provides a natural way to compare the plausibility of different parameter values: $L(\theta_1) > L(\theta_2)$ suggests that θ_1 is more plausible than θ_2. We could compare values of θ using a scale such as:

$$\begin{array}{rcll} 1 \geq & RL(\theta) & > \frac{1}{3}, & \theta \text{ strongly supported,} \\ \frac{1}{3} \geq & RL(\theta) & > \frac{1}{10}, & \theta \text{ supported,} \\ \frac{1}{10} \geq & RL(\theta) & > \frac{1}{100}, & \theta \text{ weakly supported,} \\ \frac{1}{100} \geq & RL(\theta) & > \frac{1}{1000}, & \theta \text{ poorly supported,} \\ \frac{1}{1000} \geq & RL(\theta) & > 0, & \theta \text{ very poorly supported.} \end{array}$$

Obviously these comparisons could also be made on the log scale, using $\log RL(\theta) = \ell(\theta) - \ell(\widehat{\theta})$.
Such inferences would satisfy the likelihood principle.
Such a scale is subjective, however, so is not usually regarded as an acceptable basis for inference. Also, it takes no account of the dimension of θ.
Below we see an objective way to calibrate the relative likelihood.

Pivots and significance functions

Definition 11 *Let $X \sim f(x;\theta)$. A* **pivot** *$Q = q(X, \theta)$ is a function of X and θ that has a known distribution, for all $\theta \in \Omega_\theta$.*

Definition 12 *Let $X \sim f(x;\theta)$ be a continuous scalar random variable that has taken value x_{obs}. The corresponding* **significance function** *is*

$$\Pr(X \leq x_{\text{obs}}; \theta).$$

Pivots are used to perform tests and to find confidence sets, using the argument that if q_α is the α quantile of Q, then $\Pr\{q(Y, \theta) \leq q_{1-2\alpha}\} = 1 - 2\alpha$, and so

$$\{\theta \in \Omega_\theta : q(Y, \theta) \leq q_{1-2\alpha}\}$$

is a $(1 - 2\alpha)$ confidence set for the true θ. Likewise for significance functions: a $(1 - 2\alpha)$ confidence set for θ is

$$\{\theta : \alpha < \Pr(X \leq x_{\text{obs}}; \theta) \leq 1 - \alpha\}.$$

Examples

Example 13 *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, *show that* $n(\overline{Y} - \mu)^2/S^2$ *and* S^2/σ^2 *are pivots.*

Example 14 *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(\theta)$, *find a significance function based on the MLE* $\widehat{\theta}$.

Repeated sampling inference

Below we consider the properties of likelihood inferences when the data $Y \sim f(y;\theta)$ are generated from f with particular value $\theta = \theta^0$.

We suppose that the data are $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta^0)$, so

- $\widehat{\theta}$ is a random variable, called the **maximum likelihood estimator**;
- $J(\theta^0)$ and $J(\widehat{\theta})$ are random variables; and
- $I(\theta^0)$ is an unknown constant.

The key results come from showing that in large samples, $\widehat{\theta} \approx \theta^0$, and from finding a central limit theorem for the score

$$U(\theta^0) = \frac{\partial \ell(\theta^0)}{\partial \theta} = \sum_{j=1}^{n} \frac{\partial \log f(Y_j;\theta^0)}{\partial \theta}.$$

More precisely ...

Definitions

Definition 15 *Let* $X, X_1, X_2, \ldots$ *be random variables with distribution functions* $F, F_1, F_2, \ldots$. *Then as* $n \to \infty$,

(a) X_n *converges* **almost surely** *to* X, $X_n \xrightarrow{\text{a.s.}} X$, *if*

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1;$$

(b) X_n *converges to* X **in mean square**, $X_n \xrightarrow{2} X$, *if*

$$\lim_{n\to\infty} \mathrm{E}\{(X_n - X)^2\} = 0, \quad \textit{where} \quad \mathrm{E}(X_n^2), \mathrm{E}(X^2) < \infty;$$

(c) X_n *converges to* X **in probability**, $X_n \xrightarrow{\text{P}} X$, *if for any* $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| \le \varepsilon) = 1;$$

(d) X_n *converges to* X **in distribution** *(or* **in law***)*, $X_n \xrightarrow{\text{D}} X$, *if*

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \textit{at every } x \textit{ where } F(x) \textit{ is continuous.}$$

Slutsky's lemma

Lemma 16 (Slutsky) *Let $y_0 \in \mathbb{R}$ be constant, and let X, Y, $\{X_n\}_{n=1}^{\infty}$, $\{Y_n\}_{n=1}^{\infty}$ be random variables. Then*

$$\left.\begin{array}{l} X_n \xrightarrow{D} X, \\ Y_n \xrightarrow{P} y_0, \end{array}\right\} \Rightarrow \left\{\begin{array}{r} X_n + Y_n \xrightarrow{D} X + y_0, \\ X_n Y_n \xrightarrow{D} X y_0. \end{array}\right.$$

Convergence theorems

Theorem 17 (Weak law of large numbers, WLLN) *Let $\{X_j\}_{j=1}^{\infty}$ be independent identically distributed random variables with common finite mean $\mu = \mathrm{E}(X_j)$. Then*

$$\overline{X} = n^{-1} \sum_{j=1}^{n} X_j \xrightarrow{P} \mu.$$

Theorem 18 (Strong law of large numbers, SLLN) *Let $\{X_j\}_{j=1}^{\infty}$ be independent identically distributed random variables. Then*

$$\overline{X} = n^{-1} \sum_{j=1}^{n} X_j \xrightarrow{\text{a.s.}} \mu$$

if and only if $\mathrm{E}(|X_1|) < \infty$, and then $\mathrm{E}(X_1) = \mu$.

Theorem 19 (Central limit theorem, CLT) *If $\{X_j\}_{j=1}^{\infty} \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ and $Z \sim \mathcal{N}(0,1)$, then*

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z.$$

Large-sample likelihood results

Theorem 20 *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$, where $\theta_{p\times 1} \in \Omega_\theta$, and f is a* **regular** *density. Suppose that the value of θ that generated the data is $\theta^0 \in \Omega_\theta$. Then as $n \to \infty$,*

(a) the maximum likelihood estimator $\widehat{\theta} \xrightarrow{\text{a.s.}} \theta^0$;

(b) the standardised score statistic $I(\theta^0)^{-1/2} U(\theta^0) \xrightarrow{D} \mathcal{N}_p(0,1)$;

(c) the Wald statistic $T(\theta^0) = J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta^0) \xrightarrow{D} \mathcal{N}_p(0,1)$;

(d) the likelihood ratio statistic $W(\theta^0) \xrightarrow{D} \chi^2_p$; and

(e) if θ is scalar, the likelihood root $R(\theta^0) = \mathrm{sign}(\widehat{\theta} - \theta^0)\sqrt{W(\theta^0)} \xrightarrow{D} \mathcal{N}(0,1)$.

Thus the standardised score statistic, Wald statistic, likelihood ratio statistic, and likelihood root, are approximate pivots, for large n.

Implications

Speaking informally, we have:

(c) implies that for large n, we have $\widehat{\theta} \dot{\sim} \mathcal{N}_p\left(\theta^0, J(\widehat{\theta})^{-1}\right)$;

(d) implies that for large n, we have $W(\theta^0) \dot{\sim} \chi^2_p$.

We can base tests and/or confidence intervals for θ^0 on these results, even when the sample is 'small' (e.g. $n = 20$ or less). Often the approximation is good, especially for (d), and it can be improved by more refined techniques.

Example 21 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \exp(\lambda)$, use (b) to construct a test of the hypothesis $\lambda = \lambda^0$.*

Example 22 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, use (c) to give a confidence interval for μ, first supposing that σ^2 is known, and then that it is unknown. Give also a test of the hypothesis $\sigma^2 = 1$. Use (d) to give a joint confidence set for (μ, σ^2).*

Choice of standardisation for $\widehat{\theta}$

We have

$$T(\theta^0) = J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta^0) \xrightarrow{D} \mathcal{N}_p(0, 1).$$

In many books $J(\widehat{\theta})$ is replaced by $J(\theta^0)$, by $I(\theta^0)$, or by $I(\widehat{\theta})$. These are equivalent to first order of approximation, but we use $J(\widehat{\theta})$ because:

- ☐ $I(\theta^0), J(\theta^0)$ involve the unknown θ^0, so can't be used directly to set confidence limits;
- ☐ $J(\widehat{\theta})$ just involves numerical derivatives and so is easy to compute;
- ☐ it automatically keeps ancillary statistics fixed, because if there is a reduction $Y \mapsto S \mapsto (T, A)$ by sufficiency and ancillarity, then
$$\log f_Y(y; \theta) = \log f_{T|A}(t \mid a; \theta) + \log f_{Y|S}(y \mid s) + \log f_A(a),$$
and so the observed information based on the full data equals that based on $T \mid A$;
- ☐ unlike $I(\theta)$ it involves no expectations over sample spaces, and so gives inferences that respect the likelihood principle more nearly;
- ☐ higher order computations show that it gives more accurate inferences;
- ☐ it gives inferences that are closer to Bayesian (later).

Likelihood calibration

The likelihood ratio statistic may be written

$$W(\theta) = 2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\} = -2 \log RL(\theta) \dot{\sim} \chi^2_p.$$

Hence a 'plausible' set of values of θ at the $(1 - 2\alpha)$ level is

$$\{\theta \in \Omega_\theta : -2 \log RL(\theta) \leq c_p(1 - 2\alpha)\}$$

and this is

$$\left\{\theta \in \Omega_\theta : \ell(\theta) \geq \ell(\widehat{\theta}) - \tfrac{1}{2}c_p(1 - 2\alpha)\right\}.$$

This is an objective way to calibrate the likelihood, but since it depends on repeated sampling and hence involves a sample space, it also seems to violate the likelihood principle.

Interest and nuisance parameters

In practice θ usually divides into

- **interest parameters** $\psi_{q\times 1}$ central to the problem (often $q = 1$ in practice);
- **nuisance parameters** $\lambda_{p-q\times 1}$ whose values are not of real importance.

Let $\widehat{\lambda}_\psi$ denote the maximum likelihood estimator of λ when ψ is fixed:

$$\ell(\psi, \widehat{\lambda}_\psi) \geq \ell(\psi, \lambda),$$

and define the **(generalised) likelihood ratio statistic**

$$W_p(\psi) = 2\left\{\ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi, \widehat{\lambda}_\psi)\right\} = 2\left\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\right\},$$

say. Then if ψ^0 is the value of ψ that generated the data from a **regular model**,

$$W_p(\psi) \xrightarrow{D} \chi^2_q, \quad n \to \infty,$$

on which confidence intervals and **likelihood ratio tests** for ψ^0 can be based.

Interest-preserving reparametrization

Desire invariance to **interest-preserving reparametrizations**: $(\psi, \lambda) \leftrightarrow (g(\psi), h(\psi, \lambda))$.

Example 23 *If $X \sim N(\mu, \sigma^2)$, then the log-normal variable $Y = \exp(X)$ has mean and variance*

$$\begin{aligned} \mathrm{E}(Y) &= \exp(\mu + \sigma^2/2) = \psi, \\ \mathrm{var}(Y) &= \exp(2\mu + \sigma^2)\{\exp(\sigma^2) - 1\} = \lambda, \end{aligned}$$

say. A confidence interval (ψ_-, ψ_+) for ψ should transform to $(\log\psi_-, \log\psi_+)$ if the model is expressed in terms of $\eta = \log\psi$ and (say) $\zeta = \zeta(\psi, \lambda) = \sigma^2$.

Likelihood ratio statistic

Sometimes there is no need to identify λ and ψ: to test the hypothesis $\theta \in \Omega^0_\theta \subset \Omega_\theta$ we compute

$$2\left\{\max_{\theta\in\Omega_\theta} \ell(\theta) - \max_{\theta\in\Omega^0_\theta} \ell(\theta)\right\} \dot{\sim} \chi^2_q,$$

where q is the difference in dimension between Ω_θ and Ω^0_θ.

Example 24 *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Compute the likelihood ratio statistic when the interest parameter is μ, and give the corresponding confidence interval.*

Main theorem

Theorem 25 *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$, where $\theta_{p\times 1} = (\psi_{q\times 1}, \lambda_{(p-q)\times 1}) \in \Omega_\theta$, and f is a* **regular** *density. Suppose that the value of ψ that generated the data is ψ^0, and let $\widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$ be the constrained maximum likelihood estimator. Let $I_{\psi\lambda}(\theta)$ denote the component of $I(\theta)$ corresponding to (ψ, λ), etc. Then as $n \to \infty$,*

(a) the standardised score statistic $\left\{I_{\psi\psi}(\widehat{\theta}_{\psi^0}) - I_{\psi\lambda}(\widehat{\theta}_{\psi^0}) I_{\lambda\lambda}(\widehat{\theta}_{\psi^0})^{-1} I_{\lambda\psi}(\widehat{\theta}_{\psi^0})\right\}^{-1/2} U_\psi(\widehat{\theta}_{\psi^0}) \xrightarrow{D} \mathcal{N}_q(0,1)$;

(b) the generalised likelihood ratio statistic $W(\psi^0) \xrightarrow{D} \chi^2_q$; *and*

(c) if ψ is scalar, the likelihood root $R(\widehat{\theta}_{\psi^0}) = \text{sign}(\widehat{\psi} - \psi^0)\sqrt{W(\psi^0)} \xrightarrow{D} \mathcal{N}(0,1)$.

Example 26 *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Gamma}(\lambda, \alpha)$. Compute a score test of the hypothesis that $\alpha = 1$, and give its limiting distribution.*

Comments

- Confidence interval based on maximum likelihood estimator
  - requires only maximisation of ℓ, computation of $\widehat{\psi}$ and of $J(\widehat{\theta})$;
  - symmetric around $\widehat{\psi}$, so widely used in practice;
  - can contain values of ψ outside Ω_θ;
  - is not transformation-invariant.
- Confidence interval based on likelihood ratio statistic
  - involves maximization of ℓ on a grid of ψ values;
  - can be asymmetric and/or non-convex;
  - is transformation-invariant;
  - is preferable to MLE interval, if available.
- Score statistic
  - is generally used only for tests;
  - involves maximization only for $\psi = \psi^0$.

Regular model (scalar θ)

Definition 27 *A statistical model $f(y;\theta)$ is* **regular (for likelihood inference)** *if*

1. *the true value θ^0 of θ is interior to the parameter space $\Omega_\theta \subset \mathbb{R}$;*
2. *the densities defined by any two different values of θ are distinct;*
3. *there is an open interval $\mathcal{I} \subset \Omega_\theta$ containing θ^0 within which the first three derivatives of the log likelihood with respect to θ exist almost surely, and $|\partial^3 \log f(Y_j;\theta)/\partial\theta^3| \le g(Y_j)$ uniformly for $\theta \in \mathcal{I}$, where $0 < \text{E}_0\{g(Y_j)\} = K < \infty$; and*
4. *the function $\partial \log f(y;\theta)/\partial\theta$ is almost surely monotonic decreasing in θ;*
5. *for $\theta \in \mathcal{I}$ we have*

$$\frac{\partial^r}{\partial\theta^r}\int f(y;\theta)\,dy = \int \frac{\partial^r}{\partial\theta^r} f(y;\theta)\,dy \quad r = 1, 2.$$

Comments

Condition

1. is needed so that $\widehat{\theta}$ can lie on both sides of θ^0 and hence can have a limiting normal distribution, once standardized;
2. is needed to be able to identify the model;
3. ensures the validity of Taylor series expansions of $\ell(\theta)$;
4. allows a simple proof of consistency of $\widehat{\theta}$; and
5. ensures that the score statistic has a limiting normal distribution.

These conditions are sufficient to give simple proofs of the results we want, but they are not necessary. Essentially similar conditions apply in the vector case—see below.
The results are also true under much weaker conditions, for non-identically distributed and dependent data.

Regular model (vector θ)

Definition 28 *A statistical model $f(y;\theta)$ is* **regular (for likelihood inference)** *if*

1. *the true value θ^0 of θ is interior to the parameter space $\Omega_\theta \subset \mathbb{R}^p$;*
2. *the densities defined by any two different values of θ are distinct;*
3. *there is an open set $\mathcal{I} \subset \Omega_\theta$ containing θ^0 within which the first three derivatives of the log likelihood with respect to elements of θ exist almost surely, and $|\partial^3 \log f(Y_j;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| \leq g(Y_j)$ uniformly for $\theta \in \mathcal{I}$, where $0 < \mathrm{E}_0\{g(Y_j)\} = K < \infty$; and*
4. *for $\theta \in \mathcal{I}$ we can interchange differentation with respect to θ and integration, that is,*

$$\frac{\partial}{\partial\theta}\int f(y;\theta)\,dy = \int \frac{\partial f(y;\theta)}{\partial\theta}\,dy, \quad \frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\int f(y;\theta)\,dy = \int \frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\,dy.$$

Theorem 20, (a)

We need Jensen's inequality: if $h : \mathbb{R} \to \mathbb{R}$ is convex and X is a real-valued random variable, then $\mathrm{E}\{h(X)\} \geq h\{\mathrm{E}(X)\}$, with equality if and only if X is degenerate.
Let $Y_1, \ldots, Y_n$ be a random sample from a density $f(y;\theta)$, where θ is scalar with true value θ^0, and let $\overline{\ell}(\theta) = n^{-1}\sum \log f(Y_j;\theta)$. Now

$$\begin{aligned} \mathrm{E}\left\{\overline{\ell}(\theta) - \overline{\ell}(\theta^0)\right\} &= \mathrm{E}\left[\log\left\{\frac{f(Y_1;\theta)}{f(Y_1;\theta^0)}\right\}\right] \\ &\leq \log \mathrm{E}\left\{\frac{f(Y_1;\theta)}{f(Y_1;\theta^0)}\right\} \\ &= \log \int \frac{f(y;\theta)}{f(y;\theta^0)} f(y;\theta^0)\, dy = 0, \end{aligned} \tag{2}$$

where we have applied Jensen's inequality to the convex function $-\log x$. The inequality is strict unless the density ratio is constant, so that the densities are the same, and condition 2 implies that this may occur only if $\theta = \theta^0$. As $n \to \infty$, the strong law of large numbers gives

$$\overline{\ell}(\theta) - \overline{\ell}(\theta^0) \xrightarrow{\text{a.s.}} \int \log\left\{\frac{f(y;\theta)}{f(y;\theta^0)}\right\} f(y;\theta^0)\, dy = -D(f_\theta, f_{\theta^0}),$$

say. This is negative unless $\theta = \theta^0$. The quantity $D(f,g) \geq 0$ is known as the *Kullback–Leibler discrepancy* between f and g; it is minimized when $f = g$.
Now for any $\delta > 0$, $\overline{\ell}(\theta^0 - \delta) - \overline{\ell}(\theta^0)$ and $\overline{\ell}(\theta^0 + \delta) - \overline{\ell}(\theta^0)$ converge with probability one to the negative quantities $-D(f_{\theta^0-\delta}, f_{\theta^0})$ and $-D(f_{\theta^0+\delta}, f_{\theta^0})$. Hence for any sequence of random variables $Y_1, \ldots, Y_n$ there is an n' such that for $n > n'$, $\overline{\ell}(\theta)$ has a local maximum in the interval $(\theta^0 - \delta, \theta^0 + \delta)$. If we let $\widehat{\theta}$ denote the value at which this local maximum occurs, then $\Pr(\widehat{\theta} \to \theta^0) = 1$ and $\widehat{\theta}$ is said to be a *strongly consistent* estimate of θ^0.
As this proof does not require $f(y;\theta)$ to be smooth it is very general. It says nothing about uniqueness of $\widehat{\theta}$, merely that a strongly consistent local maximum exists, but if $\ell(\theta)$ has just one maximum, then $\widehat{\theta}$ must also be the global maximum. A more delicate argument is needed when θ is vector, because it is then not enough to consider only the two values $\theta^0 \pm \delta$.

Theorem 20, (b)

(b) The data $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta^0)$ come from the 'true' model, and we will write $\mathrm{E}_0, \mathrm{var}_0$ to denote expectation/variance under this model. However the score and other quantities can be computed for any $\theta \in \Omega_\theta$, so we write

$$\ell(\theta) = \sum_{j=1}^{n} \log f(Y_j;\theta), \quad U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \sum_{j=1}^{n} \frac{\partial \log f(Y_j;\theta)}{\partial \theta}, \quad \theta \in \Omega_\theta.$$

Below we think of the score $U(\theta)$, the MLE $\widehat{\theta}$, and the observed information $J(\widehat{\theta})$ as being random variables, functions of $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta^0)$. Thus the expected value of the score is

$$\mathrm{E}_0\{U(\theta)\} = \sum_{j=1}^{n} \mathrm{E}_0 \left\{ \frac{\partial \log f(Y_j;\theta)}{\partial \theta} \right\} = n \int \frac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta^0)\, dy,$$

and when $\theta = \theta^0$ this equals

$$\mathrm{E}_0\{U(\theta^0)\} = \int \frac{\partial \log f(y;\theta^0)}{\partial \theta} f(y;\theta^0)\, dy = \int \frac{\partial f(y;\theta^0)}{\partial \theta} \frac{1}{f(y;\theta^0)} f(y;\theta^0)\, dy = \frac{\partial}{\partial \theta} \int f(y;\theta^0)\, dy = \frac{\partial}{\partial \theta} 1 = 0. \quad (3)$$

The variance of $U_0(\theta^0)$ is

$$n\mathrm{E}_0 \left[\left\{ \frac{\partial \log f(y;\theta^0)}{\partial \theta} \right\}^2 \right] = n \int \left\{ \frac{\partial \log f(y;\theta^0)}{\partial \theta} \right\}^2 f(y;\theta^0)\, dy.$$

A further differentiation of (3) gives

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta} \int \frac{\partial \log f(y;\theta^0)}{\partial \theta} f(y;\theta^0)\, dy &= \int \left\{ \frac{\partial^2 \log f(y;\theta^0)}{\partial \theta^2} f(y;\theta^0) + \frac{\partial \log f(y;\theta^0)}{\partial \theta} \frac{\partial f(y;\theta^0)}{\partial \theta} \right\} dy, \\ &= -i(\theta^0) + \mathrm{var}_0 \left\{ \frac{\partial \log f(y;\theta^0)}{\partial \theta} \right\}, \end{aligned}$$

which implies that $\mathrm{var}_0\{U(\theta^0)\} = I(\theta^0) = ni(\theta^0) > 0$. The CLT then gives that the centered and scaled version of $U(\theta^0)$, that is, $I(\theta^0)^{-1/2} U(\theta^0)$ has a limiting $\mathcal{N}(0,1)$ distribution.

Theorem 20, (c)

(c) The MLE $\widehat{\theta}$ is determined by the equation $U(\widehat{\theta}) = 0$, expansion of which gives

$$0 = U(\theta^0) - (\widehat{\theta} - \theta^0)J(\theta_n^*), \tag{4}$$

where $0 < |\theta_n^* - \theta^0| < |\widehat{\theta} - \theta^0|$ and so the sequence of random variables $\theta_n^* \xrightarrow{P} \theta^0$. We rearrange (4) to yield

$$J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta^0) = \sqrt{\frac{J(\widehat{\theta})}{J(\theta^0)}} \times \sqrt{\frac{J(\theta^0)}{J(\theta_n^*)}} \times \sqrt{\frac{I(\theta^0)}{J(\theta_n^*)}} \times I(\theta^0)^{-1/2}U(\theta^0), \tag{5}$$

and note that the final term on the right $\xrightarrow{D} \mathcal{N}(0,1)$, by (b). We shall show that each of the other terms on the right $\xrightarrow{P} 1$ as $n \to \infty$, and then apply Slutsky's lemma. Let $\Pr_0$ denote a probability calculated under the true model, and let $\mathcal{A}_n$ denote the event $\widehat{\theta} \in \mathcal{I}$, where $\Pr_0(\mathcal{A}_n) \to 1$ as $n \to \infty$, by (a). If $\mathcal{A}_n$ has occurred, then

$$n^{-1}|J(\widehat{\theta}) - J(\theta^0)| = \left| \frac{\ell''(\widehat{\theta}) - \ell''(\theta^0)}{n} \right| \leq |\widehat{\theta} - \theta^0| \frac{\sum_{j=1}^n g(Y_j)}{n},$$

where the terms on the right $\xrightarrow{P} 0$ and $\xrightarrow{P} K$ respectively, and so $n^{-1}\{J(\widehat{\theta}) - J(\theta^0)\} \xrightarrow{P} 0$. This implies that

$$\frac{J(\widehat{\theta})}{J(\theta^0)} - 1 = \frac{J(\widehat{\theta}) - J(\theta^0)}{n} \times \left\{ \frac{J(\theta^0)}{n} \right\}^{-1} \xrightarrow{P} 0,$$

because the first term $\xrightarrow{P} 0$ and the second $\xrightarrow{P} i(\theta^0)^{-1} > 0$ as $n \to \infty$.
If $\varepsilon > 0$ is arbitrary and $\mathcal{B}_n$ denotes the event $|J(\widehat{\theta})/J(\theta^0) - 1| < \varepsilon$, then

$$\Pr_0(\mathcal{B}_n) = \Pr_0(\mathcal{B}_n \mid \mathcal{A}_n)\Pr_0(\mathcal{A}_n) + \Pr_0(\mathcal{B}_n \cap \mathcal{A}_n^c) \geq \Pr_0(\mathcal{B}_n \mid \mathcal{A}_n)\Pr_0(\mathcal{A}_n),$$

and we have shown that both terms on the right tend to 1 as $n \to \infty$. Thus $J(\widehat{\theta})/J(\theta^0) \xrightarrow{P} 1$ as $n \to \infty$, and this implies that the first term on the right of (5) tends to 1 in probability. Similar arguments apply to the other terms, giving the required result.

Theorem 20, (d), (e)

(d) We write

$$\begin{aligned} W(\theta^0) &= 2\left\{\ell(\widehat{\theta}) - \ell(\theta^0)\right\} \\ &= 2\left[\ell(\widehat{\theta}) - \left\{\ell(\widehat{\theta}) + (\theta^0 - \widehat{\theta})U(\widehat{\theta}) - \tfrac{1}{2}(\theta^0 - \widehat{\theta})^2 J(\theta_n^*)\right\}\right] \\ &= J(\widehat{\theta})(\widehat{\theta} - \theta^0)^2 \times \frac{J(\theta^0)}{J(\widehat{\theta})} \times \frac{J(\theta_n^*)}{J(\theta^0)}, \end{aligned}$$

where $|\theta^* - \theta^0| < |\widehat{\theta} - \theta^0|$, and then note by (c) that the first term on the right $\xrightarrow{D} Z^2$, where $Z \sim \mathcal{N}(0,1)$, and the others $\xrightarrow{P} 1$ using the same arguments as in (c); now use Slutsky's lemma and note that $Z^2 \sim \chi_1^2$.
(e) Just note that (d) implies that

$$\begin{aligned} R(\theta^0) &= \operatorname{sign}(\widehat{\theta} - \theta^0)\sqrt{W(\theta^0)} \\ &= \operatorname{sign}(\widehat{\theta} - \theta^0)\sqrt{J(\widehat{\theta})(\widehat{\theta} - \theta^0)^2 \times \frac{J(\theta^0)}{J(\widehat{\theta})} \times \frac{J(\theta_n^*)}{J(\theta^0)}} \\ &= J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta^0) \times \sqrt{\frac{J(\theta^0)}{J(\widehat{\theta})} \times \frac{J(\theta_n^*)}{J(\theta^0)}} \end{aligned}$$

and apply the same arguments as in (c).

References

- Azzalini (1996) (first four chapters)
- Cox and Hinkley (1974)
- Davison (2003)
- Knight (2000)

Regular model

Definition 29 *A statistical model $f(y;\theta)$ is* **regular (for likelihood inference)** *if*

1. *the true value θ^0 of θ is interior to the parameter space $\Omega_\theta \subset \mathbb{R}^p$;*
2. *the densities defined by any two different values of θ are distinct;*
3. *there is an open set $\mathcal{I} \subset \Omega_\theta$ containing θ^0 within which the first three derivatives of the log likelihood with respect to elements of θ exist almost surely, and $|\partial^3 \log f(Y_j;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| \leq g(Y_j)$ uniformly for $\theta \in \mathcal{I}$, where $0 < \mathrm{E}_0\{g(Y_j)\} = K < \infty$; and*
4. *for $\theta \in \mathcal{I}$ we can interchange differentation with respect to θ and integration, that is,*

$$\frac{\partial}{\partial\theta}\int f(y;\theta)\,dy = \int \frac{\partial f(y;\theta)}{\partial\theta}\,dy, \quad \frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\int f(y;\theta)\,dy = \int \frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\,dy.$$

Possible problems

Condition

1. is needed so that $\widehat{\theta}$ can lie on both sides of θ^0 and hence can have a limiting normal distribution, once standardized—**fails** if, for example, we want to test for a variance parameter $\sigma^2 = 0$, or a probability equal to zero, or if the parameter has a discrete component (e.g. changepoint $\gamma \in \{1,\ldots,n\}$);
2. is needed to be able to identify the model—can **fail** if, for example, $f = (1-\gamma)f_0 + \gamma f_1$ is a mixture;
3. ensures the validity of Taylor series expansions of $\ell(\theta)$—not usually a problem;
4. ensures that the score statistic has a limiting normal distribution—can **fail** in endpoint models (e.g. $Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} U(0,\theta)$) — sometimes good news, leading to faster convergence than $n^{-1/2}$;
5. non-independent data can lead to convergence of $n^{-1}J(\theta^0)$ to random variable (or to zero), so need to use random asymptotic variance;
6. all the above assumes the postulated model is correct!

Examples

Example 30 *Consider $t_{1/\psi}$ density for long-tailed data, with $0 < \psi < 1$:*

$$f(y;\mu,\sigma^2,\psi) = \frac{\Gamma\{(\psi^{-1}+1)/2\}\psi^{1/2}}{(\sigma^2\pi)^{1/2}\Gamma\{1/(2\psi)\}}\left\{1+\psi(y-\mu)^2/\sigma^2\right\}^{-(\psi^{-1}+1)/2},$$

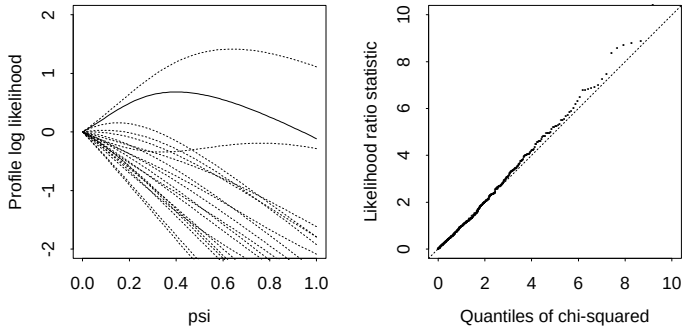
where $\sigma > 0$ and $-\infty < \mu, y < \infty$. Get $\mathcal{N}(\mu,\sigma^2)$ density when $\psi = 0$, and Cauchy density when $\psi = 1$. Note that $\psi = 0$ is a boundary case.
Profile log likelihood $\ell_p(\psi)$ based on maize data ($n = 15$)

$$-67, -48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75$$

shown on next page.
Distribution of likelihood ratio statistic is of form $p\chi^2_0 + (1-p)\chi^2_1$; for (very) large n have $p = 1/2$.
Unusual interpretation of values of score statistic: $\partial\ell_{\mathrm{p}}(\psi)/\partial\psi|_{\psi=0} < 0$ implies that $\psi = 0$.

Maize data

Left: profile log likelihoods for ψ for maize data (solid), and for 19 simulated normal samples (dots); $\psi = 0$ corresponds to the $N(\mu, \sigma^2)$ density. Right: χ^2_1 probability plot for the 1237 positive values of the likelihood ratio statistic $W_p(0)$ observed in 5000 simulated normal samples of size 15; the rest had $W_p(0) = 0$.



Example 30

To understand this, we expand $\log f(\mu, \sigma^2, \psi)$ about $\psi = 0$, giving

$$-\tfrac{1}{2}\left\{z^2 + \log(2\pi\sigma^2)\right\} + \tfrac{\psi}{4}(z^4 - 2z^2 - 1) + \tfrac{\psi^2}{2}(\tfrac{1}{2}z^4 - \tfrac{1}{3}z^6) + \tfrac{\psi^3}{24}(3z^8 - 4z^6 - 1) + O(\psi^4),$$

where $z = (y - \mu)/\sigma$. The first and second derivatives that involve ψ are
$\partial \log f/\partial\psi = (z^4 - 2z^2 - 1)/4$ and

$$\frac{\partial^2 \log f}{\partial\psi^2} = \tfrac{1}{2}z^4 - \tfrac{1}{3}z^6, \quad \frac{\partial^2 \log f}{\partial\psi\partial\mu} = (z - z^3)/\sigma, \quad \frac{\partial^2 \log f}{\partial\psi\partial\sigma^2} = (z^2 - z^4)/(2\sigma^2)$$

evaluated at $\psi = 0$, while Example 8 gives the other derivatives needed. When $\psi = 0$,
$Z = (Y - \mu)/\sigma \sim N(0, 1)$, with odd moments zero and first three even moments 1, 3, and 15, so $\mathrm{cov}(Z^4, Z^4) = 96$, $\mathrm{cov}(Z^2, Z^4) = 12$, and $\mathrm{var}(Z^2) = 2$. The expected information matrix,

$$i(\mu, \sigma^2, 0) = \begin{pmatrix} \sigma^{-2} & 0 & 0 \\ 0 & \frac{1}{2}\sigma^{-4} & \sigma^{-2} \\ 0 & \sigma^{-2} & \frac{7}{2} \end{pmatrix},$$

equals the covariance matrix of the score statistic, and the third derivatives of $\log f$ are well-behaved, so the large-sample distribution of the score vector when $\psi = 0$ is normal with mean zero and covariance matrix $ni(\mu, \sigma^2, 0)$. On setting $\lambda = (\mu, \sigma^2)$ and $\psi = 0$, Theorem 25 part (a) entails

$$\frac{\partial \ell(\widehat{\mu}_0, \widehat{\sigma}_0^2, 0)}{\partial\psi} \mathrel{\dot\sim} N(0, 3n/2).$$

In large samples this derivative is negative with probability $\frac{1}{2}$, and then $W_{\mathrm{p}}(0) = 0$; while if it is positive the usual Taylor series expansion applies and $W_{\mathrm{p}}(0) \sim \chi^2_1$. Thus the limiting distribution of $W_{\mathrm{p}}(0)$ is $\frac{1}{2} + \frac{1}{2}\chi^2_1$, giving

$$\Pr\{W_{\mathrm{p}}(0) \le 1.366\} = \tfrac{1}{2} + \tfrac{1}{2}\Pr(\chi^2_1 \le 1.366) = 0.88.$$

The asymptotic distribution of $\widehat{\psi}$ puts mass $\frac{1}{2}$ at $\psi = 0$, with the remaining probability spread as a normal density confined to the positive half-line.
To assess the quality of such approximations, 5000 normal samples of size $n = 15$ were generated. Just 1237 of the $W_{\mathrm{p}}(0)$ were positive, but those that were had distribution close to χ^2_1, as the right panel of the figure above shows. Hence

$$\Pr\{W_{\mathrm{p}}(0) \le 1.366\} \doteq (3763/5000) + (1237/5000)\Pr(\chi^2_1 \le 1.366) = 0.94,$$

stronger though not decisive evidence for the t model. Large-sample results are unreliable even with $n = 100$, when $\Pr\{W_{\mathrm{p}}(0) = 0\} \doteq 0.37$.
Such problems also arise if the favoured model is close to the boundary. For example, despite being normal in large samples, when n is small the distribution of $\widehat{\psi}$ would have a point mass at $\psi = 0$. If several parameters lie on their boundaries, then asymptotics become yet more cumbersome. Simulation seems preferable.

Example

Example 31 (Normal mixture) *Discuss inference based on a random sample from the normal mixture density*

$$\frac{\gamma}{(2\pi)^{1/2}\sigma_1}\exp\left\{-\frac{(y-\mu_1)^2}{2\sigma_1^2}\right\}+\frac{1-\gamma}{(2\pi)^{1/2}\sigma_2}\exp\left\{-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right\},\quad 0\le\gamma\le 1,$$

with the means and variances in their usual ranges.

Example 32 (Poisson birth process) *Consider a sequence $Y_0,\ldots,Y_n$ such that given the values of $Y_0,\ldots,Y_{j-1}$, the variable Y_j has a Poisson density with mean θY_{j-1}, and $\mathrm{E}(Y_0)=\theta$. Show that the log likelihood and observed information are*

$$\ell(\theta)\equiv\sum_{j=0}^{n}Y_j\log\theta-\theta\left(1+\sum_{j=0}^{n-1}Y_j\right),\quad J(\theta)=\theta^{-2}\sum_{j=0}^{n}Y_j,$$

and discuss inference for this model.

Example 31

For an example of a non-smooth likelihood, let $L(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\gamma)$ be the likelihood for a random sample $y_1,\ldots,y_n$ from the mixture of normal densities

$$\frac{\gamma}{(2\pi)^{1/2}\sigma_1}\exp\left\{-\frac{(y-\mu_1)^2}{2\sigma_1^2}\right\}+\frac{1-\gamma}{(2\pi)^{1/2}\sigma_2}\exp\left\{-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right\},\quad 0\le\gamma\le 1,$$

with the means and variances in their usual ranges. This corresponds to taking observations in proportions γ, $1-\gamma$ from two normal populations, not knowing from which they come.
Note first that this model is not identifiable, as switching $(\mu_1,\sigma_1,\gamma)\leftrightarrow(\mu_2,\sigma_2,1-\gamma)$ gives the same density.
If $\gamma\neq 0,1$, then for each y_j

$$\lim_{\sigma_1\to 0}L(y_j,\mu_2,\sigma_1^2,\sigma_2^2,\gamma)=\lim_{\sigma_2\to 0}L(\mu_1,y_j,\sigma_1^2,\sigma_2^2,\gamma)=+\infty,$$

so L is a smooth surface pocked with singularities, each of which corresponds to estimating the mean and variance of one of the populations from a single observation. For large n the strong consistency result guarantees the existence of a smooth local maximum of L near the true parameter values. When finding this numerically a careful choice of starting values can help one avoid ending up at a spike instead, but it is worth asking why they occur.
The issue is rounding. The fiction that data are continuous is usually harmless and convenient. Here it is not harmless, however, because it results in infinite likelihoods. The spikes can be removed by accounting for the rounding of the y_j. If they are rounded to multiples of δ, then
$\Pr(Y=k\delta)=F(k\delta+\delta/2)-F(k\delta-\delta/2)$, where

$$F(y)=\gamma\Phi\left(\frac{y-\mu_1}{\sigma_1}\right)+(1-\gamma)\Phi\left(\frac{y-\mu_2}{\sigma_2}\right).$$

As $0<F(y_j)<1$, the largest possible contribution to L is then finite.

Example 32

The expected value of Y_j, given Y_{j-1}, is θY_{j-1}, so its unconditional expectation is θ^{j+1}. Hence the expected information is $I(\theta) = \theta^{-2}(\theta + \cdots + \theta^{n+1})$. If $\theta \geq 1$, then $I(\theta) \to \infty$ as $n \to \infty$, but if not, $I(\theta)$ is asymptotically bounded. In fact, as $n \to \infty$, the process is certain to become extinct — that is, there will be an n_0 such that $Y_{n_0} = Y_{n_0+1} = \cdots = 0$ — unless $\theta > 1$, and even then there is a non-zero probability of extinction. Hence $J(\theta)$ remains finite with probability one unless $\theta > 1$, and remains finite with non-zero probability for any θ. Thus the maximum likelihood estimator $\widehat{\theta} = (Y_0 + \cdots + Y_n)/(1 + Y_0 + \cdots + Y_{n-1})$ is neither consistent nor asymptotically normal if $\theta \leq 1$. From a practical viewpoint, this failure of standard asymptotics is less critical than it might appear, but we can still use standard approximations if they can be justified by other means. Inference is not impossible, merely more difficult than with independent data.

Example

Example 33 (Shifted exponential density) *Discuss inference based on a random sample from the shifted exponential density*

$$f(y; \phi, \theta) = \theta^{-1} \exp\{-(y - \phi)/\theta\}, \quad y > \phi, \theta > 0.$$

Lemma 33

The corresponding random variables $Y_1, \ldots, Y_n$ have the same distribution as $\phi + \theta E_1, \ldots, \phi + \theta E_n$, where $E_1, \ldots, E_n$ is a random sample from the standard exponential density. The log likelihood contribution from a single observation $y > \phi$ is $\ell(\phi, \theta) = -\log\theta - (y - \phi)/\theta$, so

$$\frac{\partial \ell(\phi, \theta)}{\partial \phi} = \begin{cases} \theta^{-1}, & y > \phi, \\ 0, & \text{otherwise.} \end{cases}$$

For a regular model this would have mean zero, but here the interchange of differentiation and integration fails because the support of the density depends on ϕ, and $\mathrm{E}(\partial\ell/\partial\phi) = \theta^{-1}$.
The likelihood is $L(\phi, \theta) = \theta^{-n} \exp\{-n(\overline{y} - \phi)/\theta\}$ for $y_1, \ldots, y_n > \phi$ and $\theta > 0$, and for any θ this increases as $\phi \uparrow \min y_j$ and is zero thereafter. Thus ϕ has maximum likelihood estimate $\widehat{\phi} = y_{(1)}$, while $\widehat{\theta} = \overline{y} - \widehat{\phi} = \overline{y} - y_{(1)}$.
To find limiting distributions of $\widehat{\phi}$ and $\widehat{\theta}$, recall that the rth order statistic $E_{(r)}$ of a standard exponential random sample may be written $\sum_{j=1}^{r} (n + 1 - j)^{-1} E_j$, where $E_1, \ldots, E_n$ is an exponential random sample. As $Y_{(r)} \stackrel{D}{=} \phi + \theta E_{(r)}$, we see that $Y_{(1)} \stackrel{D}{=} \phi + n^{-1}\theta E_1$, implying that $n\theta^{-1}(\widehat{\phi} - \phi) \stackrel{D}{=} E_1$: the rescaled endpoint estimate $\widehat{\phi}$ has a non-normal limit distribution. Moreover it converges faster than usual because $\widehat{\phi} - \phi$ must be multiplied by n rather than $n^{1/2}$ in order to give a non-degenerate limit.
For the distribution of $\widehat{\theta}$, note that as $\overline{Y} - Y_{(1)} = n^{-1} \sum_{r=1}^{n} Y_{(r)} - Y_{(1)}$,

$$\widehat{\theta} \stackrel{D}{=} n^{-1} \left\{ n\phi + \theta \sum_{r=1}^{n} \sum_{j=1}^{r} \frac{E_j}{n + 1 - j} - n\phi - \theta E_1 \right\} = n^{-1}(n-1)\theta\overline{E},$$

with $\overline{E}$ the average of $E_2, \ldots, E_n$. The central limit theorem implies that $n^{1/2}(\widehat{\theta} - \theta)/\theta \stackrel{D}{\longrightarrow} N(0, 1)$, so standard asymptotics apply to $\widehat{\theta}$ despite their failure for $\widehat{\phi}$, which converges so fast that its randomness has no impact on the limiting distribution of $\widehat{\theta}$.
In this problem exact inference is possible for any n, but the general conclusion is that endpoints must be treated gingerly.

Wrong model

Suppose the true model is g, that is, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, but we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. The log likelihood $\ell(\theta)$ will be maximised at $\widehat{\theta}$, and

$$\overline{\ell}(\widehat{\theta}) = n^{-1}\ell(\widehat{\theta}) \xrightarrow{\text{a.s.}} \int \log f(y;\theta_g) g(y)\, dy, \quad n \to \infty,$$

where θ_g minimizes the Kullback–Leibler discrepancy

$$D(f_\theta, g) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)\, dy.$$

θ_g gives the density $f(y;\theta_g)$ closest to g in this sense, and $\widehat{\theta}$ is determined by the finite-sample version of $\partial D(f_\theta, g)/\partial\theta$, i.e.

$$0 = n^{-1}\sum_{j=1}^{n} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial\theta}. \tag{6}$$

Wrong model II

Theorem 34 *Suppose the true model is g, that is, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, but we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. Then under mild regularity conditions the maximum likelihood estimator $\widehat{\theta}$ satisfies*

$$\widehat{\theta} \;\dot{\sim}\; N_p\left\{\theta_g, I(\theta_g)^{-1}K(\theta_g)I(\theta_g)^{-1}\right\},$$

where f_{θ_g} is the density minimising the Kullback–Leibler discrepancy between f_θ and g, I is the Fisher information, and K is the variance of the score statistic.
For large n, the likelihood ratio statistic $W(\theta_g) \;\dot{\sim}\; \sum \lambda_r V_r$, where the $V_1, \ldots, V_p \overset{\text{iid}}{\sim} \chi^2_1$, and the λ_r are eigenvalues of $K(\theta_g)^{1/2} I_g(\theta_g)^{-1} K(\theta_g)^{1/2}$.

Note that under the correct model, $\theta_g = \theta^0$, $K(\theta_g) = I(\theta_g)$, and we recover the usual results.

Theorem 34

Expansion of (6) about θ_g yields

$$\widehat{\theta} \doteq \theta_g + \left\{ -n^{-1} \sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\theta_g)}{\partial\theta\partial\theta^{\mathrm{T}}} \right\}^{-1} \left\{ n^{-1} \sum_{j=1}^{n} \frac{\partial \log f(y_j;\theta_g)}{\partial\theta} \right\}$$

and a modification of the previous derivation gives

$$\widehat{\theta} \;\dot{\sim}\; N_p\left\{\theta_g, I(\theta_g)^{-1} K(\theta_g) I(\theta_g)^{-1}\right\}, \tag{7}$$

where the *information sandwich* variance matrix depends on

$$\begin{aligned} K(\theta_g) &= n \int \frac{\partial \log f(y;\theta)}{\partial\theta} \frac{\partial \log f(y;\theta)}{\partial\theta^{\mathrm{T}}} g(y)\, dy, \\ I_g(\theta_g) &= -n \int \frac{\partial^2 \log f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} g(y)\, dy. \end{aligned} \tag{8}$$

If $g(y) = f(y;\theta)$, so that the supposed density is correct, then θ_g is the true θ, then $K(\theta_g) = I_g(\theta_g) = I(\theta)$, and (7) reduces to the usual approximation.
In practice $g(y)$ is of course unknown, and then $K(\theta_g)$ and $I_g(\theta_g)$ may be estimated by

$$\widehat{K} = \sum_{j=1}^{n} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial\theta} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial\theta^{\mathrm{T}}}, \quad \widehat{J} = -\sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\widehat{\theta})}{\partial\theta\partial\theta^{\mathrm{T}}}; \tag{9}$$

the latter is just the observed information matrix. We may then construct confidence intervals for θ_g using (7) with variance matrix $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$.
For future reference we give the approximate distribution of the likelihood ratio statistic. Taylor series approximation gives

$$\begin{aligned} 2\left\{\ell(\widehat{\theta}) - \ell(\theta_g)\right\} &\doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}} \left\{ -\sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\theta_g)}{\partial\theta\partial\theta^{\mathrm{T}}} \right\} (\widehat{\theta} - \theta_g) \\ &\doteq n(\widehat{\theta} - \theta_g)^{\mathrm{T}} I_g(\theta_g)(\widehat{\theta} - \theta_g) \end{aligned}$$

and the normal distribution (7) of $\widehat{\theta}$ implies that the likelihood ratio statistic has a distribution proportional to χ^2_p, but with mean $\mathrm{tr}\{I_g(\theta_g)^{-1}K(\theta_g)\}$. If the model is correct, $I_g(\theta_g) = K(\theta_g)$, giving the previous mean, p.

Example

Example 35 *Let $f(y;\theta)$ be the exponential density with mean θ, while in fact $Y = e^{\sigma Z}$, where Z is standard normal. Then Y is log-normal, with mean $e^{\sigma^2/2}$ and variance $e^{\sigma^2}(e^{\sigma^2} - 1)$. Discuss inference based on a random sample $Y_1, \ldots, Y_n$.*

Example 35

Let $f(y;\theta)$ be the exponential density with mean θ, while in fact $Y = e^{\sigma Z}$, where Z is standard normal. Then Y is log-normal, with mean $e^{\sigma^2/2}$ and variance $e^{\sigma^2}(e^{\sigma^2} - 1)$.
The presumed log likelihood is $-\log\theta - y/\theta$, so that

$$\int \log f(y;\theta) g(y)\, dy = -\log\theta - \theta^{-1} \int y g(y)\, dy = -\log\theta - \theta^{-1} e^{\sigma^2/2},$$

and differentiation of this with respect to θ gives $\theta_g = e^{\sigma^2/2}$. Here the 'least bad' exponential model has the same mean as the true log-normal distribution, which must always exceed one. Further calculation gives $I(\theta_g) = \theta_g^{-2}$ and $K(\theta_g) = 1 - \theta_g^{-2}$,
The maximum likelihood estimate of θ is $\widehat{\theta} = \overline{Y}$, and either directly or using the information sandwich we see that $\mathrm{var}(\widehat{\theta}) = n^{-1}\theta_g^2(\theta_g^2 - 1)$. Note that replacement of θ_g with its estimate $\widehat{\theta}$ could result in a negative variance. This is not the case if we use the empirical variance — simple calculations give $\widehat{J} = n/\overline{y}^2$ and $\widehat{K} = \overline{y}^{-4} \sum (y_j - \overline{y})^2$, so $\widehat{J}^{-2}\widehat{K} = n^{-2} \sum (y_j - \overline{y})^2$. Reassuringly, this is a consistent estimate of the variance of the average of a random sample from any distribution with finite variance. As $I_g(\theta_g)^{-1} K(\theta_g) = e^{\sigma^2} - 1 = \theta_g^2 - 1$, the likelihood ratio statistic may be over- or under-dispersed relative to the χ^2_1 distribution.

References

- Barndoff-Nielsen and Cox (1994)
- Basawa and Scott (1981)
- Self and Liang (1987)
- Smith (1989)
- Various papers in economics (White estimator, sandwich variance estimator)

Thomas Bayes (1702–1761)

Bayes (1763/4) *Essay towards solving a problem in the doctrine of chances*. Philosophical Transactions of the Royal Society of London.

Bayesian inference

Parametric model for data y assumed to be realisation of $Y \sim f(y;\theta)$, where $\theta \in \Omega_\theta$.
Frequentist viewpoint:

- there is a true value of θ that generated the data;
- this 'true' value of θ is to be treated as an unknown constant;
- probability statements concern randomness in the data, possibly conditioned on an ancillary statistic.

Bayesian viewpoint:

- ignorance about unknowns such as θ should (and can) be expressed using probability distributions;
- Bayes' theorem can be used to convert prior beliefs $\pi(\theta)$ about θ into posterior beliefs $\pi(\theta \mid y)$;
- probability statements concern randomness about unknowns, conditioned on all known quantities.

Mechanics

- Separate from data, we have prior information about scalar θ summarised in density $\pi(\theta)$
- Data model $f(y \mid \theta) \equiv f(y; \theta)$
- Posterior density given by Bayes' theorem:

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int \pi(\theta) f(y \mid \theta)\, d\theta}.$$

- $\pi(\theta \mid y)$ contains all information about θ, conditional on data value y
- Posterior confidence bound for θ is quantile of $\pi(\theta \mid y)$:

$$\Pr\{\theta \leq \theta^{\alpha}(y) \mid y\} = \int_{-\infty}^{\theta^{\alpha}(y)} \pi(\theta \mid y)\, d\theta = \alpha,$$

  giving $(1 - 2\alpha)$ posterior credible interval $(\theta^{\alpha}(y), \theta^{1-\alpha}(y))$.

Conjugate priors

Certain combinations of data and prior give posterior densities of the same form as the prior. Example: $s \sim B(n, \theta)$ gives

$$\theta \sim \text{Beta}(a, b) \xrightarrow{s,n} \theta \mid y \sim \text{Beta}(a + s, b + n - s).$$

The beta density is the **conjugate prior** for binomial data.

Lemma 36 *Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y; \theta)$ is a random sample from an exponential family. Then the conjugate prior for θ has form*

$$\pi(\theta) = \exp\{\xi\theta - \nu\kappa(\theta)\}\, k(\xi, \nu),$$

where the **hyperparameters** *ξ, ν determine the prior.*

Lemma 36

Suppose that $y_1, \ldots, y_n$ is a random sample from the exponential family

$$f(y \mid \theta) = \exp\{y\theta - \kappa(\theta)\}\, c(y),$$

so that in terms of $s = \sum y_j$, the likelihood is proportional to

$$\exp\{s\theta - n\kappa(\theta)\}. \tag{10}$$

If the prior density for θ depends on the quantities ξ and ν and has form

$$\pi(\theta) = \exp\{\xi\theta - \nu\kappa(\theta)\}\, k(\xi, \nu),$$

then the posterior density is proportional to

$$\exp\{(\xi + s)\theta - (\nu + n)\kappa(\theta)\}.$$

Provided this is integrable the posterior density therefore must be

$$\pi(\theta \mid y) = \exp\{(\xi + s)\theta - (\nu + n)\kappa(\theta)\}\, k(\xi + s, \nu + n).$$

Thus the prior parameters (ξ, ν) are updated to $(\xi + s, \nu + n)$ by the data. One interpretation of the *hyperparameters* ξ and ν is that the prior information is equivalent to ν prior observations summing to ξ.

Arguments for/against Bayes

For:

- provides unified approach to inference—all unknowns, data, parameters, predictands are treated on the same footing;
- argument based on axioms of 'rational behaviour' under uncertainty leads to 'coherent' (i.e. internally consistent) Bayes inference;
- satisfies likelihood principle (recall that tests and confidence intervals don't);
- simple recipe (in principle)—just apply Bayes' theorem.

Against:

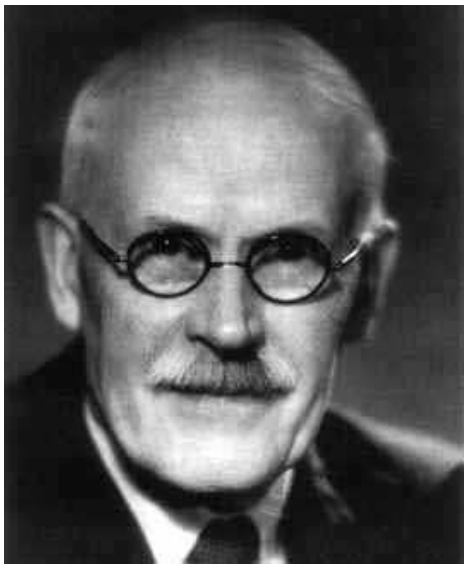
- is it always appropriate to treat data (whose model is checkable) on the same footing as the prior? Surely this should depend on the context?
- Different priors will give different answers. Which is to be believed by a third party?
- external validity with respect to real world is more important than internal consistency (one can be consistently wrong!)

Potted history

- Bayes' theorem published posthumously in 1763, but he was concerned about finding 'objective' prior for binomial parameter
- Early-mid 19th century: Laplace advocated Bayesian viewpoint
- Late 19th century: Venn argued against Bayes on grounds that objective prior could not be found
- Fisher (1920s–1962) strongly anti-Bayes, on same grounds
- Jeffreys (1939, 1962) advocated Bayes with 'objective' priors
- Savage (1950s) advocated use of 'subjective' or 'personalist' priors
- Welch/Peers (1963) suggested matching priors—make Bayes and frequentist inference agree
- de Finetti (1970s)—'there is no such thing as probability' (subjective degree of belief only)
- Bernardo/Berger (1979–) suggest use of 'reference' priors
- Empirical Bayes (1960s–) uses priors estimated from data
- Most modern applications use priors chosen for convenience, and vary them for sensitivity analysis

Two giants

Left: Harold Jeffreys (1891–1989), a geophysicist and astronomer who developed a (failed) theory of objective inference based on noninformative prior distributions.
Right: Ronald Alymer Fisher (1890–1962), a geneticist and statistician who developed a (failed) theory of objective inference based on the 'fiducial' distribution.



'Ignorance'

What prior density represents ignorance about θ?

Definition 37 *(a) A prior is called* **uniform** *if* $\pi(\theta) \propto 1$, $\forall \theta$.
(b) A prior is called **improper** *if it cannot be renormalised to have finite integral.*
(c) The **Jeffreys prior** *for the parameter* θ *of a regular statistical model is defined to be* $\pi(\theta) \propto |i(\theta)|^{1/2}$, *where* $i(\theta)$ *is the Fisher information for* θ.

Example 38 *Let* $0 < \theta < 1$. *Show that a uniform prior for* θ *yields a non-uniform prior for* $\psi = \log\{\theta/(1-\theta)\}$.

Lemma 39 *The Jeffreys prior is invariant to smooth reparametrizations* $\theta = \theta(\psi)$.

Example 38

The probability of success in a Bernoulli trial lies in the interval $[0, 1]$, so if we are completely ignorant of its true value, the obvious prior to use is uniform on the unit interval: $\pi(\theta) = 1$, $0 \le \theta \le 1$. But if we are completely ignorant of θ, we are also completely ignorant of $\psi = \log\{\theta/(1-\theta)\}$, which takes values in the real line. The density implied for ψ by the uniform prior for θ is

$$\pi(\psi) = \pi\{\psi(\theta)\} \times \left|\frac{d\theta}{d\psi}\right| = \frac{e^{\psi}}{(1+e^{\psi})^2}, \quad -\infty < \psi < \infty :$$

the standard logistic density. Far from expressing ignorance about ψ, this density asserts that the prior probability of $|\psi| < 3$ is about 0.9.

Lemma 39

For a smooth reparametrization $\theta = \theta(\psi)$ in terms of ψ, the expected information for ψ is

$$i(\psi) = -\mathrm{E}\left[\frac{d^2\ell\{\theta(\psi)\}}{d\psi^2}\right] = -\mathrm{E}\left\{\frac{d^2\ell(\theta)}{d\theta^2}\right\} \times \left|\frac{d\theta}{d\psi}\right|^2 = i(\theta) \times \left|\frac{d\theta}{d\psi}\right|^2 .$$

Consequently $|i(\theta)|^{1/2} d\theta = |i(\psi)|^{1/2} d\psi$: the Jeffreys prior does behave consistently under reparametrization; furthermore such priors give widely-accepted solutions in some standard problems. When θ is vector, $|i(\theta)|$ is taken to be the determinant of $i(\theta)$.
This prior was initially proposed with the aim of giving an 'objective' basis for inference, but after further paradoxes emerged its use was suggested for convenience, a matter of scientific convention rather than as a logically unassailable expression of ignorance about the parameter.

Edgeworth series

Definition 40 *Let $X_1, \ldots, X_n$ be a random sample of continuous variables with cumulant-generating function $K(u)$ and finite cumulants κ_r, let $\rho_r = \kappa_r/\kappa_2^{r/2}$ denote the rth standardized cumulant, and let $Z_n = (S_n - n\kappa_1)/(n\kappa_2)^{1/2}$ denote the standardized version of $S_n = X_1 + \cdots + X_n$. Also let*

$$\begin{aligned} H_1(z) &= z,\ H_2(z) = z^2 - 1,\ H_3(z) = z^3 - 3z,\ H_4(z) = z^4 - 6z^2 + 3, \\ H_5(z) &= z^5 - 10z^3 + 15z,\ H_6(z) = z^6 - 15z^4 + 45z^2 - 15 \end{aligned}$$

denote the Hermite polynomials. Then the **Edgeworth series** *for the distribution of Z_n is*

$$F_{Z_n}(z) = \Phi(z) - \phi(z)\left[\frac{\rho_3}{6n^{1/2}}H_2(z) + \frac{1}{n}\left\{\frac{\rho_4}{24}H_3(z) + \frac{\rho_3^2}{72}H_5(z)\right\} + O(n^{-3/2})\right],$$

and **Cornish–Fisher inversion** *yields that the α quantile of $F_{Z_n}(z)$ equals*

$$z_\alpha + \frac{\rho_3}{6n^{1/2}}H_2(z_\alpha) + \frac{1}{n}\left\{\frac{\rho_4}{24}H_3(z_\alpha) + \frac{\rho_3^2}{36}(5z_\alpha - 2z_\alpha^3)\right\} + O(n^{-3/2}).$$

Mechanics

- Separate from data, we have prior information about scalar θ summarised in density $\pi(\theta)$
- Data model $f(y \mid \theta) \equiv f(y; \theta)$
- Posterior density given by Bayes' theorem:

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int \pi(\theta) f(y \mid \theta)\, d\theta}.$$

- $\pi(\theta \mid y)$ contains all information about θ, conditional on data value y
- Posterior confidence bound for θ is quantile of $\pi(\theta \mid y)$:

$$\Pr\{\theta \le \theta^\alpha(y) \mid y\} = \int_{-\infty}^{\theta^\alpha(y)} \pi(\theta \mid y)\, d\theta = \alpha,$$

  giving $(1 - 2\alpha)$ posterior credible interval $(\theta^\alpha(y), \theta^{1-\alpha}(y))$.

Matching priors

Definition 41 *A* **matching prior** *is one for which Bayesian posterior probability statements about the parameter may also be interpreted as confidence statements in the sampling model.*

Lindley (1958), Welch and Peers (1963), Peers (1965), ...
Hope to:

- provide a Bayes/frequentist compromise
- provide default priors for routine Bayesian use
- provide basis for assessment of robustness of inference using a specified prior

Basic setup

Suppose $Y_1, \ldots, Y_n$ random sample with joint density $f(y \mid \theta)$, with prior $\pi(\theta)$ and $\theta \in \mathbb{R}^p$, and let $\widehat{\theta}$ be the MLE and $\widehat{\sigma}^2/n = J(\widehat{\theta})^{-1}$ be its asymptotic variance based on the observed information.

Definition 42 *Under the above setup and if θ is scalar, let $\theta^{1-\alpha}$ denote the $(1-\alpha)$ posterior quantile of θ, which satisfies*

$$\Pr_{\theta|Y}\left\{\theta \leq \theta^{1-\alpha}(y) \mid y\right\} = \int_{-\infty}^{\theta^{1-\alpha}} \pi(\theta \mid y)\, d\theta = 1-\alpha.$$

If

$$\Pr_{Y|\theta}\left\{\theta^{1-\alpha}(Y) \geq \theta\right\} = \int I\{\theta^{1-\alpha}(y) \geq \theta\} f(y \mid \theta)\, dy = 1-\alpha,$$

then Bayes and frequentist inference would agree perfectly.

Matching: scalar θ

- Obtain Edgeworth series for $n^{1/2}(\theta - \widehat{\theta})/\widehat{\sigma}$, conditional on data—so $\widehat{\theta}(y), \widehat{\sigma}(y)$ are constants
- Obtain corresponding Cornish–Fisher series for $\theta^{1-\alpha}$:

$$\theta^{1-\alpha}(y) = \widehat{\theta} + \frac{\widehat{\sigma}}{n^{1/2}} z_\alpha + \frac{\widehat{\sigma}}{n}\left\{(z_\alpha^2+2)A_3(y) + A_1(y)\right\} + O(n^{-3/2}).$$

- Use this expansion in the expression

$$\Pr_{Y|\theta}\left\{\theta^{1-\alpha}(Y) \geq \theta\right\} = \int I\{\theta^{1-\alpha}(y) \geq \theta\} f(y \mid \theta)\, dy$$

and obtain

$$1-\alpha + \frac{\phi(z_\alpha)}{n^{1/2}} T_1(\pi,\theta) + \frac{z_\alpha \phi(z_\alpha)}{n} T_2(\pi,\theta) + O(n^{-3/2}),$$

where

$$T_1(\pi,\theta) = \frac{1}{\pi(\theta)} \frac{d}{d\theta}\left\{\frac{\pi(\theta)}{i(\theta)^{1/2}}\right\}, \qquad T_2 = 0 \text{ iff } \frac{d}{d\theta}\left\{\frac{\mathrm{E}_{Y|\theta}\{U(\theta)\}^3}{i(\theta)^{3/2}}\right\} = 0,$$

and the condition on T_2 here holds when $\pi(\theta) \propto i(\theta)^{1/2}$

Discussion: scalar θ

- $T_1(\pi,\theta) \equiv 0$ if and only if

$$\pi(\theta) \propto i(\theta)^{1/2},$$

so the Jeffreys prior is matching to order n^{-1}

- with the Jeffreys prior, T_2 vanishes if and only if a condition

$$\frac{d}{d\theta}\left\{\frac{\mathrm{E}_{Y|\theta}\{U(\theta)\}^3}{i(\theta)^{3/2}}\right\} = 0$$

on the model is satisfied. Hence higher-order matching is only possible in special cases

Matching: vector θ

- Suppose $\theta = (\theta_1, \ldots, \theta_p)$, and define posterior quantile of θ_1 by

$$\int_{-\infty}^{\theta_1^{1-\alpha}} \pi(\theta_1 \mid y)\, d\theta_1 = 1 - \alpha,$$

  where $\pi(\theta_1 \mid y)$ is the marginal posterior for θ_1.
- Same steps as before (but nastier) give

$$\Pr_{Y|\theta}\left\{\theta_1^{1-\alpha}(Y) \geq \theta_1\right\} = 1 - \alpha + \frac{\phi(z_\alpha)}{n^{1/2}} T_1'(\pi, \theta) + O(n^{-1}),$$

  where $T_1' \equiv 0$ if and only if

$$\sum_{r=1}^{p} \frac{\partial}{\partial \theta_r} \left\{ i^{11}(\theta)^{-1/2} i^{r1}(\theta) \pi(\theta) \right\} = 0,$$

  where $i^{ab}(\theta)$ is the (a, b) element of $i(\theta)^{-1}$.
- If θ_1 is orthogonal to $(\theta_2, \ldots, \theta_p)$, then $T_1 = 0$ if and only if

$$\pi(\theta) \propto i_{11}^{1/2}(\theta) \times g(\theta_2, \ldots, \theta_p)$$

Discussion: vector θ

- In general cannot orthogonalise all parameters at once, so impossible to obtain matching for all parameters simultaneously—need separate priors for each element of θ
- Higher order matching requires data-dependent priors
- Most prominent alternative default prior is reference prior (Kass and Wasserman, 1996, JASA)

References

- Barndorff-Nielsen and Cox (1989)
- Cox (2006)
- Jeffreys (1961)
- Kass and Wasserman (1996)
- Any modern Bayesian book