Statistical Inference

Anthony Davison

http://stat.epfl.ch

Starting point

- ☐ We start with a concrete question, e.g.,
  - Does the Higgs boson exist?
  - Is fraud taking place at this factory?
  - Are these two satellites likely to collide soon?
  - Do lockdowns reduce Covid transmission?
- ☐ We aim
  - to use **data**
  - to provide **evidence** bearing on the question,
  - to draw a **conclusion** or sometimes reach a **decision**.
- ☐ Here we mostly discuss how to express the evidence, but the choice and quality of the data, and how they were obtained, affect the evidence and the clarity of any decision.
- ☐ The data typically display both **structure** and **haphazard variation**, so any conclusion reached is uncertain, i.e., is an **inference**.
- ☐ A **statistical inference** uses probability models to express the variation in the data.

Theory

- ☐ The goal of statistical theory is to clarify the foundations of statistical inference, which need not involve deep mathematics.
- ☐ Some reasons for studying theory:
  - to clarify underlying issues;
  - to guide strategy for applied work;
  - to elucidate the common structure of diverse applied procedures and thus to formulate general approaches;
  - to transfer ideas between different application domains;
  - to give a basis for developing new methods (including adaptations for specific applied problems)
  - to give a basis (or bases) for comparing competing procedures; and
  - because it's interesting.

Mathematics, asymptotics and all that

- Statistical science is a mathematical science (like physics, ...): problems are often formulated in mathematical terms, but success or failure is judged in consequences for the real world, not in an abstract realm.
- Very often we argue as sample size n (or some other measure of information) becomes large, because
  - exact calculations are impossible or too onerous to be worthwhile;
  - asymptotic arguments clarify the essential structure of a problem (and in particular the necessary assumptions), and thus help generalisation;
  - a procedure that fails when $n = \infty$ will be suspect for finite n.
- Such arguments are often used to generate approximations useful for some fixed finite n, so we should always ask 'will this approximation be adequate in this context?' Often this involves
  - Monte Carlo studies with a realistic sample size;
  - general knowledge (e.g., Laplace approximations generally work better than Edgeworth approximation);
  - previous experience (hard to teach).

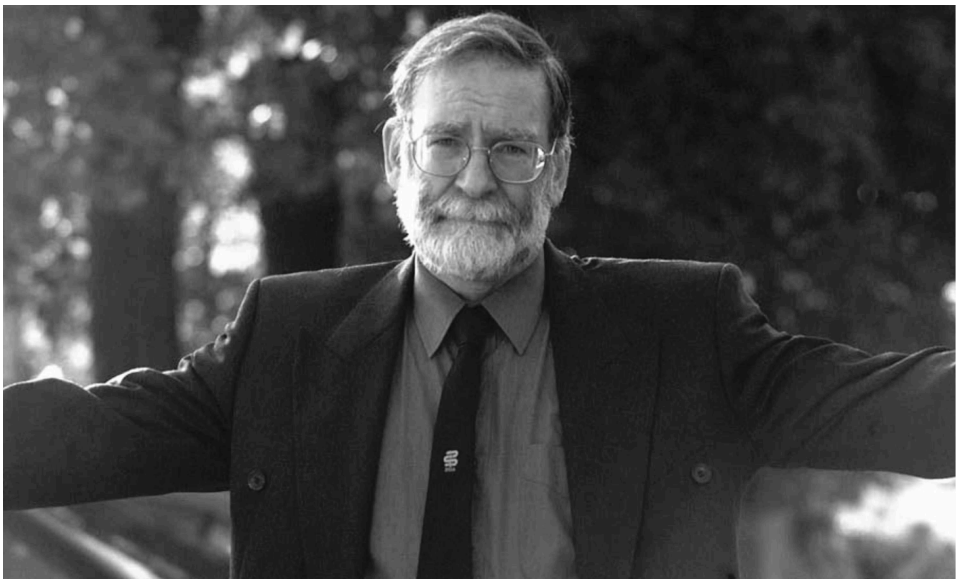
Types of statistical analysis

- Conventionally divided into
  - **descriptive statistics/exploratory data analysis** — informal (often graphical) methods used to gain insight into data, used for initial data analysis and for presenting conclusions;
  - **statistical inference** — more formal approaches using probability models to clarify the support for particular statements about the underlying situation;

    to which we nowadays add

  - **algorithmic methods** — machine learning algorithms, generally complex and computationally demanding, often used for prediction/decision-making.
- Even the first and the last are studied using probability models (e.g., formulation of a boxplot, 'is that difference significant?', analysis of neural networks).
- Need to be clear what type of analysis is being conducted, otherwise conclusions may be horribly biased.

Dr Harold Shipman



Descriptive statistics

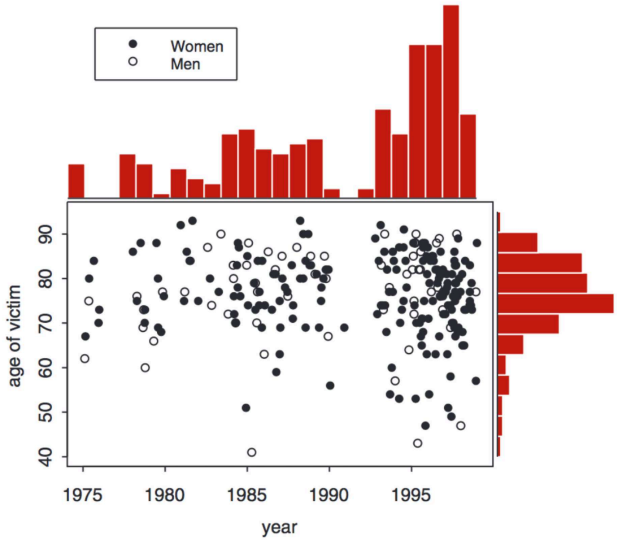
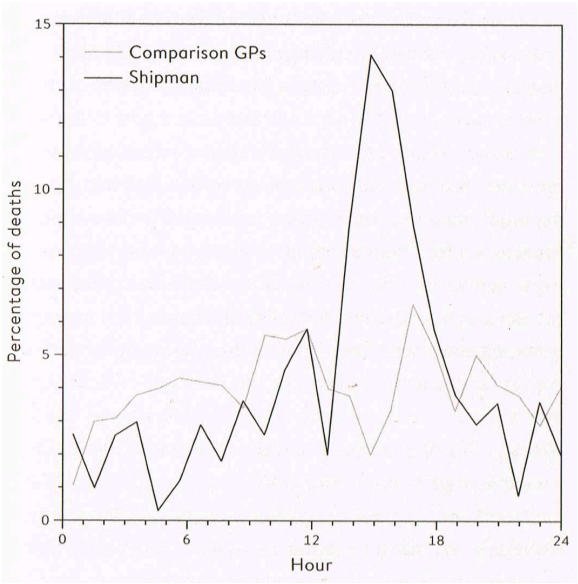


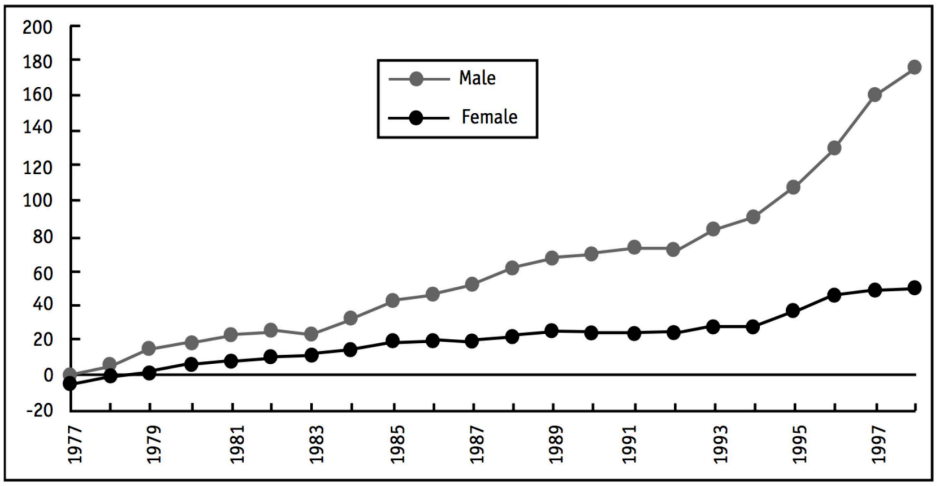
Figure 1. Age and sex of Shipman's victims by date of death

More descriptive statistics



Excess mortality statistics

Figure 2. Cumulative excess death certificates signed by Shipman, for people older than 64 and who died at home or in his practice



Data

- Theoretical discussion generally takes observed data as given, but
  - to get the data we may need to **plan an investigation**, perhaps **design an experiment** largely controlled by the investigator — not considered here but often crucial to obtaining strong data and hence secure conclusions; or
  - to use data from an **observational study** (the investigator has little or no control over data collection).
- In both cases the data used may be selected from those available, and especially if we have 'found data' we must ask
  - why am I seeing these data?
  - what exactly was measured, and how?
  - can the observations actually shed light on the problem?
  - will using a function of the available data give more insight?
- Below we assume that these questions have suitable answers, and we can continue with our analysis.

Example: Satellite conjunction

- Goal is to say whether two 'space objects' will collide, based on repeated noisy measurements of their velocities v and positions p.
- In a 'close conjunction' the problem is usually simplified:
  - only the latest measurements are used,
  - the relative position and velocity of the second object relative to the first are considered, giving $p_2 - p_1$ (km), $v_2 - v_1$ (km/s),
  - the relative motion is taken as rectilinear (Newton I), the objects are represented as spheres and their radii are added to give a 'combined hard-body radius' (HBR),

  so we ask

  based on noisy measurements of $p_2 - p_1$ and $v_2 - v_1$, what is the evidence that the true path of object 2 will pass inside the HBR?
- A simple probability model for the relative position and velocity is

$$\begin{pmatrix} p_2 - p_1 \\ v_2 - v_1 \end{pmatrix} \sim \mathcal{N}_6 \left\{ \begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \Omega & \Omega' \\ \Omega'^{\mathrm{T}} & \Omega'' \end{pmatrix} \right\}$$

  where μ and ν represent the unknown true relative position and velocity vectors and the 3×3 matrices Ω, Ω' and Ω'' are treated as known.

Multivariate normal distribution

A random variable $X_{n\times 1}$ with real components has the **multivariate normal distribution,** $X \sim \mathcal{N}_n(\mu, \Omega)$, if $a^{\mathrm{T}}X \sim \mathcal{N}(a^{\mathrm{T}}\mu, a^{\mathrm{T}}\Omega a)$ for any constant vector $a_{n\times 1}$, and then

- the mean vector and covariance matrix of X are

$$\mathrm{E}(X) = \mu_{n\times 1}, \quad \mathrm{var}(X) = \Omega_{n\times n},$$

  where Ω is symmetric semi-positive definite with real components;

- for any matrix $A_{m\times n}$ and vector $b_{m\times 1}$ of constants,

$$AX + b \sim \mathcal{N}_m\left(A\mu + b, A\Omega A^{\mathrm{T}}\right);$$

- X has a density on $\mathbb{R}^n$ iff Ω is positive definite (i.e., has rank n), and then

$$f(x; \mu, \Omega) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^n; \tag{1}$$

- If $X^{\mathrm{T}} = (X_1^{\mathrm{T}}, X_2^{\mathrm{T}})$, where X_1 is $m \times 1$, and μ and Ω are partitioned correspondingly, then the marginal and conditional distributions of X_1 are also multivariate normal:

$$X_1 \sim \mathcal{N}_m(\mu_1, \Omega_{11}), \quad X_1 \mid X_2 = x_2 \sim \mathcal{N}_m\left\{\mu_1 + \Omega_{12}\Omega_{22}^{-1}(x_2 - \mu_2), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right\}.$$

Example: Satellite conjunction II

- When the velocity measurements can be treated as noiseless compared to the position measurements, i.e., $\nu = v_2 - v_1$ (equivalently $\Omega' = \Omega'' = 0$), then we can write

$$X = p_2 - p_1 \sim \mathcal{N}_3(\mu, \Omega), \quad \psi_{\min} = \mathrm{HBR} > 0,$$

  and object 2 traverses the line $\{\mu + t\nu : t \geq 0\}$.

- We use a matrix $A_{2\times 3}$ to project X and μ into the plane $\mathcal{P}$ perpendicular to ν and passing through the origin, and obtain the two-dimensional model

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = Y = AX \sim \mathcal{N}_2(\xi, D^{-1}), \quad D^{-1} = A\Omega A^{\mathrm{T}} = \begin{pmatrix} d_1^{-1} & 0 \\ 0 & d_2^{-1} \end{pmatrix},$$

  say, with known d_1, d_2 (as Ω is known and A depends on the known ν), and unknown

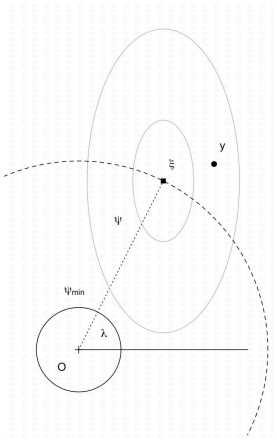
$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \psi\cos\lambda \\ \psi\sin\lambda \end{pmatrix}, \quad \psi > 0, 0 \leq \lambda < 2\pi.$$

- We observe a value x of X and ask

  based on the projection $y = Ax$ of x into $\mathcal{P}$, what is the evidence that $\xi = A\mu$ lies inside the HBR, or equivalently that $\psi = \|\xi\| < \psi_{\min}$?

Example: Satellite conjunction III

- We seek inferences about the scalar $\psi = \|\xi\|$ based on the bivariate observation y.
- A natural estimate of ψ is $\|y\| = (y_1^2 + y_2^2)^{1/2}$, but geometrically it is obvious that $\mathrm{P}(\|y\| > \psi) > 1/2$, i.e., $\|y\|$ will tend to be larger than ψ.
- The sample size is $n = 1$, and any asymptotics are as $d_1, d_2 \to \infty$, i.e., $\mathrm{var}(Y) \to 0$.



Some definitions

- We use the term **(probability) density** for both the PMF of a discrete random variable, and the PDF of a continuous random variable.
- A **statistical model** for data y is a probability density $f(y)$ defined for $y \in \mathcal{Y}$.
- A **parametric model** $f \equiv f(y; \theta)$ is determined by **parameters** $\theta \in \Theta \subset \mathbb{R}^d$. If no such θ exists, f is **nonparametric**.
- Sometimes we use the term **family of models** to stress that there are many possibilities, $\{f(y; \theta) : \theta \in \Theta\}$.
- Generally $\theta = (\psi, \lambda)$ splits into
  - **parameters of interest (interest parameters)** ψ, usually scalar, that we focus on,
  - **nuisance parameters** λ, usually vector, needed to complete the model but not of main interest,

  though different elements of θ may become of interest during an investigation.
- By convention we (try to) use
  - letters like $c, d, \ldots$ for (known) constants,
  - Roman letters for random variables $X, Y, \ldots$ and their realisations $x, y, \ldots$,
  - Greek letters $\mu, \nu, \psi, \lambda, \Omega, \Delta, \ldots$ for unknown parameters.

Some shortcuts

- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ means that the X_j are independent and all have the density f, and we call the X_j a **random sample of size n from f**.
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} f_1, \ldots, f_n$ means that the X_j are independent and $X_j \sim f_j$.
- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ means that the X_j are independent with mean μ and variance σ^2 (we assume that $0 < \sigma^2 < \infty$). Here we do not assume that the X_j are normal.
- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2)$ means that the X_j are independent with means μ_j and variances σ_j^2 (where $0 < \sigma_j^2 < \infty$).
- We write $\mathcal{N}(0,1)$ for the standard normal distribution, which has density and distribution functions

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad u \in \mathbb{R}, \qquad \Phi(z) = \int_{-\infty}^{z} \phi(u)\,\mathrm{d}u, \quad z \in \mathbb{R}.$$

  and write z_p for the p quantile of the standard normal distribution, i.e., $\Phi(z_p) = p$, where $0 < p < 1$.

Example: Satellite conjunction IV

- If $Y \sim \mathcal{N}_n(\mu, \Omega)$, then $\theta = (\mu, \Omega) \in \Theta \subset \mathbb{R}^n \times \mathcal{S}$, where $\mathcal{S}$ denotes the set of all $n \times n$ symmetric positive definite matrices, so this is a parametric model.
- In the satellite example, we assume that the observed position y is a realisation of

$$Y \sim \mathcal{N}_2 \left\{ \begin{pmatrix} \psi \cos\lambda \\ \psi \sin\lambda \end{pmatrix}, \begin{pmatrix} d_1^{-1} & 0 \\ 0 & d_2^{-1} \end{pmatrix} \right\},$$

  where $d_1, d_2 > 0$, $\theta = (\psi, \lambda) \in \Theta = \mathbb{R}_+ \times [0, 2\pi)$, the density function is

$$\frac{(d_1 d_2)^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2} \left\{ d_1 (y_1 - \psi\cos\lambda)^2 + d_2 (y_2 - \psi \sin\lambda)^2 \right\} \right], \quad y_1, y_2 \in \mathbb{R}.$$

- Given the constants d_1, d_2 and observed data y_1, y_2,
  - the interest parameter is the length ψ of ξ,
  - the nuisance parameter is the angle λ, and

  we want to know the evidence that $\psi < \psi_{\min}$.
- A natural estimate of ψ is $\|y\| = (y_1^2 + y_2^2)^{1/2}$, but this will tend to be too long.

Objectives

- Given a specified family of models $f(y;\theta)$ with $\theta = (\psi, \lambda)$ and observed data y, the objectives of a statistical analysis might be to
  - give intervals (or more generally sets) of values within which ψ is 'likely to lie';
  - assess the consistency of y with a particular ψ_0;
  - predict as-yet unobserved random variables from the system that generated y;
  - use y to choose one of a specified set of decisions — requires the specification of the consequences of the decisions.
- We might also seek to check whether the family of models adequately represents the data.

Probability

- Two distinct roles of probability in statistical analysis:
  - as a description of **variation** in data ('aleatory probability', 'chance'), using a probability model and treating the observed data y as an outcome of that model;
  - to formulate **uncertainty** ('epistemic probability') about the reality modelled in terms of the random experiment, based on y.

Probability models for variation

- Two broad types of probability model:
  - **substantive** — based on fundamental subject-matter theory (e.g., particle physics, Mendelian genetics, Navier–Stokes equations);
  - **empirical** — a convenient, adequately realistic, representation of data variation;
  - and of course there is a spectrum between them.
- The satellite example is partly substantive (using Newtonian mechanics for the ideal trajectory $\mu + t\nu$) and partly empirical (normal distributions for measurement error).
- We aim that
  - primary questions/issues are encapsulated in the parameter(s) of interest;
  - secondary aspects can be taken into account, often via nuisance parameters;
  - variation in the data is realistically modelled, leading to reasonable statements of uncertainty;
  - any special feature of the data collection process is represented;
  - different approaches to analysis can if necessary be compared.
- Such models are always provisional and should if possible be checked against data.

Uncertainty

- Essentially three bases for statements of uncertainty:
  - a **Bayesian (inverse probability) inference** expresses it via a prior probability density and uses Bayes' theorem to update this in light of the data;
  - a **frequentist (sampling theory) inference** compares y with the set $\mathcal{S}$ of other data that might have been observed in a hypothetical sampling experiment;
  - in a designed experiment, clinical trial, sample survey or similar the investigator uses **randomisation** to generate a distribution against which y is compared.
- There are many variants of the first two approaches.
- In particular, a frequentist should choose $\mathcal{S}$ thoughtfully:
  - what doctors would be a suitable comparison group for Shipman?

Example 1 (Measuring machines) *A physical quantity θ can be measured with two machines, both giving normal observations Y such that $\mathrm{E}(Y) = \theta$. A measurement from machine 1 has variance 1, and one from machine 2 has variance 100. A machine is chosen by tossing a fair coin, giving $M = 1, 2$ with equal probabilities.*
If we observe $m = 1$ and $y = 2$, then clearly we can ignore the fact that we might have observed $m = 2$, i.e., we should take $\mathcal{S} = \{(y, 1) : y \in \mathbb{R}\}$ rather than $\mathcal{S} = \{(y, m) : y \in \mathbb{R}, m \in \{1, 2\}\}$.

Bayesian inference

- Our observed data y^{o} are assumed to be a realisation from a density $f(y \mid \psi)$.
- If we can summarise information about ψ, separately from y^{o}, in a **prior density** $f(\psi)$, then we can use Bayes' theorem to obtain the **posterior density**

$$f(\psi \mid y^{\mathrm{o}}) = \frac{f(y^{\mathrm{o}} \mid \psi) f(\psi)}{\int f(y^{\mathrm{o}} \mid \psi) f(\psi) \, \mathrm{d}\psi},$$

  and base all our uncertainty statements on this.
- For example, if ψ_p satisfies $\mathrm{P}(\psi \leq \psi_p \mid y^{\mathrm{o}}) = p$ for any $p \in (0, 1)$, we could give a $(1 - 2\alpha)$ **posterior credible interval** $\mathcal{I}_{1-2\alpha} = (\psi_\alpha, \psi_{1-\alpha})$ such that

$$\mathrm{P}(\psi \in \mathcal{I}_{1-\alpha} \mid y^{\mathrm{o}}) = 1 - 2\alpha;$$

  here ψ is regarded as random and y^{o} as fixed.
- Likewise if there is a nuisance parameter, we require a prior density $f(\psi, \lambda)$ and compute the **marginal posterior density of ψ**,

$$f(\psi \mid y^{\mathrm{o}}) = \frac{\int f(y^{\mathrm{o}} \mid \psi, \lambda) f(\psi, \lambda) \, \mathrm{d}\lambda}{\iint f(y^{\mathrm{o}} \mid \psi, \lambda) f(\psi, \lambda) \, \mathrm{d}\lambda \mathrm{d}\psi}.$$

Example

Example 2 *In the satellite problem a natural prior density for $\xi \in \mathbb{R}^2$ is the* **improper prior**

$$f(\xi_1, \xi_2) \propto c > 0, \quad \xi_1, \xi_2 \in \mathbb{R},$$

used to express total ignorance about ξ, and if $\mathcal{D} = \{(\xi_1, \xi_2) : \xi_1^2 + \xi_2^2 \leq \psi_{\min}\}$, then

$$\mathrm{P}(\psi < \psi_{\min} \mid y^{\mathrm{o}}) = \iint_{\mathcal{D}} \frac{(d_1 d_2)^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2}\left\{d_1(y_1^{\mathrm{o}} - \xi_1)^2 + d_2(y_2^{\mathrm{o}} - \xi_2)^2\right\}\right] \mathrm{d}\xi_1 \mathrm{d}\xi_2$$

also equals (in polar coordinates)

$$\int_0^{\psi_{\min}} \int_0^{2\pi} \frac{(d_1 d_2)^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2}\left\{d_1(y_1^{\mathrm{o}} - \psi\cos\lambda)^2 + d_2(y_2^{\mathrm{o}} - \psi\sin\lambda)^2\right\}\right] \psi \, \mathrm{d}\lambda \mathrm{d}\psi$$

corresponding to the prior

$$f(\psi, \lambda) \propto \psi > 0, \quad \psi > 0, 0 \leq \lambda < 2\pi,$$

on the polar coordinates.
How does this prior express ignorance about ξ?

Note to Example 2

- The posterior density for ξ_1, ξ_2 is proportional to

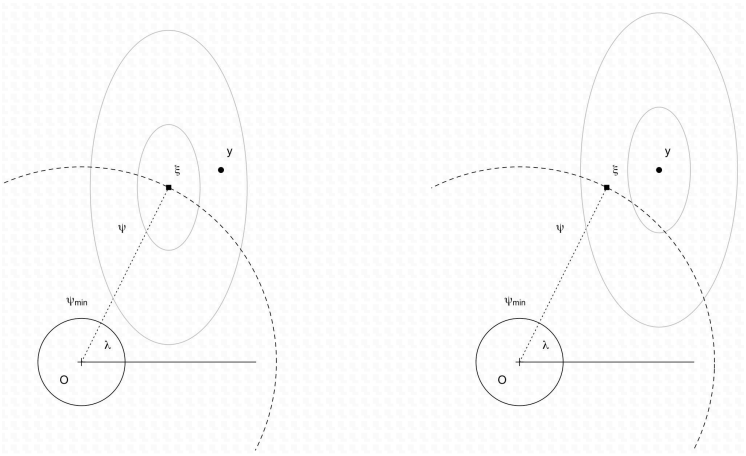
$$c \times \frac{(d_1 d_2)^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2}\left\{d_1(y_1^{\mathrm{o}} - \xi_1)^2 + d_2(y_2^{\mathrm{o}} - \xi_2)^2\right\}\right],$$

  and the constant of proportionality is just the integral of this expression with respect to ξ_1, ξ_2. By symmetry of the quadratic form in the exponent, this is just c, so the posterior density for ξ_1, ξ_2 is just

$$\frac{(d_1 d_2)^{1/2}}{2\pi} \exp\left[-\tfrac{1}{2}\left\{d_1(\xi_1 - y_1^{\mathrm{o}})^2 + d_2(\xi_2 - y_2^{\mathrm{o}})^2\right\}\right], \quad \xi_1, \xi_2 \in \mathbb{R},$$

  i.e., $\xi \mid y^o \sim \mathcal{N}_2(y^{\mathrm{o}}, D^{-1})$. Compare the densities on the next slide to see how the sampling model for $y \mid \xi$ on the left is updated by Bayes' theorem to the posterior density of $\xi \mid y = y^{\mathrm{o}}$ on the right.
- The Jacobian for changing from the Cartesian coordinates ξ_1, ξ_2 to the polar coordinates $\psi = (\xi_1^2 + \xi_2^2)^{1/2}$, $\lambda = \tan^{-1}(\xi_2/\xi_2)$ is ψ, which gives the second expression.
- As the prior is uniform the prior probability that ξ lies in any set is the ratio of the measure of that set to the measure of $\mathbb{R}^2$, and thus is zero for any bounded set, and thus in particular for any disk around the origin.

Conjunction



Left: sampling model $y \mid \xi \sim \mathcal{N}_2(\xi, D^{-1})$. Right: posterior density $\xi \mid y \sim \mathcal{N}_2(y, D^{-1})$ based on the constant prior.

Comments

- ☐ Bayesian inference
  - requires the specification of a prior distribution on unknowns, separate from the data;
  - implies that we regard prior information as equivalent to data, putting uncertainty and variation on the same footing;
  - reduces inference to computation of probabilities, so in principle is simple and direct.
- ☐ Specifying prior 'ignorance' in an objective way is problematic and can lead to paradoxes, especially in high-dimensional settings.
- ☐ (Approximate) Bayesian computation can be performed using
  - conjugate prior distributions (exact computations in simple cases),
  - integral approximations (e.g., Laplace's method),
  - deterministic methods (e.g., variational approximation),
  - simulation, especially Markov chain Monte Carlo.

Sampling theory

- ☐ Sampling theory inference treats the observed data y^{o} as a realisation from some model $f(y;\theta)$, and calculates probabilities using hypothetical samples from $f(y;\theta)$.
- ☐ We assess the plausibility of different values of θ by asking how well they explain y^{o}, often using a pivot.

Definition 3 *If Y has density $f(y;\theta)$, then a* **pivot (or pivotal quantity)** *$Q = q(Y,\theta)$ is a function of Y and θ that has a known distribution (i.e., does not depend on θ). Often it is convenient if Q is monotone in θ for each Y.*

Example 4 *If $M = \max(Y_1,\ldots,Y_n)$, where $Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} U(0,\theta)$, show that $Q = M/\theta$ is a pivot.*

Note to Example 4

Q_1 is a function of the data and the parameter, and

$$\mathrm{P}(M \le x) = F_Y(x)^n = (x/\theta)^n, \quad 0 < x < \theta,$$

so

$$\mathrm{P}(Q_1 \le q) = \mathrm{P}(M/\theta \le q) = \mathrm{P}(M \le \theta q) = (\theta q/\theta)^n = q^n, \quad 0 < q < 1.$$

which is known and does not depend on θ. Hence Q_1 is a pivot.

Confidence intervals

Definition 5 *Let $Y = (Y_1,\ldots,Y_n)$ be data from a parametric statistical model with scalar parameter θ. A* **confidence interval (CI)** *(L,U)* **for** *θ with lower confidence bound L and upper confidence bound U is a random interval that contains θ with a specified probability, called the* **(confidence) level** *of the interval.*

- ☐ $L = l(Y)$ and $U = u(Y)$ are statistics that can be computed from the data. They do not depend on θ.
- ☐ In a continuous setting (so $<$ gives the same probabilities as $\le$), and if we write the probabilities that θ lies below and above the interval as

$$\mathrm{P}(\theta < L) = \alpha_L, \quad \mathrm{P}(U < \theta) = \alpha_U,$$

then (L,U) has confidence level

$$\mathrm{P}(L \le \theta \le U) = 1 - \mathrm{P}(\theta < L) - \mathrm{P}(U < \theta) = 1 - \alpha_L - \alpha_U.$$

- ☐ Often we seek an interval with equal probabilities of not containing θ at each end, with $\alpha_L = \alpha_U = \alpha/2$, giving an **equi-tailed** $(1-\alpha)\times 100\%$ **confidence interval**.
- ☐ We often take standard values of α, such that $1-\alpha = 0.9, 0.95, 0.99, \ldots$

Construction of a CI

- We use pivots to construct CIs:
  - we find a pivot $Q = q(Y, \theta)$ involving θ;
  - we obtain the quantiles q_{α_U}, $q_{1-\alpha_L}$ of Q;
  - then we transform the equation

$$\mathrm{P}\{q_{\alpha_U} \leq q(Y, \theta) \leq q_{1-\alpha_L}\} = (1 - \alpha_L) - \alpha_U$$

    into the form

$$\mathrm{P}(L \leq \theta \leq U) = 1 - \alpha_L - \alpha_U,$$

    where the bounds L, U depend on Y, $q_{1-\alpha_L}$ and q_{α_U}, but not on θ.
- Going from quantiles of Q to confidence limits for θ is known as **inverting the pivot**.
- In many cases, the bounds are of a standard form (see below).

Example 6 *In Example 4, find a CI based on Q.*

Note to Example 6

The p quantile of $Q_1 = M/\theta$ is given by $p = \mathrm{P}(Q_1 \leq q_p) = q_p^n$, so $q_p = p^{1/n}$. Thus

$$\mathrm{P}\{\alpha_U^{1/n} \leq M/\theta \leq (1 - \alpha_L)^{1/n}\} = 1 - \alpha_L - \alpha_U,$$

and a little algebra gives that

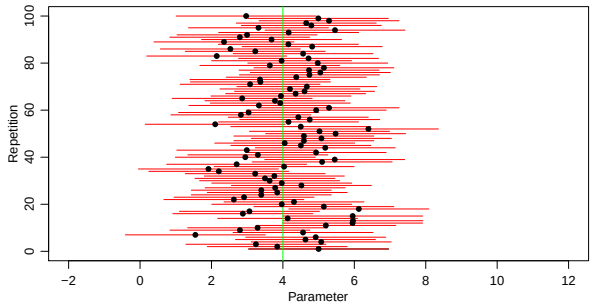
$$\mathrm{P}\{M/(1 - \alpha_L)^{1/n} \leq \theta \leq M/\alpha_U^{1/n}\} = 1 - \alpha_L - \alpha_U,$$

so

$$L = M/(1 - \alpha_L)^{1/n}, \quad U = M/\alpha_U^{1/n}.$$

Interpretation of a CI

- (L, U) is a random interval that contains θ with probability $1 - \alpha$.
- We imagine an infinity of possible datasets from the experiment that resulted in (L, U).
- Our particular CI is regarded as randomly chosen from the corresponding infinity of CIs.
- Although we do not know whether our particular CI contains θ, the event $\theta \in (L, U)$ has probability $1 - \alpha$ across these datasets.
- In the figure below, the parameter θ (green line) is contained (or not) in realisations of the 95% CI (red). The black points show the corresponding estimates.



One- and two-sided intervals

- A **two-sided confidence interval** (L, U) is generally used, but **one-sided confidence intervals**, $(-\infty, U)$ or (L, ∞), are sometimes required instead.
- For one-sided CIs, we take $\alpha_U = 0$ or $\alpha_L = 0$, giving (L, ∞) or $(-\infty, U)$.
- For a one-sided $(1 - \alpha) \times 100\%$ interval, we compute a two-sided interval with $\alpha_L = \alpha_U = \alpha$, and then replace the unwanted limit by $\pm\infty$ (or another value if required in the context).

Comments

- We assume that y^{o} is just one of many possible datasets $y \in \mathcal{S}$ that might have been generated from $f(y;\theta)$, and the probability calculations are with respect to $\mathcal{S}$.
- We choose the **reference set** $\mathcal{S}$ to ensure that the probability calculation is **relevant** to the data actually observed. For example, if y^{o} has n observations, we usually insist that every element of $\mathcal{S}$ also has n observations.
- The repeated sampling principle ensures that (if we use an exact pivot) inferences are **calibrated**, for example, a $(1-\alpha)$ confidence interval (L, U) satisfies

$$\mathrm{P}(L < \theta \leq U) = 1 - \alpha,$$

  for every $\theta \in \Theta$ and every $\alpha \in (0, 1)$. Hence if such an interval is used repeatedly, then the probability it does not contain θ is exactly α.
- Calibration guarantees that the procedure, if repeated, has the stated error probability, and any particular interval either does or does not contain θ.
- Bayesians object that inferences should only be based on the dataset y^{o} actually observed, so the reference set $\mathcal{S}$ is irrelevant.

Example 7 *What would the confidence intervals look like in Example 1? How would the image on slide 30 change? What hypothetical repetitions form the reference set?*

Randomisation

- To compare how **treatments** affect a **response**, they are **randomised** to experimental **units**:
  - **treatments** are clearly-defined procedures, one of which is applied to each unit;
  - a **unit** is the smallest division of the raw material such that two different units might receive two different treatments;
  - the **response** is a well-defined variable measured for each unit-treatment combination.
- Examples are agricultural trials, industrial experiments, clinical trials, ...
- The experiment is 'under the control' of the investigator, making strong inferences possible.
- Main goals of randomisation:
  - avoidance of systematic error (eliminating bias);
  - estimation of baseline variation (e.g., by use of replication and/or blocking);
  - realistic statement of uncertainty of final conclusions;
  - providing a basis for exact inferences using the randomisation distribution.

Example: Shoe data

- Shoe wear in an paired comparison experiment in which materials A (expensive) and B (cheaper) were randomly assigned to the soles of the left (L) or right (R) shoe of each of $m = 10$ boys.
- The $m = 10$ differences $d_1, \ldots, d_m$ have average $\overline{d} = 0.41$.

Boy	Material		Difference
	A	B	d
1	13.2 (L)	14.0 (R)	0.8
2	8.2 (L)	8.8 (R)	0.6
3	10.9 (R)	11.2 (L)	0.3
4	14.3 (L)	14.2 (R)	–0.1
5	10.7 (R)	11.8 (L)	1.1
6	6.6 (L)	6.4 (R)	–0.2
7	9.5 (L)	9.8 (R)	0.3
8	10.8 (L)	11.3 (R)	0.5
9	8.8 (R)	9.3 (L)	0.5
10	13.3 (L)	13.6 (R)	0.3

Example: Shoe data II

- A unit is a foot, a treatment is the type of sole, and the response is the amount of wear.
- This is **paired comparison** experiment, as there are **blocks** of two similar units, each of which is given one treatment at random, according to the scheme

Treatment for boy j	Left foot	Right foot
A	l_j	r_j
B	$\psi + l_j$	$\psi + r_j$

- We observe either $(\psi + l_j, r_j)$ or $(l_j, r_j + \psi)$ so the difference D_j of B and A for boy j is $\psi + l_j - r_j$ or $\psi + r_j - l_j$. These are equally likely, so we can write $D_j = \psi + I_j c_j$, where
  - ψ is the unknown (extra wear) effect of B compared to A,
  - $I_j = 1$ if the left shoe of boy j has material B and otherwise equals -1, and
  - $c_j = l_j - r_j$ is the unobserved baseline difference in wear between the left and right feet of boy j.
- If we observe $(\psi + l_j, r_j)$ for boy j, then we cannot observe $(l_j, \psi + r_j)$, which is said to be **counterfactual**.

Example: Shoe data III

- There are 2^m equally-likely treatment allocations, and the observed $\overline{d}$ is a realisation of the random variable

$$\overline{D} = \frac{1}{m}\sum_{j=1}^{m} D_j = \frac{1}{m}\sum_{j=1}^{m} \psi + I_j c_j = \psi + \frac{1}{m}\sum_{j=1}^{m} I_j c_j,$$

  where $I_j = \pm 1$ with equal probabilities, so

$$\mathrm{E}(I_j) = 0, \quad \mathrm{var}(I_j) = 1.$$

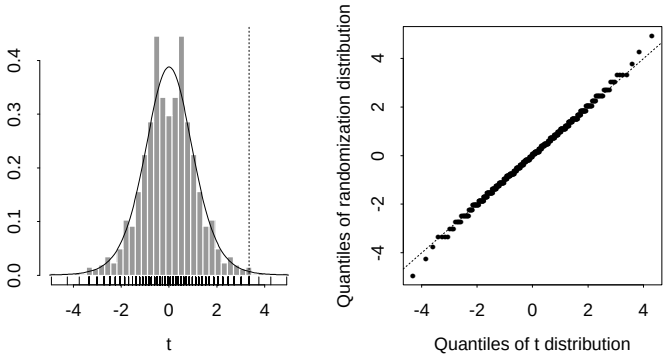
- Hence $\mathrm{E}(\overline{D}) = \psi$ and $\mathrm{var}(\overline{D}) = m^{-2}\sum_{j=1}^{m} c_j^2$, which is unknown because the c_j are unknown, is estimated by (exercise)

$$S^2 = \frac{1}{m(m-1)}\sum_{j=1}^{m}(D_j - \overline{D})^2.$$

- $\overline{D}$ and S^2 can be computed from the observed data, so the standardized quantity $Z = (\overline{D} - \psi)/S$ is an approximate pivot.
- If there was no difference between B and A (i.e., $\psi = 0$), then $T = \overline{D}/S$ would be symmetrically distributed, as positive and negative values of $\overline{D}$ would be equally likely.

Example: Shoe data IV

Randomization distribution of $T = \overline{D}/S$ for the shoes data, i.e., setting $\psi = 0$, together with a t_9 distribution. Left: histogram and rug for the values of T, with the t_9 density overlaid; the observed value is given by the vertical dotted line. Right: probability plot of the randomization distribution against t_9 quantiles.



Comments

- **Systematic error** is reduced by randomisation,
  - but if material A had by chance been allocated to all the left feet, then we might have re-randomised;
  - we could have used a design in which A appeared on left feet exactly 5 times.
- **Baseline variation** was reduced by blocking, i.e., using two treatments for each boy, and is estimated by S^2, based only on the observed values $D_1, \ldots, D_m$.
- S^2 also allows a statement of **uncertainty** for $\overline{D}$ and hence for estimates of ψ.
- If $\psi = 0$, then the observed value of $\overline{D}$ is highly unlikely: just 3 values of $\overline{D}$ exceed $\overline{d} = 0.41$, so if $\psi = 0$ then **exact calculation** gives

$$\mathrm{P}(\overline{D} \geq \overline{d}) = 7/2^{10} \doteq 0.007,$$

  which seems unlikely enough to suggest that $\psi > 0$.
- Normal distribution theory suggests that $Z \stackrel{\cdot}{\sim} t_9$, and the QQ-plot shows that this would work well even here. The symmetry induced by randomisation justifies the widespread use of normal errors in designed experiments.

Wrapping up

- Statistical inference involves (a family of) **probability models** from which observed data are assumed to be drawn.
- These models express **variation** inherent in the data, but we also wish to express our **uncertainty** about the underlying situation.
- Uncertainty is formulated using
  - a **Bayesian approach**, which requires that 'prior information' on unknown quantities be expressed as a probability distribution, or
  - a **repeated sampling (frequentist) approach**, which invokes hypothetical repetitions of the data-generating mechanism, or
  - a **randomisation approach**, in which the model and hypothetical repetitions are controlled by the investigator.
- The last is the strongest approach, but it is not always applicable.

Likelihood

- Given observed data y thought to come from a parametric model $f_Y(y;\theta)$ for which $\theta \in \Theta$, the **likelihood** and the **log likelihood** are

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

  we regard these as functions of θ for fixed y. The log likelihood is often more convenient to work with because if y consists of independent observations $y_1, \ldots, y_n$, then

$$\ell(\theta) = \log f_Y(y;\theta) = \log \prod_{j=1}^{n} f(y_j;\theta) = \sum_{j=1}^{n} \log f(y_j;\theta), \quad \theta \in \Theta,$$

  so laws of large numbers and other limiting results apply directly to $n^{-1}\ell(\theta)$.
- The posterior density based on data y and prior $f(\theta)$ is proportional to $L(\theta) \times f(\theta)$.

Likelihood quantities

- The **maximum likelihood estimate (MLE)** $\widehat{\theta}$ satisfies

$$\ell(\widehat{\theta}) \geq \ell(\theta) \quad \text{or equivalently} \quad L(\widehat{\theta}) \geq L(\theta), \qquad \theta \in \Theta.$$

- Often $\widehat{\theta}$ is unique and satisfies the **score (or likelihood) equation**

$$\nabla \ell(\theta) = \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = 0,$$

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

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

  these are $d \times d$ matrices if θ has dimension d and otherwise are scalars.
- To evaluate $I(\theta)$ we replace y by the random variable Y and take expectations.

Example 8 (Poisson sample) *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{Pois}(\theta)$, *for* $\theta > 0$, *find the log likelihood, the MLE, the observed information and the Fisher information.*

Note to Example 8

- The log likelihood is

$$\ell(\theta) = \sum_{j=1}^{n} \log f(y_j;\theta) = \sum_{j=1}^{n} \log(\theta^{y_j} e^{-\theta}/y_j!) \equiv s\log\theta - n\theta, , \quad \theta > 0,$$

  where $s = \sum_j y_j$ and $\equiv$ means that we have dropped additive constants from the log likelihood.
- It follows that $\ell'(\theta) = s/\theta - n$ and $\ell''(\theta) = -s/\theta^2$, so provided $s > 0$ we have $\widehat{\theta} = s/n = \overline{y}$ and $J(\theta) = s/\theta^2$. If $s = 0$ then a sketch of $\ell(\theta)$ shows that $\widehat{\theta} = 0 = s/n$, which is on the boundary of the parameter space. Also $\ell''(\widehat{\theta})$ is only defined as a limit from the right, and equals zero.
- As $\mathrm{E}(S) = n\mathrm{E}(Y_j) = n\theta$, the Fisher information is computed as

$$\mathrm{E}(S/\theta^2) = n/\theta.$$

Invariance

- Seek invariance to (smooth) 1–1 transformations of data and/or parameter.
- If $Z = z(Y)$ is a 1–1 function of a continuous variable Y and the transformation does not depend on θ, then $f_Z(z;\theta) = f_Y\{y^{-1}(z);\theta\}|\mathrm{d}y/\mathrm{d}z|$, so (in an explicit notation)

$$\ell(\theta; z) = \log f_Z(z;\theta) \equiv \ell(\theta; y) = \log f_Y(y;\theta),$$

  where $\equiv$ means that an additive constant not depending on θ has been dropped — hence likelihood inference is the same whether we use Y or Z.
- Likewise a smooth 1–1 transformation from θ to $\phi(\theta)$ will give

$$\tilde{f}(y;\phi) = \tilde{f}\{y;\phi(\theta)\} = f(y;\theta),$$

  where the tilde denotes the density expressed using ϕ. Clearly

$$\tilde{f}(y;\widehat{\phi}) = \tilde{f}\{y;\phi(\widehat{\theta})\} = f(y;\widehat{\theta}), \quad J(\widehat{\theta}) = \frac{\partial\phi^{\mathrm{T}}}{\partial\theta}\tilde{J}(\phi)\frac{\partial\phi}{\partial\theta^{\mathrm{T}}}\bigg|_{\phi=\phi(\widehat{\theta})},$$

  so the respective maximum likelihood estimates satisfy $\widehat{\phi} = \phi(\widehat{\theta})$.
- If possible inferences on ψ should be invariant to **interest-respecting (or interest-preserving) transformations**

$$\psi, \lambda \quad \mapsto \quad \eta = \eta(\psi), \zeta = \zeta(\psi, \lambda).$$

Sufficiency

- A statistic $S = s(Y)$ is **sufficient (for θ)** under a model $f_Y(y;\theta)$ if the conditional density $f_{Y|S}(y \mid s;\theta)$ is independent of θ for any θ and s.
- This implies that
$$f_Y(y;\theta) = f_S(s;\theta) f_{Y|S}(y \mid s), \quad \ell(\theta;s) \equiv \ell(\theta;y),$$
so we can regard s as containing all the sample information about θ: if we consider Y to be generated in two steps,
  - first generate S from $f_S(s;\theta)$, and
  - then generate Y from $f_{Y|S}(y \mid s)$,

  we see that if the model holds, then the second step gives no information about θ, so we could stop after the first step.
- The conditional distribution $f_{Y|S}(y \mid s)$ allows assessment of the model without reference to θ.

Minimal sufficiency

- If $S = s(Y)$ is sufficient and $T = t(Y)$ is any other function of Y, then (S,T) contains at least as much information as S, and is also sufficient.
- To define a 'smallest' sufficient function of Y, we define a **minimal sufficient statistic** to be a function of any other sufficient statistic. This is unique up to 1–1 maps.
- Below we always use minimal sufficient statistics, identified by finding the 'lowest-dimensional function of the data' that allows us to plot the log likelihood.

Example 9 (Poisson sample) *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Pois}(\theta)$, for $\theta > 0$.*

- *Obtain a sufficient statistic and discuss how to assess the fit of the model.*
- *How does the sufficient statistic change if the sample size n is a realisation of a geometric random variable with success probability θ?*

Example 10 (Location model) *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g(y-\theta)$, with g a known continuous density, find a sufficient statistic.*

Example 11 (Uniform model) *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(\theta)$, sketch the likelihood, and find a sufficient statistic and the MLE of θ.*

Note to Example 9

- Here $Y = (Y_1, \ldots, Y_n)$ so

$$f_Y(y;\theta) = \prod_{j=1}^{n} \frac{\theta^{y_j}}{y_j!} e^{-\theta} = m(y)\theta^s e^{-n\theta},$$

  where $m(y) = 1/\prod y_j!$, and clearly we need only know n and $s = \sum_j y_j$, and of these only $s = s(y)$ depends on y. Hence $S = \sum_j Y_j$, which has a Poisson distribution with mean $n\theta$, looks like a suitable sufficient statistic. We have

$$f_{Y|S}(y \mid s;\theta) = \frac{f_Y(y;\theta)}{f_S(s;\theta)} = \frac{m(y)\theta^s e^{-n\theta}}{(n\theta)^s e^{-n\theta}/s!} = \frac{s!}{y_1! \cdots y_n!} n^{-s}, \quad y \in \mathcal{Y}_s,$$

  where $\mathcal{Y}_s = \{y : y_1, \ldots, y_n \in \{0, \ldots, s\}, \sum_j y_j = s\}$. This conditional distribution does not depend on θ, so S is sufficient for θ.
- The conditional distribution is in fact a multinomial distribution with denominator s and probability vector $(1/n, \ldots, 1/n)$, and would be used for testing the fit of the Poisson model.
- When N is geometric with success probability $\theta \in (0, 1)$, the likelihood becomes

$$f(y, n;\theta) = f(y \mid n;\theta) f(n;\theta) = \prod_{j=1}^{n} \frac{\theta^{y_j}}{y_j!} e^{-\theta} \times (1-\theta)^{n-1}\theta = m(y)(1-\theta)^{n-1}\theta^{s+1} e^{-n\theta},$$

  and now we see that we would need both n and s to sketch the likelihood. Hence the sufficient statistic is now (n, s), which is two-dimensional, although there is only one parameter. This is because the value of n now contains information about θ: very large n will suggest that θ is small.

Note to Example 10

- The density g is continuous, so all the y_j are distinct with probability one. The joint density is therefore

$$f(y;\theta) = \prod_{j=1}^{n} g(y_j - \theta) = n! \prod_{j=1}^{n} g(y_{(j)} - \theta), \quad y_{(1)} < \cdots < y_{(n)},$$

  where $s = (y_{(1)}, \ldots, y_{(n)})$ are the sample order statistics. The labels on the original data are simply a permutation of the n labels on the order statistics, but the values are the same, so

$$f(y \mid s;\theta) = \frac{f(y;\theta)}{f(s;\theta)} = \frac{1}{n!}, \quad y \in \mathcal{Y}_s,$$

  where $\mathcal{Y}_s$ is the set of permutations of $(y_1, \ldots, y_n)$ with order statistics s; clearly $|\mathcal{Y}_s| = n!$, because there are no ties.
- Here $|s| = n$ in general. In special cases (e.g., the normal distribution, $g = \phi$) we can find a sufficient statistic of of lower dimension.

Note to Example 11

- The density is $f(y;\theta) = \theta^{-1} I(0 < y < \theta)$, so since the observations are independent, the likelihood is

$$L(\theta) = \prod_{j=1}^{n} \theta^{-1} I(0 < y_j < \theta) = \theta^{-n} I(0 < y_1, \ldots, y_n < \theta) = \theta^{-n} I(\theta > m), \quad \theta > 0,$$

  where $m = \max(y_1, \ldots, y_n)$; note that $\prod_j I(0 < y_j < \theta) = I(m < \theta)$. Viewed as a function of θ this is maximised at $\widehat{\theta} = m$, which is therefore the MLE.
- Here the maximum is NOT found by differentiation of the likelihood, which is not differentiable at $\widehat{\theta}$.
- We need m and n to compute the likelihood, and since n is supposed constant, the maximum m is minimal sufficient.

Formal results

- For a less informal treatment we note that
  - any statistic $T = t(Y)$ taking values $t \in \mathcal{T}$ partitions the sample space $\mathcal{Y}$ into equivalence classes $\mathcal{C}_s = \{y' \in \mathcal{Y} : t(y') = s\}$;
  - the partition $\mathcal{C}_s$ corresponding to T is sufficient if and only if the distribution of Y within each $\mathcal{C}_s$ does not depend on θ; and
  - a minimal sufficient statistic gives the coarsest possible sufficient partition.
- We use the following two results to identify the (minimal) sufficient statistic.

Theorem 12 (Factorisation) *A statistic $S = s(Y)$ is sufficient for θ in a model $f(y;\theta)$ if and only if there exist functions g and h such that*

$$f(y;\theta) = g\{s(y);\theta\} \times h(y).$$

Theorem 13 *Let Y have density $f(y;\theta)$ and let $S = s(Y)$ be such that the ratio*

$$\frac{f(y;\theta)}{f(z;\theta)}$$

is free of θ if and only if $s(y) = s(z)$. Then S is minimal sufficient for θ.

Note to Theorem 12

- ☐ The result is 'if and only if', so we need to argue in both directions.
- ☐ If S is sufficient, then the factorisation

$$f(y;\theta) = f\{s(y);\theta\} \times f(y \mid s) = g\{s(y);\theta\} \times h(y)$$

holds.

- ☐ To prove the converse, suppose for simplicity that Y is discrete and that there is a factorisation. Then S has density

$$f(s;\theta) = \sum_{y' \in \mathcal{Y}: s(y')=s} g\{s(y');\theta\}h(y') = g(s;\theta) \sum_{y' \in \mathcal{Y}: s(y')=s} h(y'),$$

where the sum is in fact over $y' \in \mathcal{C}_s$. Thus the conditional density of Y given $S = s = s(y)$ is

$$f(y \mid s;\theta) = \frac{g\{s(y);\theta\}h(y)}{g(s;\theta)\sum_{y' \in \mathcal{C}_s} h(y')} = \frac{h(y)}{\sum_{y' \in \mathcal{C}_s} h(y')},$$

which does not depend on θ. Hence S is sufficient.

- ☐ The continuous case is similar, but the presence of a Jacobian makes the argument a bit messier.

Note to Theorem 13

- ☐ We must show that that S is sufficient and that it is minimal.
- ☐ To show sufficiency, note that every $y \in \mathcal{Y}$ lies in an element of the partition $\mathcal{C}_s$ generated by the possible values of S, and choose a representative dataset $y'_s \in \mathcal{C}_s$ for each s. For any y, $y'_{s(y)}$ is in the same equivalence set as y, so the ratio $f(y;\theta)/f(y'_{s(y)};\theta)$ does not depend on θ, by the premise of the theorem. Hence

$$f(y;\theta) = f(y'_{s(y)};\theta) \times \frac{f(y;\theta)}{f(y'_{s(y)};\theta)} = g\{s(y);\theta\} \times h(y),$$

because $y'_{s(y)}$ is a function of $s(y)$. This factorisation shows that $S = s(Y)$ is sufficient.

- ☐ To show minimality, if $T = t(Y)$ is any other sufficient statistic the factorisation theorem gives

$$f(y;\theta) = g'\{t(y);\theta\}h'(y)$$

for some g' and h'. If two datasets y and z are such that $t(y) = t(z)$, then

$$\frac{f(y;\theta)}{f(z;\theta)} = \frac{g'\{t(y);\theta\}h'(y)}{g'\{t(z);\theta\}h'(z)} = \frac{h'(y)}{h'(z)}$$

does not depend on θ, and hence $s(y) = s(z)$. This implies that

$$\{z \in \mathcal{Y} : t(z) = t(y)\} \subset \{z \in \mathcal{Y} : s(z) = s(y)\},$$

i.e., the partition generated by the values of S is coarser than that generated by the values of T, and therefore it must be minimal.

Exponential family models

- An elegant general theory puts many well-known distributions (Poisson, binomial, normal, ...) under the same roof.
- If $\theta \in \Theta \subset \mathbb{R}^d$, where $\dim \Theta = d$, and there exists a d-dimensional statistic $s = s(y)$ of data y and a parametrisation $\varphi \equiv \varphi(\theta)$, i.e., a 1–1 function of θ, such that

$$f(y;\theta) = m(y)\exp\{s^{\mathrm{T}}\varphi - k(\varphi)\} = m(y)\exp\left[s^{\mathrm{T}}\varphi(\theta) - k\{\varphi(\theta)\}\right], \quad \theta \in \Theta, y \in \mathcal{Y},$$

  then this is a **regular (or full) (d, d) exponential family** of distributions, and
  - the **canonical statistic** $S = s(Y)$ is minimal sufficient for θ,
  - the **canonical parameter** is φ,
  - the **mean parameter** $\eta = \mathrm{E}(S;\varphi) = \mathrm{d}k(\varphi)/\mathrm{d}\varphi$ is obtained by differentiating the
  - the **cumulant-generating function** $k(\varphi + t) - k(\varphi)$, and
  - the **cumulant generator** $k(\cdot)$ is convex on the set $\mathcal{N} = \{\theta : \log \int e^{s^{\mathrm{T}}\varphi} m(y)\,\mathrm{d}y < \infty\}$.
- The cumulant-generating function is the log moment-generating function for S, i.e.,

$$K_S(t) = \log M_S(t) = k(\varphi + t) - k(\varphi), \quad t \in \mathcal{T} \subset \mathbb{R}^d,$$

  where the open set $\mathcal{T}$ contains $t = 0$. Check that $\mathrm{E}(S) = \nabla k(\varphi)$, $\mathrm{var}(S) = \nabla^2 k(\varphi)$.

Note on cumulant-generating functions

- The moment-generating function for the canonical statistic S of an exponential family is

$$M_S(t) = \mathrm{E}\left\{\exp(t^{\mathrm{T}}S)\right\} = \int m(y)\exp\left\{s^{\mathrm{T}}t + s^{\mathrm{T}}\varphi - k(\varphi)\right\}\,\mathrm{d}y,$$

  and since this must equal unity when $t = 0$ we see that

$$\int m(y)\exp\left\{s^{\mathrm{T}}\varphi\right\}\,\mathrm{d}y = \exp\{k(\varphi)\},$$

  and therefore that if it is defined,

$$M_S(t) = \int m(y)\exp\left\{s^{\mathrm{T}}(t+\varphi) - k(\varphi)\right\}\,\mathrm{d}y = \exp\{k(\varphi + t) - k(\varphi)\},$$

  as required.
- To find the mean and variance we note that $M_S(0) = 1$,

$$\nabla M_S(t)|_{t=0} = \mathrm{E}(S), \quad \nabla^2 M_S(t)\big|_{t=0} - \mathrm{E}(S)\mathrm{E}(S)^{\mathrm{T}} = \mathrm{var}(S),$$

  and as $K_S(t) = \log M_S(t)$,

$$\nabla K_S(t) = M_S(t)^{-1}\nabla M_S(t), \quad \nabla^2 K_S(t) = M_S(t)^{-1}\nabla^2 M_S(t) - M_S(t)^{-2}\nabla M_S(t)\nabla^{\mathrm{T}} M_S(t),$$

  which reduce to the mean and covariance matrix when $t = 0$.

Examples

Example 14 (Poisson sample) *Are the models of Example 9 full exponential families?*

Example 15 (Satellite conjunction) *Show that the given model is an exponential family.*

Note to Example 14

☐ In the first model

$$f_y(y;\theta) = \prod_{j=1}^{n} \frac{\theta^{y_j}}{y_j!} e^{-\theta} = m(y)\exp(s\log\theta - n\theta),$$

where $m(y) = (\prod y_j)^{-1}$, $s = s(y) = \sum y_j$, $\varphi = \log\theta$, $k(\varphi) = n\theta = ne^{\varphi}$. This is a $(1,1)$ exponential family.

☐ In the second model we saw that

$$f(y,n;\theta) = m(y)(1-\theta)^{n-1}\theta^{s+1}e^{-n\theta} = m(y)\exp\{-n\theta + s\log\theta + n\log(1-\theta) + \log\theta - \log(1-\theta)\},$$

so we can take

$$s(y) = (n,s), \quad \varphi = (\log(1-\theta) - \theta, \log\theta), \quad k\{\varphi(\theta)\} = \log(1-\theta) - \log\theta,$$

with θ scalar but $s(y)$ of dimension 2. This is a $(2,1)$ curved exponential family (next slide).

Note to Example 15

☐ The multivariate normal density is

$$\begin{aligned} f(y;\mu,\Omega) &= \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^{\mathrm{T}}\Omega^{-1}(y-\mu)\right\}, \quad y \in \mathbb{R}^n \\ &= (2\pi)^{-n/2}\exp\left\{-\tfrac{1}{2}(y-\mu)^{\mathrm{T}}\Omega^{-1}(y-\mu) - \tfrac{1}{2}\log|\Omega|\right\}, \end{aligned}$$

and the data y only appear in the exponent,

$$(y-\mu)^{\mathrm{T}}\Omega^{-1}(y-\mu) = y^{\mathrm{T}}\Omega^{-1}y - 2y^{\mathrm{T}}\Omega^{-1}\mu + \mu^{\mathrm{T}}\Omega^{-1}\mu.$$

☐ When Ω is known (as in the satellite case), the only unknown parameter is μ and we can set

$$\varphi = \Omega^{-1}\mu, \quad s(y) = y, \quad m(y) = (2\pi)^{-n/2}|\Omega|^{-1/2}e^{-y^{\mathrm{T}}\Omega^{-1}y/2}, \quad k(\varphi) = \mu^{\mathrm{T}}\Omega^{-1}\mu/2 = \varphi^{\mathrm{T}}\Omega\varphi/2.$$

Note that $\nabla k(\varphi) = \Omega\varphi = \Omega\Omega^{-1}\mu = \mu = \mathrm{E}(y)$ and $\nabla^2 k(\varphi) = \Omega = \mathrm{var}(y)$, as expected.

☐ We could also have set $\varphi = \mu$ and $s(y) = \Omega^{-1}y$. Find $k(\varphi)$ in this case and find the corresponding mean and variance of the canonical statistic.

☐ If Ω is known, then y is a 1–1 transform of $s(y) = \Omega^{-1}y$, so y is sufficient also.

☐ In the satellite example $n = 2$ and $\Omega = D^{-1}$ is diagonal and known, so the expressions above simplify slightly.

Exponential family models II

- When $\dim s = k > \dim \theta = d$ the model is called a (k, d) **curved exponential family**, and the $k \times 1$ vector $\varphi(\theta)$ gives a d-dimensional manifold in $\mathbb{R}^k$.
- The joint density of a random sample $Y_1, \ldots, Y_n$ from an exponential family is

$$\prod_{j=1}^{n} f(y_j; \theta) = \prod_{j=1}^{n} m(y_j) \exp\left\{s_j^{\mathrm{T}} \varphi - k(\varphi)\right\} = m^*(y) \exp\left\{n \overline{s}^{\mathrm{T}} \varphi - n k(\varphi)\right\},$$

  say, where $n\overline{s} = \sum_j s(y_j)$, and this is also an exponential family, with canonical statistic $n\overline{s}$ and cumulant generator $nk(\varphi)$. Thus exponential families are **closed under sampling**.
- It is easy to check that the canonical statistic $S = \sum_{j=1}^{n} s(Y_j)$ of a random sample from an exponential family is minimal sufficient.
- The log likelihood $\ell(\theta) \equiv s^{\mathrm{T}} \varphi - k(\varphi)$ in a full exponential family is concave as a function of φ, so the MLEs of φ and μ are given by

$$s = \nabla k(\varphi)|_{\varphi = \widehat{\varphi}} = \widehat{\mu},$$

  and the observed and expected information quantities both equal $\nabla^2 k(\varphi)$.

Eliminating nuisance parameters

Sometimes the removal of nuisance parameters can be based on the following results.

Lemma 16 *In a statistical model $f(y; \psi, \lambda)$ let W_ψ be (minimal) sufficient for λ when ψ is regarded as fixed. Then the conditional density $f(y \mid w_\psi; \psi)$ depends only on ψ. This holds in particular if W_ψ does not depend on ψ.*

Lemma 17 *In a (d, d) exponential family in which $\varphi(\theta) = (\psi, \lambda)$ and $s = (t, w)$ is partitioned conformally with φ, the conditional density of T given $W = w^{\mathrm{o}}$ is an exponential family that depends only on ψ.*

Example 18 (2×2 table) *Apply Lemma 17 to the 2×2 table.*

Note to Lemma 16

If ψ is regarded as fixed, then

$$f(y; \psi, \lambda) = f(w_\psi; \psi, \lambda) \times f(y \mid w_\psi; \psi),$$

where the rightmost term is free of λ, with logarithm

$$\log f(y; \psi, \lambda) - \log f(w_\psi; \psi, \lambda).$$

Note to Lemma 17

In the discrete case, let $\sum_{\mathrm{o}}$ denote the sum over the set $\{y : w = w^{\mathrm{o}}\}$ and note that

$$\begin{aligned} f(w^{\mathrm{o}}; \psi, \lambda) &= \sum_{\mathrm{o}} m^*(y) \exp\{t^{\mathrm{T}}\psi + w^{\mathrm{o}\mathrm{T}}\lambda - k(\varphi)\} \\ &= \exp\{w^{\mathrm{o}\mathrm{T}}\lambda - k(\varphi)\} \sum_{\mathrm{o}} m^*(y) \exp\{t^{\mathrm{T}}\psi\} \end{aligned}$$

so

$$\begin{aligned} f(t \mid w^{\mathrm{o}}; \psi) &= \frac{m^*(y) \exp\{t^{\mathrm{T}}\psi + w^{\mathrm{o}\mathrm{T}}\lambda - k(\varphi)\}}{\exp\{w^{\mathrm{o}\mathrm{T}}\lambda - k(\varphi)\} \sum_{\mathrm{o}} m^*(y) \exp(t^{\mathrm{T}}\psi)} \\ &= m^*(y) \exp\left\{t^{\mathrm{T}}\psi - \log \sum_{\mathrm{o}} m^*(y) \exp(t^{\mathrm{T}}\psi)\right\} \\ &= m^*(y) \exp\{t^{\mathrm{T}}\psi - k(\psi; w^{\mathrm{o}})\}, \end{aligned}$$

say, where the cumulant generator for the conditional density depends on w^{o}.

Note to Example 18

☐ A 2×2 table arises when m_1 individuals are allocated to a treatment and m_0 are allocated to a control. Responses from all individuals are independent and are binary with values $0/1$, so the total number of successes for the control group $R_0 \sim B(m_0, \pi_0)$ is independent of those for the treatment group, $R_1 \sim B(m_1, \pi_1)$. If the parameter of interest is the difference in log odds of success. Here m_0 and m_1 are considered to be fixed, and R_0 and R_1 as random. If we write

$$\psi = \log\{\pi_1/(1-\pi_1)\} - \log\{\pi_0/(1-\pi_0)\} = \log\left\{\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}\right\}, \quad \lambda = \log\{\pi_0/(1-\pi_0)\},$$

then we have

$$\pi_0 = \frac{e^{\lambda}}{1+e^{\lambda}}, \quad \pi_0 = \frac{e^{\lambda+\psi}}{1+e^{\lambda+\psi}}, \quad \psi, \lambda \in \mathbb{R}$$

and the joint density of the data reduces to

$$\binom{m_0}{r_0}\pi_0^{r_0}(1-\pi_0)^{m_0-r_0} \times \binom{m_1}{r_1}\pi_1^{r_1}(1-\pi_1)^{m_1-r_1} = \binom{m_0}{r_0}\binom{m_1}{r_1}\frac{e^{r_1\psi+(r_0+r_1)\lambda}}{(1+e^{\lambda})^{m_0}(1+e^{\lambda+\psi})^{m_1}},$$

which is a $(2,2)$ exponential family with $\varphi = (\psi, \lambda)$, $s = (r_1, r_0+r_1)$, and

$$m^*(y) = \binom{m_0}{r_0}\binom{m_1}{r_1}, \quad k(\varphi) = -m_0 \log\left(1+e^{\lambda}\right) - m_1 \log\left(1+e^{\lambda+\psi}\right).$$

☐ The result above implies that conditioning on $W = R_0 + R_1$ will eliminate λ, and

$$\mathrm{P}(W=w) = \sum_{r=r_-}^{r_+} \binom{m_0}{w-r}\binom{m_1}{r}\frac{e^{r\psi+w\lambda}}{(1+e^{\lambda})^{m_0}(1+e^{\lambda+\psi})^{m_1}},$$

where $r_- = \max(0, w-m_0)$, $r_+ = \min(w, m_1)$, and hence the conditional density of $T = R_1$ given $W = R_1 + R_0 = w$ is the **non-central hypergeometric density**

$$\mathrm{P}(T=t \mid W=w; \psi) = \frac{\binom{m_0}{w-t}\binom{m_1}{t}e^{t\psi}}{\sum_{r=r_-}^{r_+}\binom{m_0}{w-r}\binom{m_1}{r}e^{r\psi}}, \quad t \in \{r_-, \ldots, r_+\}.$$

Simple frequentist inference

☐ A basic frequentist recipe for inference on a parameter of interest:

- find the likelihood function for the data Y;
- find a sufficient statistic $S = s(Y)$ of the same dimension as θ;
- find a function T of S whose distribution depends only on ψ;
- invert the distribution of T to find confidence limits for ψ for arbitrary α;
- (use the conditional distribution of Y given S to assess model adequacy).

Example 19 (Exponential sample) *Apply the recipe above to inference for the mean of an exponential random sample.*

Example 20 (2×2 table) *Apply the recipe above to the 2×2 table.*

Note to Example 19

- ☐ Here $\theta = \psi$, so we can replace ψ by θ.
- ☐ The likelihood equals the joint density of $y_1, \ldots, y_n$,

$$L(\theta) = \prod_{j=1}^{n} \theta^{-1} \exp(-y_j/\theta) = \theta^{-n} \exp(-s/\theta), \quad y_1, \ldots, y_n > 0, \theta > 0,$$

so $\ell(\theta) = -n \log \theta - s/\theta$ for $\theta > 0$.

- ☐ The scalar statistic $s = \sum_{j=1}^{m} y_j$ is clearly minimal sufficient for θ, and its distribution is gamma with parameters θ and n, i.e.,

$$f_S(s; \theta) = \frac{s^{n-1}}{\theta^n \Gamma(n)} \exp(-s/\theta), \quad s > 0, \theta > 0.$$

Note that $Q = S/\theta$ has density

$$f_Q(q; \theta) = f_S(s; \theta)|_{s=\theta q} \left| \frac{\mathrm{d}s}{\mathrm{d}w} \right| = \frac{q^{n-1}}{\Gamma(n)} \exp(-q), \quad q > 0,$$

so Q is a pivot, with quantiles q_p that satisfy $p = \int_0^{q_p} f_Q(u)\,\mathrm{d}u$, for $p \in (0, 1)$; these are just the quantiles of the gamma distribution with parameters 1 and n.

- ☐ For $\alpha \in (0, 1)$, the equation

$$1 - \alpha = \mathrm{P}(q_{\alpha/2} < S/\theta \leq q_{1-\alpha/2}) = \mathrm{P}(S/q_{1-\alpha/2} \leq \theta \leq S/q_{\alpha/2}),$$

yields exact $(1 - \alpha)$ confidence interval $(L, U) = (S/q_{1-\alpha/2}, S/q_{\alpha/2})$.

- ☐ The conditional distribution of $Y_1, \ldots, Y_n$ given $S = s$ is

$$\frac{f_Y(y; \theta)}{f_S(s; \theta)} = \frac{\theta^{-n} \exp(-s/\theta)}{\frac{s^{n-1}}{\theta^n \Gamma(n)} \exp(-s/\theta)} = \frac{\Gamma(n)}{s^{n-1}}, \quad (y_1, \ldots, y_n)/s \in \mathcal{S}_n,$$

where $\mathcal{S}_n$ denotes the simplex $\{(y_1, \ldots, y_n) : y_1, \ldots, y_n \geq 0, \sum_j y_j = 1\}$, and it follows that the joint distribution of $(Y_1/S, \ldots, Y_n/S)$ is uniform on $\mathcal{S}_n$ and does not depend on θ. Model assessment could be based on this.

Note to Example 20

- ☐ In this case

$$\mathrm{P}(T \leq t \mid W = w; \psi) = \sum_{r=r_-}^{t} \frac{\binom{m_0}{w-r}\binom{m_1}{r} e^{r\psi}}{\sum_{r=r_-}^{r_+} \binom{m_0}{w-r}\binom{m_1}{r} e^{r\psi}}, \quad t \in \{r_-, \ldots, r_+\},$$

and we can vary ψ to solve the equations

$$\mathrm{P}(T \leq t \mid W = w; \psi_-) = \alpha/2, \quad \mathrm{P}(T \leq t \mid W = w; \psi_+) = 1 - \alpha/2,$$

thus giving a $(1 - \alpha)$ confidence interval.

Ancillary statistics

Sometimes the dimension of the problem can be reduced by writing $S = (T, A)$ where $A = a(Y)$ is an **ancillary statistic**: a function of the minimal sufficient statistic whose distribution does not depend on the parameter. Then

$$f_Y(y; \theta) = f_{Y|S}(y \mid s) f_S(s; \theta) = f_{Y|S}(y \mid s) \times f_{T|A}(t \mid a; \theta) \times f_A(a),$$

and inference on θ is based on the second term only, with the other terms used for model-checking.

Example 21 (Location model) *Show that writing*

$$T = Y_{(1)}, \quad A = (0, Y_{(2)} - Y_{(1)}, \ldots, Y_{(n)} - Y_{(1)}),$$

leads to inference based on the conditional density

$$f(t \mid a; \theta) = \frac{\prod_{j=1}^n g(t - \theta + a_j)}{\int \prod_{j=1}^n g(u + a_j) \, \mathrm{d}u}.$$

Note to Example 21

- Write $y'_j = y_{(j)}$ for simplicity of notation, and note that

$$y'_1 = t, \quad y'_j = y'_1 + (y'_j - y'_1) = t + a_j, \quad j = 2, \ldots, n,$$

so the Jacobian for the transformation is

$$\frac{\partial(y'_1, \ldots, y'_n)}{\partial(t, a_2, \ldots, a_n)} = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix} = 1,$$

and thus (setting $a_1 = 0$ for simplicity) the density of the **configuration** A is

$$f_A(a) = \int \prod_{j=1}^n g(t + a_j - \theta) \, \mathrm{d}t = \int \prod_{j=1}^n g(u + a_j) \, \mathrm{d}u,$$

where we put $u = t - \theta$ in the second integral. We see that $Q = T - \theta$ is a pivot, because

$$\mathrm{P}(Q \leq q \mid A = a) = \mathrm{P}(T - \theta \leq q \mid A = a) = \frac{\int^q \prod_{j=1}^n g(u + a_j) \, \mathrm{d}u}{\int \prod_{j=1}^n g(u + a_j) \, \mathrm{d}u},$$

and using the quantiles $q_{\alpha/2}(a)$ and $q_{1-\alpha/2}(a)$ will give conditional confidence limits.

- Assessment of model fit (i.e., of g) can be based on QQ plots of the values of a. We are familiar with this in regression problems.

Comments

- The essence of the recipe on slide 51 is to base an exact pivot $Q = q(Y; \psi)$ on a minimal sufficient statistic and use the **significance (or p-value) function**

$$\mathrm{P}\{q(Y;\psi) \leq q_p\}, \quad p \in (0,1)$$

  to invert Q and thus make inference on ψ using the quantiles of Q.
- The difficulties are that:
  - finding the sufficient statistic and a function of it that depend exactly only on ψ are typically possible only in simple models;
  - finding the exact distribution of the pivot may be difficult; and
  - assessment of model fit using the conditional distribution is difficult in general.
- Nevertheless the recipe suggests how to proceed in more general settings, by basing **approximate pivots** on likelihood-based statistics, which will automatically depend on the minimal sufficient statistic.
- To finish this section we outline the two main likelihood-based approaches (to be justified later).

Maximum likelihood estimator

- In large samples from a **regular model** in which the true parameter is $\theta^0_{d\times 1}$, the maximum likelihood estimator $\widehat{\theta}$ has an approximate normal distribution,

$$\widehat{\theta} \dot{\sim} \mathcal{N}_d\left\{\theta^0, J(\widehat{\theta})^{-1}\right\}.$$

- If we write v_{rr} for the rth diagonal element of the $d \times d$ matrix $J(\widehat{\theta})^{-1}$, then the **Wald statistic (better, Wald pivot)**

$$\frac{\widehat{\theta}_r - \theta^0_r}{v_{rr}^{1/2}} \dot{\sim} \mathcal{N}(0,1), \quad r = 1, \ldots, d,$$

  is an approximate pivot involving the rth component θ^0_r of θ^0, for which an approximate $(1-2\alpha)$ confidence interval is

$$\widehat{\theta}_r \pm z_\alpha v_{rr}^{1/2}.$$

Profile log likelihood

- The Wald pivot is not invariant to interest-respecting transformations, but the **profile log likelihood** for ψ,

$$\ell_{\mathrm{p}}(\psi) = \max_{\lambda} \ell(\psi, \lambda) = \ell(\psi, \widehat{\lambda}_{\psi}),$$

  is invariant and provides a better (but computationally more demanding) approach to removing the nuisance parameter.
- If the model is regular and ψ is scalar then the **likelihood root**

$$r(\psi^0) = \mathrm{sign}(\widehat{\psi} - \psi^0) \left[2\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi^0)\}\right]^{1/2} \quad \dot{\sim} \quad \mathcal{N}(0,1),$$

  and therefore is an approximate pivot.
- A $(1-\alpha)$ confidence set for the value ψ^0 generating the data is

$$\{\psi : z_{\alpha/2} \leq r(\psi) \leq z_{1-\alpha/2}\}.$$

- We will say what 'regular' means and justify these results later.

Discovery of the top quark (Abe et al., 1995, PRL)

Here are two extracts from the article announcing the discovery:

TABLE I. Number of lepton + jet events in the 67 pb^{-1} data sample along with the numbers of SVX tags observed and the estimated background. Based on the excess number of tags in events with ≥3 jets, we expect an additional 0.5 and 5 tags from $t\bar{t}$ decay in the 1- and 2-jet bins, respectively.

N_{jet}	Observed events	Observed SVX tags	Background tags expected
1	6578	40	50 ± 12
2	1026	34	21.2 ± 6.5
3	164	17	5.2 ± 1.7
≥4	39	10	1.5 ± 0.4

The numbers of SVX tags in the 1-jet and 2-jet samples are consistent with the expected background plus a small $t\bar{t}$ contribution (Table I and Fig. 1). However, for the $W+ \geq$3-jet signal region, 27 tags are observed compared to a predicted background of 6.7 ± 2.1 tags [8]. The probability of the background fluctuating to ≥27 is calculated to be 2×10^{-5} (see Table II) using the procedure outlined in Ref. [1] (see [9]). The 27 tagged jets are in 21 events; the six events with two tagged jets can be compared with four expected for the top + background hypothesis and ≤1 for background alone. Figure 1 also shows the decay lifetime distribution

Performing a test

- ☐ There's a **null hypothesis** to be tested:

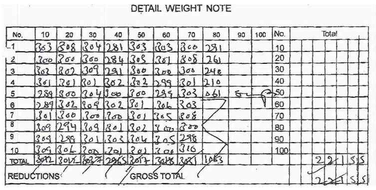
  H_0: the top quark does not exist.

  This seems counter-intuitive, but as one cannot prove a hypothesis, we attempt to refute its opposite — '**proof by (stochastic) contradiction**'.
- ☐ We obtain data, $y_{\text{obs}} = 27$ events on the 3-jet, 4-jet, ... channels.
- ☐ We compare y_{obs} with its distribution P_0 supposing that H_0 is true.
- ☐ Here P_0 is $\mathrm{Poiss}(\lambda_0 = 6.7)$ and represents the baseline noise under H_0.
- ☐ We compute the **P-value**

$$p_{\text{obs}} = \mathrm{P}_0(Y \geq y_{\text{obs}}) = \sum_{y=y_{\text{obs}}}^{\infty} \frac{\lambda_0^y}{y!} e^{-\lambda_0} = 3 \times 10^{-9},$$

  so
  - either H_0 is true but a (very) rare event has occurred,
  - or H_0 is false and the top quark exists.
- ☐ Abe et al. announced a discovery, but if they had found $p_{\text{obs}} \approx 0.001$, maybe they would have decided that H_0 could not (yet) be rejected, and not published their work.

Industrial fraud?



- ☐ $n = 92$ weighings of sacks upon the delivery of a commodity C:

```
261 289 291 265 281 291 285 283 280 261 263 281 291 289 280
292 291 282 280 281 291 282 280 286 291 283 282 291 293 291
300 302 285 281 289 281 282 261 282 291 291 282 280 261 283
291 281 246 249 252 253 241 281 282 280 261 265 281 283 280
242 260 281 261 281 282 280 241 249 251 281 273 281 261 281
282 260 281 282 241 245 253 260 261 281 280 261 265 281 241
260 241
```

- ☐ Their last digits are

0	1	2	3	4	5	6	7	8	9
14	42	14	9	0	6	2	0	0	5

- ☐ How can we tell if fraud has taken place?

Benford's law

Definition 22 *For $x \in \mathbb{R}$, let $d(x, j)$ denote the 'jth significant digit function (base 10)', so $d(31.4, 1) = 3$, $d(0.314, 2) = 1$ and $d(314, 3) = 4$.*

Definition 23 *If $x \in \mathbb{R}$ and $D_j = d(x, j)$, for $j = 1, 2, \ldots$, then (discarding any leading zeros) the D_r follow Benford's law if*

$$\mathrm{P}(D_1 = d_1, D_2 = d_2, \ldots, D_k = d_k) = \log_{10}\left\{1 + \left(\sum_{j=1}^{k} d_j \times 10^{k-j}\right)^{-1}\right\}, \quad d_j \in \{0, \ldots, 9\}.$$

- For example, $\mathrm{P}(D_1 = 3, D_2 = 1, D_3 = 4) = \log_{10}\{1 + (314)^{-1}\} \approx 0.0014$.
- This is an excellent model for the distribution of all sorts of digits.
- Frequencies (%) of the last digits D_3 for three-digit integers.

Digit	0	1	2	3	4	5	6	7	8	9
Uniform	10	10	10	10	10	10	10	10	10	10
Benford	10.178	10.137	10.097	10.057	10.017	9.978	9.940	9.901	9.864	9.826

- Apparently the last digits of the weighings should be approximately uniform.
- To detect forged deliveries we test the null hypothesis

$$H_0: \text{ the last digits of the weighings are uniformly distributed on 0,...,9.}$$

Pearson's statistic

Definition 24 *If $O_1, \ldots, O_k$ are the numbers of observations from a random sample of size n falling in categories $1, \ldots, k$, where*

$$\mathrm{E}(O_i) = E_i > 0, \quad i = 1, \ldots, k, \sum_{I=1}^{k} E_i = n,$$

then Pearson's statistic (aka the 'χ^2 statistic') is

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

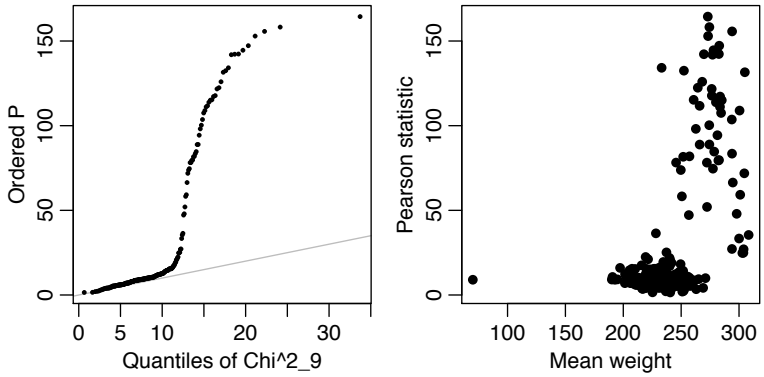
- If $O_1, \ldots, O_k$ are multinomially distributed with total n and probabilities $p_1 = E_1/n, \ldots, p_k = E_k/n$, then $T \dot{\sim} \chi^2_{k-1}$ (approximation OK if average $E_i \geq 5$).
- We use T to check whether data $O_1, \ldots, O_k$ agree with specified probabilities $p_1, \ldots, p_k$.
- For the original dataset we found $t_{\text{obs}} = 158.2$ and hence

$$p_{\text{obs}} = \mathrm{P}_0(T > t_{\text{obs}}) \doteq \mathrm{P}(\chi^2_9 \geq 158.2) \doteq 0,$$

which is essentially impossible for uniformly distributed digits.

For 250 'deliveries' ...

- ☐ Left: Q-Q plot of t_{obs} for 250 different deliveries.
- ☐ Right: mean weights for deliveries and values of t_{obs}. Deliveries with the largest values of t_{obs} tend also to be heavier, also suggestive of fraud.



Elements of a test

- ☐ A **null hypothesis** H_0 to be tested.
- ☐ A **test statistic** T, large values of which will suggest that H_0 is false, and with observed value t_{obs}.
- ☐ A **P-value**

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}),$$

where the **null distribution** $\mathrm{P}_0(\cdot)$ denotes a probability computed under H_0.

- ☐ The smaller p_{obs} is, the more we doubt that H_0 is true.
- ☐ If H_0 is true, then we can consider that p_{obs} is a realisation of a uniform random variable $P \sim U(0,1)$, and then

$$\mathrm{P}_0(P \leq p_{\text{obs}}) = p_{\text{obs}}.$$

- ☐ If I decide that H_0 is false, when in fact it is true, then I make an error whose probability under H_0 is exactly p_{obs} — so my uncertainty is quantified, because I know the probability of declaring a "**false positive**".

Note: Why is a P-value uniform?

- Let T be a test statistic whose distribution is $F_0(t)$ when the null hypothesis is true. Then the corresponding P-value is
$$\mathrm{P}_0(T \geq t_{\text{obs}}) = 1 - F_0(t_{\text{obs}}),$$
and if the value of t_{obs} is a realisation of T_{obs} (because the null hypothesis is true), then we can write the random value of p_{obs} seen in repetitions of the experiment as
$$P_{\text{obs}} = 1 - F_0(T_{\text{obs}}),$$
or equivalently $T_{\text{obs}} = F_0^{-1}(1 - P_{\text{obs}})$. Hence for $x \in [0, 1]$,
$$\begin{aligned} \mathrm{P}_0(P_{\text{obs}} \leq x) &= \mathrm{P}_0\left\{1 - F_0(T_{\text{obs}}) \leq x\right\} \\ &= \mathrm{P}_0\left\{1 - x \leq F_0(T_{\text{obs}})\right\} \\ &= \mathrm{P}_0\left\{T_{\text{obs}} \geq F_0^{-1}(1-x)\right\} \\ &= 1 - F_0\left\{F_0^{-1}(1-x)\right\} \\ &= x, \end{aligned}$$
which shows that $P_{\text{obs}} \sim U(0, 1)$.
- The above proof works for any continuous T_{obs}, but is only approximate if T_{obs} is discrete (e.g., has a Poisson distribution). In such cases P_{obs} can only take a finite or countable number of values known as the **achievable significance levels**.

Some asides

- If we say that a hypothesis is **true**, we mean 'it is reasonable to proceed as if the hypothesis was true' — any model is an idealisation, so it cannot be exactly 'true'.
- If we have a **discrete test statistic**, p_{obs} has at most a countable number of 'achievable significance levels'. This is only problematic when comparing tests, though randomisation has (unfortunately) sometimes been proposed to overcome it.
- We may consider a **two-sided test**, with both unusually large and unusually small values of T of interest. We can then define
$$p_+ = \mathrm{P}_0(T \geq t_{\text{obs}}), \quad p_- = \mathrm{P}_0(T \leq t_{\text{obs}}), \quad p_{\text{obs}} = 2\min(p_-, p_+),$$
so $p_- + p_+ = 1 + \mathrm{P}_0(T = t_{\text{obs}})$, which equals 1 unless T is discrete;
- We sometimes avoid minor problems due to discreteness by computing '**continuity-corrected**' P-values
$$p_+ = \sum_{t > t_{\text{obs}}} \mathrm{P}_0(T = t) + \tfrac{1}{2}\mathrm{P}_0(T = t_{\text{obs}}), \quad p_- = \sum_{t < t_{\text{obs}}} \mathrm{P}_0(T = t) + \tfrac{1}{2}\mathrm{P}_0(T = t_{\text{obs}}).$$
- The top quark and fraud examples illustrate **pure significance tests**, where the situation if H_0 is false is not explicitly considered. We look at the effect of alternatives now.

Testing as decision-making

Formulate testing as deciding between two hypotheses (Neyman–Pearson approach):

- the **null hypothesis** H_0, which represents a baseline situation;
- the **alternative hypothesis** H_1, which represents what happens if H_0 is false.
- We choose H_1 and 'reject' H_0 if p_{obs} is lower than some $\alpha \in (0, 1)$.
- For given α we partition the sample space $\mathcal{Y}$ as

$$\mathcal{Y}_0 = \{y \in \mathcal{Y} : p_{\text{obs}}(y) > \alpha\}, \quad \mathcal{Y}_1 = \{y \in \mathcal{Y} : p_{\text{obs}}(y) \leq \alpha\},$$

  where the notation $p_{\text{obs}}(y)$ indicates that the P-value depends on the data, or equivalently

$$\mathcal{Y}_0 = \{y \in \mathcal{Y} : t(y) < t_{1-\alpha}\}, \quad \mathcal{Y}_1 = \{y \in \mathcal{Y} : t(y) \geq t_{1-\alpha}\},$$

  where t_p denotes the p quantile of the test statistic $T = t(Y)$ under H_0
- We call $\mathcal{Y}_1$ the **size α critical region** of the test, and we reject H_0 in favour of H_1 if $Y \in \mathcal{Y}_1$, or equivalently if the test statistic exceeds the **size α critical point** $t_{1-\alpha}$.
- Critical regions of different sizes for the same test should be nested, i.e., (in an obvious notation) if $\alpha' > \alpha$, then

$$\mathcal{Y}_1^{\alpha} \subset \mathcal{Y}_1^{\alpha'} \quad \text{and} \quad t_{1-\alpha} > t_{1-\alpha'}.$$

False positives and negatives

		Decision	
		Accept H_0	Reject H_0
State of Nature	H_0 true	Correct choice (True negative)	Type I Error (False positive)
	H_1 true	Type II Error (False negative)	Correct choice (True positive)

- We can make two sorts of wrong decisions:

  Type I error (false positive): H_0 is true, but we wrongly reject it (and choose H_1);

  Type II error (false negative): H_1 is true, but we wrongly choose H_0.
- Notice that the consequences of bad decisions are not really taken into account in this framework.
- Statistics books and papers call
  - the **Type I error/false positive probability** the **size** $\alpha = \mathrm{P}_0(Y \in \mathcal{Y}_1)$, and
  - the **true positive probability** the **power** $\beta = \mathrm{P}_1(Y \in \mathcal{Y}_1)$.

Example 25 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with σ^2 known, and $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$, find the Type II error as a function of the Type I error.*

Note to Example 31

- The minimal sufficient statistic for the normal model with both parameters unknown is $(\overline{Y}, S^2)$, and it is easy to check that if σ^2 is known the minimal sufficient statistic reduces to $\overline{Y}$, which has a $\mathcal{N}(\mu_0, \sigma^2/n)$ distribution under H_0. Hence we take the test statistic T to be $\overline{Y}$, and $\mathcal{Y} = \mathbb{R}^n$.
- If $\mu_1 > \mu_0$, then clearly we will take

$$\mathcal{Y}_0 = \{y : \overline{y} < t_{1-\alpha}\}, \quad \mathcal{Y}_1 = \{y : \overline{y} \geq t_{1-\alpha}\};$$

this can be justified using the Neyman–Pearson lemma (below). Now

$$\mathrm{P}_0(Y \in \mathcal{Y}_0) = \mathrm{P}_0(\overline{Y} < t_{1-\alpha}) = \mathrm{P}_0\{\sqrt{n}(\overline{Y}-\mu_0)/\sigma < \sqrt{n}(t_{1-\alpha}-\mu_0)/\sigma\} = \Phi\left\{\sqrt{n}(t_{1-\alpha} - \mu_0)/\sigma\right\},$$

because $Z = \sqrt{n}(\overline{Y} - \mu_0)/\sigma \sim \mathcal{N}(0,1)$ under H_0, and for this probability to equal $1-\alpha$ we must take $t_{1-\alpha} = \mu_0 + \sigma n^{-1/2} z_{1-\alpha}$; this gives Type I error α.

- Note that although the form of $\mathcal{Y}_0$ above was determined by H_1, the value of $t_{1-\alpha}$ is given by calculations under H_0.
- $Z = \sqrt{n}(\overline{Y} - \mu_1)/\sigma \sim \mathcal{N}(0,1)$ under H_1, so the Type II error is

$$\begin{aligned} \mathrm{P}_1(Y \in \mathcal{Y}_0) &= \mathrm{P}_1(\overline{Y} < t_{1-\alpha}) \\ &= \mathrm{P}_1(\overline{Y} < \mu_0 + \sigma n^{-1/2} z_{1-\alpha}) \\ &= \mathrm{P}_1\{\sqrt{n}(\overline{Y} - \mu_1)/\sigma < \sqrt{n}(\mu_0 + \sigma n^{-1/2} z_{1-\alpha} - \mu_1)/\sigma\} \\ &= \Phi(z_{1-\alpha} - \delta), \end{aligned}$$

where $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$. Hence the Type II error equals $1-\alpha$ when $\mu_1 = \mu_0$ and decreases as a function of δ. We would expect this, because as μ_1 increases, the distribution of $\overline{Y}$ under H_1 shifts to the right and we are less likely to make a false negative error.

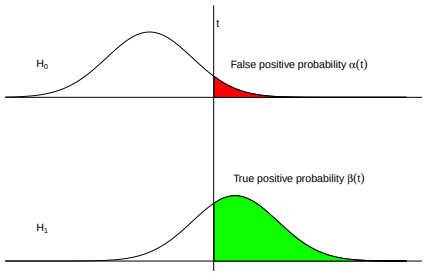
True and false positives: Example

- It is traditional to fix α and choose T (or equivalently $\mathcal{Y}_1$) to maximise β, but usually more informative to consider $\mathrm{P}_0(T \geq t)$ and $\mathrm{P}_1(T \geq t)$ as functions of t.
- In Example 31 we would
  - reject H_0 incorrectly (false positive) with probability

$$\alpha(t) = \mathrm{P}_0(T \geq t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma\},$$

  - reject H_0 correctly (true positive) with probability

$$\beta(t) = \mathrm{P}_1(T \geq t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma - \delta\}.$$



ROC curve

Definition 26 *The* **receiver operating characteristic (ROC) curve** *of a test plots* $\beta(t)$ *against* $\alpha(t)$ *as* t *varies, i.e., it shows* $(\mathrm{P}_0(T \geq t), \mathrm{P}_1(T > t))$*, when* $t \in \mathbb{R}$.

- As μ increases, it becomes easier to detect when H_0 is false, because the densities under H_0 and H_1 become more separated, and the ROC curve moves 'further north-west'.
- When H_0 and H_1 are the same, i.e., $\mu = 0$, then the curve lies on the diagonal. Then the hypotheses cannot be distinguished.
- A common summary measure of the overall quality of a test is the **area under the curve**,

$$\mathrm{AUC} = \int_0^1 \beta(\alpha)\,\mathrm{d}\alpha,$$

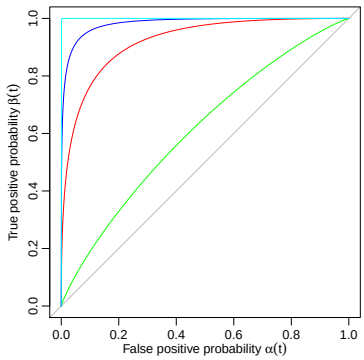
  which ranges between 0.5 for a useless test and 1.0 for a perfect test.

Example

☐ In Example 31 $\alpha(t) = 1 - \Phi\{n^{1/2}(t-\mu_0)/\sigma\}$ and $\beta(t) = 1 - \Phi\{n^{1/2}(t-\mu_0)/\sigma - \delta\}$, so equivalently we graph

$$\beta(t) = 1 - \Phi(-z_{1-\alpha} - \delta) = \Phi(\delta + z_\alpha) \equiv \beta(\alpha) \text{ against } \alpha \in (0,1).$$

☐ Here is the ROC curve with $\mu = 2$ (in red). Also shown are curves for $\mu = 0, 0.4, 3, 6$. Which is which?



Neyman–Pearson lemma

Definition 27 *A* **simple hypothesis** *entirely fixes the distribution of the data Y, whereas a* **composite hypothesis** *does not fix the distribution of Y.*

Definition 28 *The* **critical region** *of a hypothesis test is the subset $\mathcal{Y}_1$ of the sample space $\mathcal{Y}$ for which $Y \in \mathcal{Y}_1$ implies that the null hypothesis is rejected.*

We aim to choose $\mathcal{Y}_1$ to maximise the power of the test for a given size, i.e., such that $\mathrm{P}_1(Y \in \mathcal{Y}_1)$ is the largest possible such that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$.

Lemma 29 (Neyman–Pearson) *Let $f_0(y)$, $f_1(y)$ be the densities of Y under simple null and alternative hypotheses. Then if it exists, the set*

$$\mathcal{Y}_1 = \{y \in \mathcal{Y} : f_1(y)/f_0(y) > t\}$$

such that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$ maximises $\mathrm{P}_1(Y \in \mathcal{Y}_1)$ amongst all $\mathcal{Y}_1'$ for which $\mathrm{P}_0(Y \in \mathcal{Y}_1') \leq \alpha$. Thus the test of size α with maximal power rejects H_0 when $Y \in \mathcal{Y}_1$.

Example 30 *Construct an optimal test for testing $H_0 : \varphi = \varphi_0$ against $H_1 : \varphi = \varphi_1$ based on a random sample from a canonical exponential family.*

Note to Lemma 29

Suppose that a region $\mathcal{Y}_1$ such that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$ exists and let $\mathcal{Y}_1'$ be any other critical region of size α or less. Then for any density f,

$$\int_{\mathcal{Y}_1} f(y)\,\mathrm{d}y - \int_{\mathcal{Y}_1'} f(y)\,\mathrm{d}y, \tag{2}$$

equals

$$\int_{\mathcal{Y}_1\cap\mathcal{Y}_1'} f(y)\,\mathrm{d}y + \int_{\mathcal{Y}_1\cap\mathcal{Y}_0'} f(y)\,\mathrm{d}y - \int_{\mathcal{Y}_1'\cap\mathcal{Y}_1} f(y)\,\mathrm{d}y - \int_{\mathcal{Y}_1'\cap\mathcal{Y}_0} f(y)\,\mathrm{d}y,$$

where $\mathcal{Y}_0$ and $\mathcal{Y}_0'$ are the complements of $\mathcal{Y}_1$ and $\mathcal{Y}_1'$ in the sample space, and this is

$$\int_{\mathcal{Y}_1\cap\mathcal{Y}_0'} f(y)\,\mathrm{d}y - \int_{\mathcal{Y}_1'\cap\mathcal{Y}_0} f(y)\,\mathrm{d}y. \tag{3}$$

If $f = f_0$, then (2) is non-negative, because $\mathcal{Y}'$ has size at most that of $\mathcal{Y}_1$, so (3) is also non-negative, giving

$$t\int_{\mathcal{Y}_1\cap\mathcal{Y}_0'} f_0(y)\,\mathrm{d}y \geq t\int_{\mathcal{Y}_1'\cap\mathcal{Y}_0} f_0(y)\,\mathrm{d}y$$

for $t \geq 0$. But $f_1(y) > tf_0(y)$ for $y \in \mathcal{Y}_1$, and $tf_0(y) \geq f_1(y)$ for $y \in \mathcal{Y}_0$, so

$$\int_{\mathcal{Y}_1\cap\mathcal{Y}_0'} f_1(y)\,\mathrm{d}y \geq \int_{\mathcal{Y}_1'\cap\mathcal{Y}_0} f_1(y)\,\mathrm{d}y.$$

On adding $\int_{\mathcal{Y}_1\cap\mathcal{Y}_1'} f_1(y)\,\mathrm{d}y$ to both sides we see that the power of $\mathcal{Y}_1$ is at least that of $\mathcal{Y}_1'$, as required.

Note to Example 30

- The likelihood ratio is

$$\frac{f_1(y)}{f_0(y)} = \frac{m^*(y)\exp\{\varphi_1 s^* - nk(\varphi_1)\}}{m^*(y)\exp\{\varphi_0 s^* - nk(\varphi_0)\}} = \exp\{(\varphi_1-\varphi_0)s^* + nk(\varphi_0) - nk(\varphi_1)\},$$

  say, where $s^* = \sum_{j=1}^n s(y_j)$, so

$$\mathcal{Y}_1 = \{y : f_1(y)/f_0(y) > t\} = \{y : (\varphi_1-\varphi_0)s^* + nk(\varphi_0) - nk(\varphi_1) > \log t\},$$

  and if $\varphi_1 > \varphi_0$ then

$$\mathcal{Y}_1 = \{y : s^* > [\log t + nk(\varphi_1) - nk(\varphi_0)]/(\varphi_1-\varphi_0)\},$$

  This gives the form of $\mathcal{Y}_1$ and we should choose t so that $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$, or equivalently s_α so that (in the continuous case)

$$\mathrm{P}_0(S^* > s_\alpha) = \int_{s_\alpha}^{\infty} f(s;\varphi_0)\,\mathrm{d}s = \alpha.$$

  We saw such a calculation already in Example 31 for normal data with known σ^2 and $\varphi_1 = \mu_1/\sigma^2 > \varphi_0 = \mu_0/\sigma^2$.
- If $\varphi_1 < \varphi_0$, then division by $\varphi_1 - \varphi_0 < 0$ leads to

$$\mathcal{Y}_1^* = \{y : s^* < [\log t + nk(\varphi_1) - nk(\varphi_0)]/(\varphi_1-\varphi_0)\}.$$

- The Neyman–Pearson lemma tell us that $\mathcal{Y}_1$ gives a most powerful test, but as it does not depend on the value of φ, this test is **uniformly most powerful** for all $\varphi > \varphi_0$, and likewise $\mathcal{Y}_1^*$ is **uniformly most powerful** for $\varphi_1 < \varphi_0$.

Exact and inexact tests

- Above we saw that $P \sim U(0,1)$ under the null hypothesis, exactly in continuous cases and approximately in discrete cases.
- If the null distribution of the test statistic is estimated, we have $P \dot\sim U(0,1)$ only.
- For example, if the true parameter is $\theta = (\psi_0, \lambda_0)$ and $H_0 : \psi = \psi_0$, then the P-value is

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}) = \mathrm{P}(T \geq t_{\text{obs}}; \psi_0, \lambda_0),$$

  which we estimate by

$$\widehat{p}_{\text{obs}} = \mathrm{P}(T \geq t_{\text{obs}}; \psi_0, \widehat{\lambda}_0),$$

  where $\widehat{\lambda}_0$ is the estimate of λ under H_0.
- Exact tests, with $P \sim U(0,1)$, can sometimes be obtained by using a pivot whose distribution is invariant to λ, or by removing λ by conditioning.

Example 31 *If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, show that the distribution of $T = (\overline{Y} - \mu)/\sqrt{S^2/n}$ is invariant to σ^2.*

Example 32 *Find an exact test on a canonical parameter in a logistic regression model.*

Note to Example 31

In this case the variables $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ are independent and we can write $\overline{Y} \stackrel{\mathrm{D}}{=} \mu + \sigma n^{-1/2} Z$ and $S^2 = \stackrel{\mathrm{D}}{=} \sigma^2 V/(n-1)$, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ are independent. Hence

$$T = \frac{\overline{Y} - \mu}{\sqrt{S^2/n}} \stackrel{\mathrm{D}}{=} \frac{\mu + \sigma Z/n^{-1/2} - \mu}{[\sigma^2 V/\{n(n-1)\}]^{1/2}} \stackrel{\mathrm{D}}{=} \frac{Z}{\sqrt{V/(N-1)}} \sim t_{n-1},$$

is pivotal and thus allows tests on μ without reference to σ^2.

Note to Example 32

- ☐ In a logistic regression model we have independent binary variables $Y_1, \ldots, Y_n$ each with density

$$\mathrm{P}(Y_j = y_j; \beta) = \pi_j^{y_j}(1-\pi_j)^{1-y_j} = \left(\frac{e^{x_j^{\mathrm{T}}\beta}}{1+e^{x_j^{\mathrm{T}}\beta}}\right)^{y_j} \left(\frac{1}{1+e^{x_j^{\mathrm{T}}\beta}}\right)^{1-y_j} = \frac{e^{y_j x_j^{\mathrm{T}}\beta}}{1+e^{x_j^{\mathrm{T}}\beta}},$$

  for $y_j \in \{0,1\}$, known covariate vectors $X_j \in \mathbb{R}^d$ and parameter $\beta \in \mathbb{R}^d$.
- ☐ The corresponding log likelihood is

$$\ell(\beta) = \sum_{j=1}^{n} \left\{ y_j x_j^{\mathrm{T}}\beta - \log\left(1+e^{x_j^{\mathrm{T}}\beta}\right)\right\} = y^{\mathrm{T}} X\beta - \sum_{j=1}^{n} \log\left(1+e^{x_j^{\mathrm{T}}\beta}\right), \quad \beta \in \mathbb{R}^d.$$

  This is a (d,d) exponential family with canonical statistic $S = X^{\mathrm{T}}y$, canonical parameter $\varphi = \beta$, and cumulant generator $k(\varphi) = \sum_{j=1}^{n} \log\left(1+e^{x_j^{\mathrm{T}}\varphi}\right)$.
- ☐ Hence Lemma 17 implies that if $\varphi = (\psi, \lambda)$ and $S = (T, W) = (X_1^{\mathrm{T}}y, X_2^{\mathrm{T}}y)$, where X_1 is $n \times 1$ and X_2 is $n \times (d-1)$, an exact test on ψ is obtained from the conditional distribution

$$\mathrm{P}(T = t \mid W = w^{\mathrm{o}}; \psi) = \frac{e^{t\psi}}{\sum_{y' \in \mathcal{S}_{w^{\mathrm{o}}}} e^{X_1^{\mathrm{T}}y'\psi}},$$

  where $\mathcal{S}_w = \{(y'_1, \ldots, y'_n) : X_2^{\mathrm{T}}y' = w^{\mathrm{o}}\}$, with $w^{\mathrm{o}} = X_2^{\mathrm{T}}y^{\mathrm{o}}$ and y^{o} respectively the observed data and the observed value of W.
- ☐ Calculation of this conditional density in applications may be awkward, but excellent approximations are available.

Interpretation of P-values

- ☐ Be careful about interpretation:
  - p_{obs} is a one-number summary of whether data are consistent with H_0;
  - it is NOT the probability that H_0 is true (require prior probabilities on H_0 and H_1);
  - even a tiny p_{obs} can support H_0 better than H_1 (consider $t_{\text{obs}} = 3$ when $T \sim \mathcal{N}(\mu, 1)$ with $\mu_0 = 0$, $\mu_1 = 10$);
  - the power depends on analogues of $\delta = n^{1/2}(\mu_1 - \mu_0)/\sigma$, where n is the **sample size**, $\mu_1 - \mu_0$ is the **effect size**, and σ is the **precision**, so
    - ▷ even a tiny (practically irrelevant) effect size can be detected with very large n;
    - ▷ conversely a practically important effect might be undetectable if n is small;
    - ▷ i.e., 'statistical significance' $\neq$ 'subject-matter importance'!
- ☐ A confidence interval, or estimate and its standard error, is often more informative.
- ☐ Hypothesis testing is often applied by rote — in some medical journals no statement is complete without an accompanying '$(P < 0.05)$' — and is sometimes regarded as controversial, with certain journals now refusing to publish tests and P-values.
- ☐ The replication crisis is partly due to abuse of hypothesis testing, e.g., by not correcting for multiple tests, by formulating hypotheses in light of the data, ...

Contexts of testing

- ☐ It is unwise to be too categorical about testing, because of its different uses:
  - testing a clear hypothesis of scientific interest (e.g., top quark);
  - goodness of fit of a model (e.g., industrial fraud);
  - decision-making with a clearly-specific alternative (e.g., covid testing);
  - model simplification if null hypothesis true;
  - 'dividing hypothesis' used to split parameter space into different sets with sharply different interpretations;
  - as a technical device for generating confidence intervals;
  - to flag which of many null similar hypotheses might be false.

Example 33 *The generalized Pareto distribution, with survival function*

$$\mathrm{P}(X > x) = \begin{cases} (1 + \xi x/\sigma)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-x/\sigma), & \xi = 0, \end{cases}$$

simplifies if $\xi = 0$, and has finite upper support point $x_+ = -\sigma/\xi$ when $\xi < 0$ but $x_+ = \infty$ when $\xi \geq 0$. Here $H_0 : \xi = 0$ is both a simplifying and a dividing hypothesis, of interest when the distribution is fitted to data on supercentenarians.

Generating confidence intervals

- Using a statistic T with observed value t_{obs} to test the null hypothesis $H_0 : \psi = \psi_0$ for a scalar parameter ψ gives P-value

$$p_{\text{obs}} = p(\psi_0) = \mathrm{P}(T \geq t_{\text{obs}}; \psi_0),$$

  and we regard ψ_0 as incompatible with the data if p_{obs} is too small.

- Recall that the corresponding random variable $P_{\text{obs}} \sim U(0, 1)$ under H_0. Hence we can regard values of ψ for which the **P-value (or significance) function**

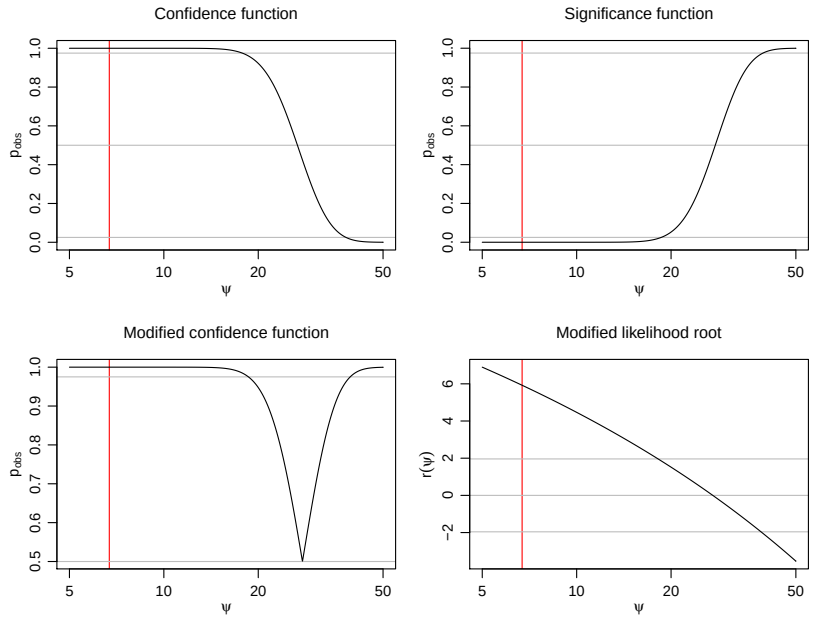
$$p(\psi) = \mathrm{P}(T \geq t_{\text{obs}}; \psi)$$

  is too extreme as incompatible with the data, leading to the (two-sided) $(1 - \alpha)$ confidence set

$$\{\psi : \alpha/2 \leq \mathrm{P}(T \geq t_{\text{obs}}; \psi) \leq 1 - \alpha/2\}.$$

- Related functions include
  - the **confidence function** $1 - p(\psi)$;
  - the **modified confidence function** $\max\{p(\psi), 1 - p(\psi)\}$; and
  - a **pivot function** showing a (usually standard normal) pivot as a function of ψ.

Significance and related functions



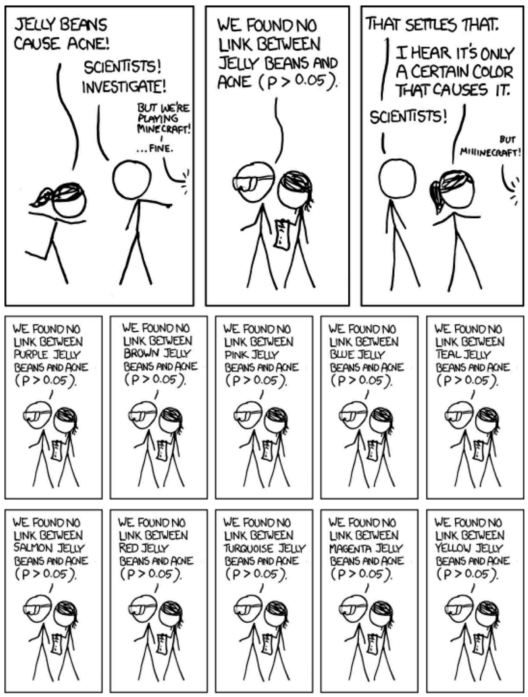
Multiple testing

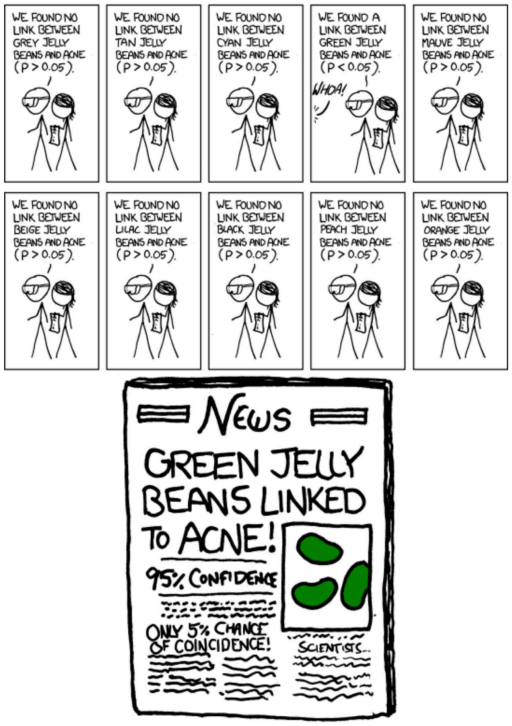
- Often require tests of several, even very many, hypotheses:
  - comparison of responses for several treatment groups with the same control group;
  - checking for a change in a series of observations;
  - screening genomic data for effects of many genes on a response.
- There are null hypotheses $H_1, \ldots, H_m$, of which
  - m_0 are true, indexed by an unknown set $\mathcal{I}$,
  - $m_1 = m - m_0$ are false, and
  - the **global null hypothesis** is $H_0 = H_1 \cap \cdots \cap H_m$.
- We apply some testing procedure and declare R hypotheses to be significant, of which FP are false positives and TP are true positives. Only R and m are known.

	Non-significant	Significant	
True nulls	TN	FP	m_0
False nulls	FN	TP	$m - m_0$
		R	m

- In the cartoon we have $m = 20$ hypotheses individually tested with $\alpha = 0.05$. We observe $R = 1$, but $\mathrm{E}(\mathrm{FP}) = m\alpha = 1$, so this is not a surprise.

The perils of multiple testing





Graphical approach

- Graphs can be helpful in suggesting which hypotheses are most suspect, and it is helpful to highlight the corresponding (i.e., smallest) P-values.
- $P \sim U(0,1)$ implies $Z = -\log_{10} P \sim \exp(\lambda)$ with $\lambda = \ln 10$.
- With this transformation small P_j become large Z_j; note that $Z_j > a$ iff $P_j < 10^{-a}$.
- If H_0 is true and the tests are independent, then $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} \exp(\lambda)$ and the **Rényi representation**

$$Z_{(r)} \overset{\text{D}}{=} \lambda^{-1} \sum_{j=1}^{r} \frac{E_j}{m+1-j}, \quad r = 1, \ldots, m, \quad E_1, \ldots, E_m \overset{\text{iid}}{\sim} \exp(1),$$

  applies to their order statistics. Then
  - a plot of the ordered empirical Z_j against their expectations should be straight;
  - outliers, very large Z_j (i.e., very small P_j), casting doubt on the corresponding H_j.
  - For very small P_j (i.e., large Z_j) the uniformity may fail even under H_0, because the null distributions give poor approximations in the extreme tail; then some form of model-fitting may be needed.
  - Similar ideas apply to z statistics (e.g., in regression): use a normal QQ-plot (excluding the intercept etc.) as a basis for discussion of significant effects.

GWAS, I

- A **genome-wide association study (GWAS)** tests the association between SNPs ('single nucleotide polymorphisms') and a phenotype such as the expression of a protein. The null hypotheses are

$$H_{0,j} : \text{no association between the expression of the protein and SNP}_j, \quad j = 1, \ldots, m.$$

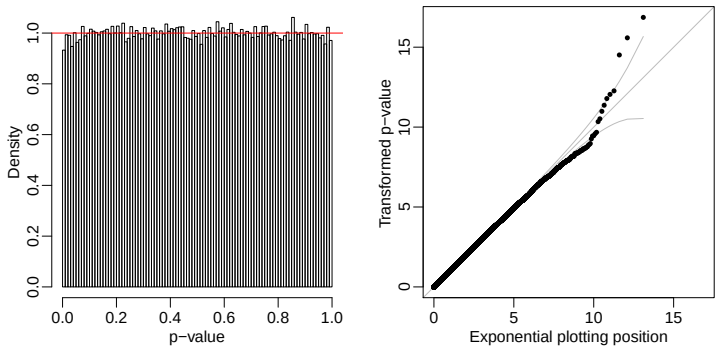
- In a simple model we construct statistics Y_j such that $Y_j \dot{\sim} \mathcal{N}(\theta_j, 1)$, where $\theta_j = 0$ under $H_{0,j}$, and we take $T_j = |Y_j|$, which is likely to be far from zero if $\theta_j \gg 0$ or$\theta_j \ll 0$.
- If $t_{\text{obs},j}$ denotes the observed value of T_j, then the P-value for association j is

$$p_{\text{obs},j} = \mathrm{P}_0(T_j > t_{\text{obs},j}) = 1 - \mathrm{P}_0(-t_{\text{obs},j} \leq Y_j \leq t_{\text{obs},j}) \doteq 2\Phi(-t_{\text{obs},j}),$$

  where the approximation comes from the fact that $Y_j \dot{\sim} \mathcal{N}(0,1)$ under $H_{0,j}$.
- Here it is reasonable to expect that the effects are **sparse**, i.e., most of the $\theta_j = 0$, and we seek a needle in a haystack.
- With many tests it is essential to ensure that the true positives are not drowned in the mass of false positives.

GWAS, II

- Left: a histogram of the P-values for tests of the association between $m = 275297$ SNPs and the expression of the protein CFAB.
- The P-values for SNPs not associated with CFAB are uniformly distributed. Is there an excess of small P-values?
- Right: exponential Q-Q plot of the $Z_j = -\log P_j$. What do you make of it?



Control

- With several tests Type I error generalises to the **familywise error rate (FWER)**, i.e., the probability of at least one false positive when the individual hypotheses are tested,

$$\mathrm{FWER} = \mathrm{P}(\mathrm{FP} \geq 1) = 1 - \mathrm{P}(\text{accept all } H_j, j \in \mathcal{I}),$$

  and we aim to control this by ensuring that $\mathrm{FWER} \leq \alpha$.
- There are different notions of control:
  - **weak control** guarantees $\mathrm{FWER} \leq \alpha$ only under H_0, i.e., $m_0 = m$;
  - **strong control** guarantees $\mathrm{FWER} \leq \alpha$ for any configuration of null and alternative hypotheses.
- If all the tests are independent and we use individual levels α, then

$$\mathrm{FWER} = 1 - \mathrm{P}(\mathrm{FP} = 0) = 1 - (1-\alpha)^{m_0} \to 1, \quad m_0 \to \infty.$$

- If conversely we fix FWER and the tests are independent we need

$$\alpha = 1 - (1 - \mathrm{FWER})^{1/m_0},$$

  so with $m_0 = 20$ and $\mathrm{FWER} = 0.05$ we need $\alpha \doteq 0.0026$, so the power for individual tests will be tiny (recall ROC curves).

Bonferroni methods

- If P_j is the P-value for the jth test and we reject H_j if $P_j < \alpha/m$, then **Boole's inequality** (the first of the **Bonferroni inequalities**) gives

$$\text{FWER} = \text{P}(\text{FP} \geq 1) = \text{P}\left(\bigcup_{j=1}^{m_0}\left\{P_j \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{j=1}^{m_0}\text{P}\left(P_j \leq \frac{\alpha}{m}\right) = m_0\frac{\alpha}{m} \leq \alpha,$$

  so we have strong control of FWER, even if the tests are dependent.
- Note that we could replace α/m for test j by α_j such that $\sum_{j=1}^{m}\alpha_j \leq \alpha$.
- The resulting **Bonferroni procedure** lacks power when m is large (because α/m is very small), but its assumptions are very weak.
- An improvement is the **Holm–Bonferroni procedure**: for given α,
  - order the P-values as $P_{(1)} \leq \cdots \leq P_{(m)}$ and the hypotheses as $H_{(1)}, \ldots, H_{(m)}$, then
  - reject $H_{(1)}, \ldots, H_{(S-1)}$, where

$$S = \min\left\{s : P_{(s)} > \frac{\alpha}{m+1-s}\right\}.$$

  This gives strong control and is more powerful than the basic Bonferroni procedure.

Note: Bonferroni–Holm procedure

If $\text{FP} \geq 1$, then we must have wrongly rejected some H_j for which $j \in \mathcal{I}$. If $H_{(s)}$ is the first such hypothesis rejected in the sequential procedure, then the $s-1$ hypotheses rejected before it must have been among the $m - m_0$ false null hypotheses, so $s - 1 \leq m - m_0$, i.e., $m_0 \leq m + 1 - s$. As $H_{(s)}$ was rejected, the corresponding P-value satisfies

$$P_{(s)} \leq \frac{\alpha}{m+1-s} \leq \frac{\alpha}{m_0}.$$

This implies that if $\text{FP} \geq 1$ then the P-value for at least one of the true null hypotheses satisfies $P_j \leq \alpha/m_0$, and so Boole's inequality gives

$$\text{FWER} = \text{P}(\text{FP} \geq 1) \leq \text{P}\left(\bigcup_{j\in\mathcal{I}}\{P_j \leq \alpha/m_0\}\right) \leq \alpha.$$

As the only assumption needed for the above argument was that the null P-values are $U(0,1)$, this procedure strongly controls the FWER at level α.

False discovery rate

- When m is large and the goal is exploratory, Bonferroni procedures are unreasonably stringent, and it seems preferable to try and control the **false discovery proportion**

$$I(R > 0)\mathrm{FP}/R,$$

where R is the number of rejected null hypotheses. The intention is to bound the proportion of false positives among the rejections.

- Control of $I(R > 0)\mathrm{FP}/R$ is impossible because $\mathcal{I}$ is unknown, so instead we try and control the **false discovery rate (FDR)**

$$\mathrm{FDR} = \mathrm{E}\{I(R > 0)\mathrm{FP}/R\}.$$

- Strong control is achieved by the **Benjamini–Hochberg procedure**: specify α, then
  - order the P-values as $P_{(1)} \leq \cdots \leq P_{(m)}$ and the hypotheses as $H_{(1)}, \ldots, H_{(m)}$,
  - reject $H_{(1)}, \ldots, H_{(R)}$, where

$$R = \max\left\{r : P_{(r)} < \frac{\alpha r}{m}\right\}.$$

This guarantees that $\mathrm{FDR} \leq \alpha$, but does not bound the actual proportion of false positives, just its expectation. Often $\alpha = 0.1$, 0.2, ...

Note: Derivation of the Benjamini–Hochberg procedure

☐ Let the P-values for the false null hypotheses be $P'_1, \ldots, P'_{m_1}$, say, independent of the true null P-values $P_1, \ldots, P_{m_0} \overset{\text{iid}}{\sim} U(0,1)$. Then

$$\{R = r\} \cap \{P_1 \leq r\alpha/m\} = \{P_1 \leq r\alpha/m\} \cap \{R_{-1} = r-1\},$$

where $\{R_{-1} = r-1\}$ is the event that there are exactly $r-1$ rejections among $H_2, \ldots, H_m$. Then the false discovery proportion is

$$\sum_{r=1}^{m} \frac{\text{FP}}{r} I(R=r) = \sum_{r=1}^{m} \frac{I(R=r)}{r} \sum_{j=1}^{m_0} I(P_j \leq r\alpha/m),$$

and by symmetry of the P_j this has the same expectation as

$$m_0 \sum_{r=1}^{m} \frac{I(R=r)}{r} I(P_1 \leq r\alpha/m) = m_0 \sum_{r=1}^{m} \frac{I(R_{-1}=r-1)}{r} I(P_1 \leq r\alpha/m).$$

Thus the false discovery rate is

$$\begin{aligned}
\text{FDR} &= m_0 \sum_{r=1}^{m} \frac{1}{r} \text{P}(R_{-1} = r-1, P_1 \leq r\alpha/m) \\
&= m_0 \sum_{r=1}^{m} \frac{1}{r} \text{P}(R_{-1} = r-1 \mid P_1 \leq r\alpha/m) \text{P}(P_1 \leq r\alpha/m) \\
&= m_0 \sum_{r=1}^{m} \frac{1}{r} \text{P}(R_{-1} = r-1) \frac{r\alpha}{m} \\
&= \frac{m_0\alpha}{m} \sum_{r=0}^{m-1} \text{P}(R_{-1} = r) \\
&= \frac{m_0\alpha}{m} \leq \alpha.
\end{aligned}$$

The main steps above successively use the definition of conditional probability, the facts that P_1 and R_{-1} are independent and $P_1 \sim U(0,1)$, and the fact that $R_{-1} \in \{0, 1, \ldots, m-1\}$.

☐ We see that the Benjamini–Hochberg procedure strongly controls the FDR under the conditions above.

☐ Note that

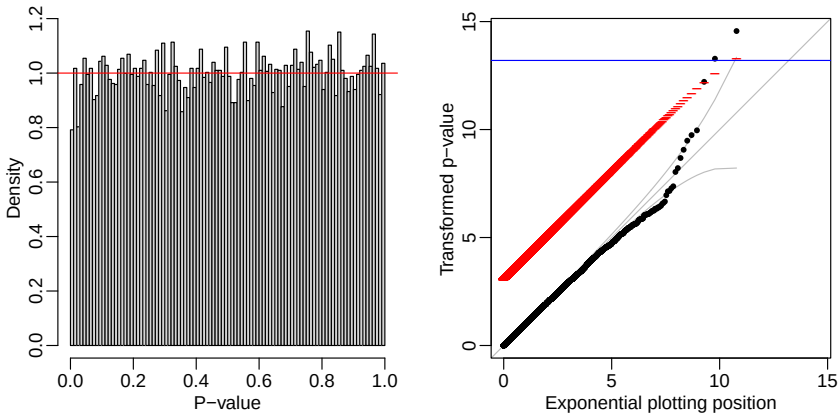
- if $m_0 \ll m$, then the last inequality may be very unequal, so the FDR may in fact be much lower than α.
- if the P-values are dependent in such a way that

$$\text{P}(R_{-1} = r-1 \mid P_1 \leq r\alpha/m) \leq \text{P}(R_{-1} = r-1),$$

then the result also holds.

GWAS, II

- Left: a histogram of $Q_j = 10P_j$ (when $P_j < 0.1$) for tests of the association between $m = 27530$ SNPs and the expression of the protein CFAB.
- Right: exponential Q-Q plot of $Z_j = -\log Q_j$, with Bonferroni cutoff (blue) and Benjamini–Hochberg cutoffs, both with $\alpha = 0.05$.



Comments

- The Bonferroni–Holm procedure compares $P_{(1)}, P_{(2)}, \ldots$ to $\alpha/m, \alpha/(m-1), \ldots$, whereas the ordinary Bonferroni procedure compares all the P_j to α/m.
- The **Simes procedure** (exercises) has exact FWER α for independent tests and then is preferable to the Bonferroni–Holm procedure.
- The Benjamini–Hochberg procedure strongly controls the false discovery rate, comparing the ordered P-values to $\alpha/m, 2\alpha/m, \ldots, \alpha$.
- The first two also give strong control when the P-values are dependent. So does the third, using the comparison

$$P_{(j)} \leq \frac{j\alpha}{mc(m)},$$

with $c(m) = 1$ when the tests are independent or positively dependent, and $c(m) = \sum_{j=1}^{m} 1/j$ under arbitrary dependence.

- Many variants exist, but these versions are simple and widely used.
- Other classical procedures for multiple testing in regression settings are named after
  - Tukey — bounds the maximum of t statistics for different tests;
  - Scheffé — simultaneously bounds all possible linear combinations of estimates $\widehat{\beta}$;
  - Dunnett — compares different treatments with the same control.

Overview

- ☐ In theoretical discussion we glibly write something like

  "Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta) \ldots$"

  but in applications this cannot be taken for granted.
- ☐ Ideally we can ensure random sampling and full measurement of observations from a well-specified population, but if not, possible complications include:
  - selection of observations based on their values, especially truncation;
  - censoring;
  - dependence;
  - missing data.
- ☐ We now briefly discuss these ...

Selection

- ☐ If the available data were selected from a population using a mechanism expressible in probabilistic terms, then the likelihood is

$$\mathrm{P}(Y = y \mid \mathcal{S}; \theta),$$

  where $\mathcal{S}$ is the selection event. If $\mathcal{S}$ is unknown or not probabilistic, only sensitivity analysis is possible (at best).
- ☐ A common example is **truncation** of independent data, where $\mathcal{S}_j = \{Y_j \in \mathcal{I}_j\}$ for some set $\mathcal{I}_j$, giving likelihood

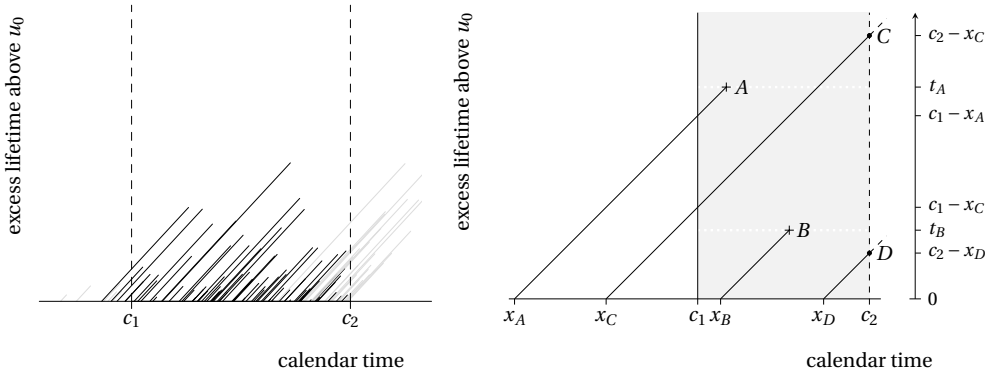
$$\prod_{j=1}^{n} f(y_j \mid y_j \in \mathcal{I}_j; \theta).$$

Example 34 *In certain demographic databases on very old persons, an individual born on calendar date x is included only if they die aged $u_0 + t$, where u_0 is a high threshold (e.g., 100 years) and $t \geq 0$, between two calendar dates c_1 and c_2. The likelihood contribution for this person is then of form*

$$\frac{f(t)}{\mathcal{F}(a) - \mathcal{F}(b)}, \quad a < t < b, \qquad [a, b] = [\max(0, c_1 - x), c_2 - x],$$

where x is the calendar date at which they reach age u_0. See the next page.

Selection in a Lexis diagram



Lexis diagrams showing age on the vertical axis and calendar time on the horizontal axis. Only ages over u_0 are shown.
Left: only the individuals with solid lines appear in the sample.
Right: explanation of the intervals for which different individuals are observed.

Censoring

- Truncation determines which observations appear in a sample, whereas censoring reduces the information available in the sample.
- **Censoring** is very common in studies on lifetime data and leads to the precise values of certain observations being unknown:
  - **right-censoring** results in $(T = \min(Y, b), D = I(Y \leq b))$ for some b;
  - **left-censoring** results in $(T = \max(Y, a), D = I(Y > a))$ for some a;
  - **interval-censoring** results in $(Y, I(a < Y \leq b))$, $(a, I(Y \leq a))$ or $(b, I(Y > b))$, or it is known only which of the disjoint intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$ contains Y.
- In each case we lose information when Y lies within some (possibly random) interval $\mathcal{I}$, often with the assumption that $Y \perp\!\!\!\perp \mathcal{I}$.
- **Rounding** is a form of interval censoring, and we have already seen (exercises) that little information is lost if the rounding is not too coarse.
- Likelihood contributions based on right- and left-censored observations are

$$f_Y(t)^d\{1 - F_Y(t)\}^{1-d}, \quad f_Y(t)^d\{F_Y(t)\}^{1-d}.$$

- Truncation and censoring can arise in the same study; see the Lexis diagram.

Dependent data

- If the joint density of $Y = (Y_1, \ldots, Y_n)$ is known, we can write

$$f(y;\theta) = f(y_1, \ldots, y_n; \theta) = f(y_1;\theta) \prod_{j=2}^{n} f(y_j \mid y_1, \ldots, y_{j-1}; \theta),$$

  the so-called **prediction decomposition**.
- This is most useful if the data arise in time order and satisfy the **Markov property**, that given the 'present' Y_{j-1}, the 'future', $Y_j, Y_{j+1}, \ldots$, is independent of the 'past', $\ldots, Y_{j-3}, Y_{j-2}$, so

$$f(y_j \mid y_1, \ldots, y_{j-1}; \theta) = f(y_j \mid y_{j-1}; \theta)$$

  and the product above simplifies to

$$f(y;\theta) = f(y_1;\theta) \prod_{j=2}^{n} f(y_j \mid y_{j-1}; \theta).$$

- Many variants of this are possible.

Example 35 (Poisson birth process) *Find the likelihood when* $Y_0 \sim \text{Poiss}(\theta)$ *and* $Y_0, \ldots, Y_n$ *are such that* $Y_{j+1} \mid Y_0 = y_0, \ldots, Y_j = y_j \sim \text{Poiss}(\theta y_j)$.

Note to Example 35

Here

$$f(y_{j+1} \mid y_j; \theta) = \frac{(\theta y_j)^{y_{j+1}}}{y_{j+1}!} \exp(-\theta y_j), \qquad y_{j+1} = 0, 1, \ldots, \quad \theta > 0.$$

If Y_0 is Poisson with mean θ, the joint density of data $y_0, \ldots, y_n$ is

$$f(y_0;\theta) \prod_{j=1}^{n} f(y_j \mid y_{j-1}; \theta) = \frac{\theta^{y_0}}{y_0!} \exp(-\theta) \prod_{j=0}^{n-1} \frac{(\theta y_j)^{y_{j+1}}}{y_{j+1}!} \exp(-\theta y_j),$$

so the likelihood is

$$L(\theta) = \left(\prod_{j=0}^{n} y_j! \right)^{-1} \exp\left(s_0 \log\theta - s_1\theta\right), \qquad \theta > 0,$$

where $s_0 = \sum_{j=0}^{n} y_j$ and $s_1 = 1 + \sum_{j=0}^{n-1} y_j$. This is a (2,1) exponential family.

Missing data

- Missing data are widespread in applications, especially those involving living subjects.
- Central problems are:
  - uncertainty increases due to missingness;
  - assumptions about missingness cannot be checked directly, so inferences are fragile.
- Suppose ideal is inference on θ based on n independent pairs (X, Y), but some Y are missing, indicated by a variable I, so we observe either $(x, y, 1)$ or $(x, ?, 0)$.
- The likelihood contributions from individuals with complete data and with y missing are respectively

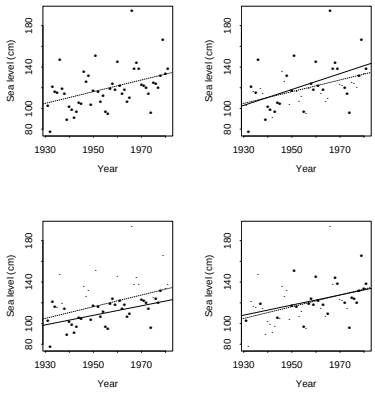
$$\mathrm{P}(I = 1 \mid x, y) f(y \mid x; \theta) f(x; \theta), \qquad \int \mathrm{P}(I = 0 \mid x, y) f(y \mid x; \theta) f(x; \theta) \, \mathrm{d}y,$$

  and there are three possibilities:
  - data are **missing completely at random**, $\mathrm{P}(I = 0 \mid x, y) = \mathrm{P}(I = 0)$;
  - data are **missing at random**, $\mathrm{P}(I = 0 \mid x, y) = \mathrm{P}(I = 0 \mid x)$; and
  - **non-ignorable non-response**, $\mathrm{P}(I = 0 \mid x, y)$ depends on y and maybe on x.

  The first two are sometimes called **ignorable non-response**, as then I has no information about θ and can (mostly) be ignored.

Example

Missing data in straight-line regression for Venice sea-level data. Clockwise from top left: original data, data with values missing completely at random, data with values missing at random — missingness depends on x but not on y, and data with non-ignorable non-response — missingness depends on both x and y. Missing values are represented by a small dot. The dotted line is the fit from the full data, the solid lines those from the non-missing data.



Example

	Truth	Average estimate (average standard error)			
		Full	MCAR	MAR	NIN
β_0	120	120 (2.79)	120 (4.02)	120 (4.73)	132 (3.67)
β_1	0.50	0.49 (0.19)	0.48 (0.28)	0.50 (0.32)	0.20 (0.25)

- Average estimates and standard errors for missing value simulation based on Venice data, for full dataset, with data missing completely at random (MCAR), missing at random (MAR) and with non-ignorable non-response (NIN) and non-response mechanisms

$$\mathrm{P}(I=0 \mid x, y) = \begin{cases} 0.5, \\ \Phi\left\{0.05(x-\overline{x})\right\}, \\ \Phi\left[0.05(x-\overline{x}) + \left\{y-\beta_0-\beta_1(x-\overline{x})\right\}/\sigma\right]; \end{cases}$$

  In each case roughly one-half of the observations are missing.
- Data loss increases the variability of the estimates but their means are unaffected when the non-response is ignorable; otherwise they become entirely unreliable.
- Standard errors for the averages for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are at most 0.16 and 0.01; those for their standard errors are at most 0.03 and 0.002.

Discussion

- Truncation, censoring and other forms of **data coarsening** are widely observed in time-to-event data and there is a huge literature on dealing with them, especially in terms of non- and semi-parametric estimation.
- Selection (especially self-selection!) can totally undermine analyses if ignored or if it can't be modelled appropriately.
- The Markov property plays a key simplifying role in inference based on time series, and generalisations are important in spatial and other types of complex data.
- Missingness is usually the most annoying of the complications above:
  - it is quite common in applications, often for ill-specified reasons;
  - when there is NIN and a non-negligible proportion of the data is missing, correct inference requires us to specify the missingness mechanism correctly;
  - in practice it is hard to tell whether missingness is ignorable, so fully reliable inference is largely out of reach;
  - sensitivity analysis and or bounds to assess how heavily the conclusions depend on plausible mechanisms for non-response is then useful.

Motivation

- ☐ Likelihood
  - provides a general paradigm for inference on parametric models, with many generalisations and variants;
  - is a central concept in both frequentist and Bayesian statistics;
  - has a simple, general and widely-applicable 'large-sample' theory; but
  - is not a panacea!
- ☐ Plan below:
  - recall some basics on convergence;
  - give (fairly) general setup for parameter;
  - prove main results for scalar parameter;
  - discussion of inference;
  - vector parameter, nuisance parameters, ...

Reminders I

Definition 36 *Let $X, X_1, X_2, \ldots$ be random variables with cumulative distribution functions $F, F_1, F_2, \ldots$. Then*
(a) X_n converges to X **in probability**, *$X_n \xrightarrow{P} X$, if $\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) = 0$ for all $\varepsilon > 0$;*
(b) X_n converges to X **in distribution**, *$X_n \xrightarrow{D} X$, if $\lim_{n\to\infty} F_n(x) = F(x)$ at each point x where $F(x)$ is continuous.*
(c) A sequence $X_1, X_2, \ldots$ of estimators of a parameter θ is **(weakly) consistent** *if $X_n \xrightarrow{P} \theta$.*

Theorem 37 *Let x_0, y_0 be constants, $X, Y, \{X_n\}, \{Y_n\}$ random variables and $g(\cdot)$ and $h(\cdot,\cdot)$ continuous functions. Then*

$$\begin{aligned} X_n \xrightarrow{P} X &\Rightarrow X_n \xrightarrow{D} X, \\ X_n \xrightarrow{D} x_0 &\Rightarrow X_n \xrightarrow{P} x_0, \\ X_n \xrightarrow{P} X &\Rightarrow g(X_n) \xrightarrow{P} g(X), \\ X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{D} y_0 &\Rightarrow h(X_n, Y_n) \xrightarrow{D} h(X, y_0). \end{aligned}$$

The last two lines are known as the **continuous mapping theorem** *and* **Slutsky's theorem**.

Reminders II

Theorem 38 (Weak law of large numbers) *If $X, X_1, X_2, \ldots$ are independent identically distributed random variables and $\mathrm{E}(X)$ is finite, then $\overline{X} = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{P} \mathrm{E}(X)$.*

Theorem 39 (Central limit theorem, CLT) *If $X_1, X_2, \ldots \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ and $0 < \sigma^2 < \infty$, then*

$$Z_n = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

Theorem 40 ('Delta method') *If $a_n(X_n - \mu) \xrightarrow{D} Z$, where $a_n, \mu \in \mathbb{R}$ for all n, $a_n \to \infty$ as $n \to \infty$, and g is continuously differentiable at μ, then $a_n\{g(X_n) - g(\mu)\} \xrightarrow{D} g'(\mu)Z$.*

- ☐ Many more general laws of large numbers and versions of the CLT exist.
- ☐ The delta method also applies with $X_n, Z \in \mathbb{R}^p$, $g(x) : \mathbb{R}^p \to \mathbb{R}^q$ continuously differentiable and $g'(\mu)$ replaced by $J_g(\mu) = \partial g(\mu)/\partial \mu^{\mathrm{T}}$.

Theorem 41 *If X is a random variable, $a > 0$ a constant, h a non-negative function and g a convex function, then*

$$\begin{aligned} \mathrm{P}\{h(X) \geq a\} &\leq \mathrm{E}\{h(X)\}/a, \quad \textit{(basic inequality)}, \\ g\{\mathrm{E}(X)\} &\leq E\{g(X)\}, \quad \textit{(Jensen's inequality)}. \end{aligned}$$

Basic setup

- ☐ Let $Y, Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, and define the **Kullback–Leibler divergence** from g to a density f,

$$\mathrm{KL}(g, f) = \mathrm{E}_g\{\log g(Y) - \log f(Y)\} = \mathrm{E}_g\left[-\log\left\{\frac{f(Y)}{g(Y)}\right\}\right] \geq 0,$$

  where the inequality follows because $-\log x$ is convex and is strict unless $f \equiv g$.
- ☐ g is the unknown model from which the Y_j are drawn, and f is a candidate model.
- ☐ In a parametric setting there is a family of models, $f \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$, so minimising $\mathrm{KL}(g, f)$ over f is equivalent to maximising $\mathrm{E}_g \log f(Y; \theta)$, which is estimated by

$$\overline{\ell}(\theta) = n^{-1} \sum_{j=1}^{n} \log f(Y_j; \theta) \xrightarrow{P} \mathrm{E}_g \log f(Y; \theta), \quad n \to \infty.$$

- ☐ $\theta_g = \arg\max_\theta \mathrm{E}_g \log f(Y; \theta)$ gives the best fit of f_θ to g.
- ☐ In an ideal case $g = f_{\theta_g}$, i.e., $g \in \mathcal{F}$, but the theory does not require this (yet).
- ☐ $\widehat{\theta} = \arg\max_\theta \overline{\ell}(\theta)$ is the natural estimator of θ_g, and we hope that $\widehat{\theta} \xrightarrow{P} \theta_g$ as $n \to \infty$.
- ☐ Unfortunately this requires conditions to restrict the variation of $\overline{\ell}$ as it converges.

Regular models

- ☐ Recall the notation $\nabla g(\theta) = \partial g(\theta)/\partial\theta$ and $\nabla^2 g(\theta) = \nabla\nabla^{\mathrm{T}} g(\theta) = \partial^2 g(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$.
- ☐ The following conditions ensure consistency and asymptotic normality of the MLE:
  - **(C1)** θ_g is interior to $\Theta \subset \mathbb{R}^d$ for some finite d, and Θ is compact;
  - **(C2)** the densities f_θ defined by any two different values of $\theta \in \Theta$ are distinct;
  - **(C3)** there is a neighbourhood $\mathcal{N}$ of θ_g within which the first three derivatives of the log likelihood with respect to θ exist almost surely, and for $r, s, t = 1, \ldots, d$ satisfy $|\partial^3 \log f(Y;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| < m(Y)$ with $\mathrm{E}_g\{m(Y)\} < \infty$; and
  - **(C4)** within $\mathcal{N}$, the $d \times d$ matrices

$$I_1(\theta) = \mathrm{E}_g\left\{-\nabla^2 \log f(Y;\theta)\right\}, \quad K_1(\theta) = \mathrm{E}_g\left\{\nabla \log f(Y;\theta)\nabla^{\mathrm{T}} \log f(Y;\theta)\right\},$$

    are finite and positive definite. When $g = f_{\theta_g}$ we shall see that $K_1(\theta_g) = I_1(\theta_g)$.
- ☐ Comments:
  - (C1) ensures that $\widehat{\theta}$ can be 'on all sides' of θ_g in the limit;
  - (C2) is essential for consistency, otherwise $\widehat{\theta}$ might not converge;
  - (C3) is a technical condition needed to bound terms of a Taylor series; and
  - (C4) ensures that the asymptotic variance of $\widehat{\theta}$ is positive definite.

Consistency of the MLE

Lemma 42 *If θ is scalar, then a sequence of maximum likelihood estimators $\widehat{\theta}$ exists such that $\widehat{\theta} \xrightarrow{P} \theta_g$.*

This result:

- ☐ does not require f_θ to be smooth, so it is quite general;
- ☐ guarantees that a consistent sequence exists, but not that we can find it;
- ☐ but if the log likelihood is convex (as in exponential families, for example), then there is (at most) one maximum for any n, and if it exists this must converge to θ_g;
- ☐ can be generalized to vector θ, but the argument is more delicate;
- ☐ van der Vaart (1998, *Asymptotic Statistics*, Chapter 5) gives more general proofs.

Note to Lemma 42

- ☐ As the θs correspond to different densities, there is precisely one θ_g that minimises $\mathrm{KL}(g, f_\theta)$.
- ☐ Take any $\varepsilon > 0$ and let $\theta_+, \theta_- = \theta_g \pm \varepsilon$, write $D_n(\theta) = \overline{\ell}(\theta_g) - \overline{\ell}(\theta)$, so $D_n(\theta_g) = 0$, and note that as $n \to \infty$,

$$D_n(\theta_+) \xrightarrow{P} \mathrm{KL}(g, f_{\theta_+}) - \mathrm{KL}(g, f_{\theta_g}) = a_+ > 0, \quad D_n(\theta_-) \xrightarrow{P} \mathrm{KL}(g, f_{\theta_-}) - \mathrm{KL}(g, f_{\theta_g}) = a_- > 0.$$

- ☐ If A_n and B_n denote the events $D_n(\theta_+) > 0$ and $D_n(\theta_-) > 0$, Boole's inequality gives

$$\mathrm{P}(A_n \cap B_n) = 1 - \mathrm{P}(A_n^c \cup B_n^c) \geq 1 - \mathrm{P}(A_n^c) - \mathrm{P}(B_n^c).$$

  Now

$$\mathrm{P}(A_n^c) = \mathrm{P}\{D_n(\theta_+) \leq 0\} = \mathrm{P}\{a_+ - D_n(\theta_+) \geq a_+\} \leq \mathrm{P}\{|D_n(\theta_+) - a_+| \geq a_+\} \to 0, \quad n \to \infty,$$

  and likewise $\mathrm{P}(B_n^c) \to 0$. Hence $\mathrm{P}(A_n \cap B_n) \to 1$.
- ☐ Hence there is a local minimum of $D_n(\theta)$, or equivalently a local maximum of $\overline{\ell}(\theta)$, inside the interval $(\theta_g - \epsilon, \theta_g + \epsilon)$ with probability one as $n \to \infty$, and as this is true for arbitrary ε, the corresponding sequence of maximisers $\widehat{\theta}$ satisfies $\mathrm{P}(|\widehat{\theta} - \theta_g| > \varepsilon) \to 0$ and therefore is consistent.

Asymptotic normality of the MLE

Theorem 43 *If θ is scalar and the regularity conditions hold, then the sequence of consistent maximum likelihood estimators $\widehat{\theta}$ satisfies*

$$n^{1/2}(\widehat{\theta} - \theta_g) \xrightarrow{D} \mathcal{N}_d\{0, I_1^{-1}(\theta_g) K_1(\theta_g) I_1^{-1}(\theta_g)\},$$

where $d = 1$ and for a single observation Y we define

$$I_1(\theta) = \mathrm{E}_g\left\{-\nabla^2 \log f(Y;\theta)\right\}, \quad K_1(\theta) = \mathrm{E}_g\left\{\nabla \log f(Y;\theta) \nabla^{\mathrm{T}} \log f(Y;\theta)\right\}.$$

- ☐ In the vector case, $d > 1$, the above **sandwich variance matrix** expression also applies.
- ☐ This implies that for large n,

$$\widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_d\{\theta_g, I^{-1}(\theta_g) K(\theta_g) I^{-1}(\theta_g)\},$$

  where $I(\theta) = nI_1(\theta)$, $K(\theta) = nK_1(\theta)$ correspond to a sample of size n.
- ☐ This provides tests and confidence intervals based on the approximate pivots

$$v_{rr}^{-1/2}(\widehat{\theta}_r - \theta_{g,r}) \mathrel{\dot\sim} \mathcal{N}(0,1), \quad r = 1, \ldots, d,$$

  where v_{rr} are the diagonal elements of an estimate of $I^{-1}(\theta_g) K(\theta_g) I^{-1}(\theta_g)$.
- ☐ When $g = f_{\theta_g}$, $I(\theta_g) = K(\theta_g)$ and the variance (matrix) becomes $I(\theta_g)^{-1}$.

Note to Theorem 43

- We first note that under the given conditions, θ_g gives a stationary point of $\mathrm{KL}(g, f_\theta)$, and therefore

$$0 = \nabla \mathrm{KL}(g, f_\theta)|_{\theta=\theta_g} = -\nabla \int \log f(y;\theta) g(y)\,\mathrm{d}y \Big|_{\theta=\theta_g} = -\int \nabla \log f(y;\theta)\Big|_{\theta=\theta_g} g(y)\,\mathrm{d}y,$$

  so $\mathrm{E}_g\{\nabla \log f(Y;\theta)\} = 0$.

- As $\widehat{\theta}$ gives a local maximum of the differentiable function $\overline{\ell}(\theta) = n^{-1}\sum_{j=1}^n \log f(Y_j;\theta)$,

$$0 = \nabla \overline{\ell}(\widehat{\theta}) = n^{-1}\sum_{j=1}^n \nabla \log f(Y_j;\widehat{\theta}),$$

  and (supposing now that θ is scalar, to simplify the expressions), Taylor series expansion gives

$$0 = \nabla \overline{\ell}(\theta_g) + (\widehat{\theta}-\theta_g)\nabla^2 \overline{\ell}(\theta_g) + \tfrac{1}{2}(\widehat{\theta}-\theta_g)^2 \nabla^3 \overline{\ell}(\theta^*),$$

  where θ^* lies between θ_g and $\widehat{\theta}$ (so $\theta^* \xrightarrow{P} \theta_g$), and hence we can write

$$n^{1/2}(\widehat{\theta}-\theta_g) = \frac{n^{1/2}\nabla \overline{\ell}(\theta_g)}{-\nabla^2 \overline{\ell}(\theta_g) - R_n/2}, \quad R_n = (\widehat{\theta}-\theta_g)\nabla^3 \overline{\ell}(\theta^*). \tag{4}$$

- Now

$$n^{1/2}\nabla \overline{\ell}(\theta_g) = n^{-1/2}\sum_{j=1}^n \nabla \log f(Y_j;\theta_g)$$

  has mean (vector) zero and variance (matrix)

$$\mathrm{var}\left\{ n^{-1/2}\sum_{j=1}^n \nabla \log f(Y_j;\theta_g)\right\} = n^{-1}\sum_{j=1}^n \mathrm{E}_g\{\nabla \log f(Y_j;\theta_g)\nabla^{\mathrm{T}} \log f(Y_j;\theta_g)\} = K_1(\theta_g).$$

  so the numerator of (4) converges in distribution to $\mathcal{N}\{0, K_1(\theta_g)\}$, using the CLT.

- Moreover the weak law of large numbers gives

$$-\nabla^2 \overline{\ell}(\theta_g) = -\frac{1}{n}\sum_{j=1}^n \nabla^2 \log f(Y_j;\theta_g) \xrightarrow{P} I_1(\theta_g).$$

- Lemma 44 shows that $R_n \xrightarrow{P} 0$, so the denominator of (4) tends in probability to $I_1(\theta_g)$.

- Putting the pieces together, we find that

$$n^{1/2}(\widehat{\theta}-\theta_g) \xrightarrow{D} \mathcal{N}_d\{0, I_1(\theta_g)^{-1}K_1(\theta_g)I_1(\theta_g)^{-1}\}, \quad n\to\infty,$$

  where the variance formula is also valid when I_1 and K_1 are $d\times d$ matrices.

- The information quantities based on a random sample of size n are $I(\theta_g) = nI_1(\theta_g)$ and $K(\theta_g) = nK_1(\theta_g)$, giving

$$\widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_d(\theta_g, I(\theta_g)^{-1}K(\theta_g)I(\theta_g)^{-1}\},$$

  in which the variance is of the usual order $1/n$.

Note: Lemma 44

Lemma 44 *Under the conditions of Theorem 43,* $R_n = (\widehat{\theta} - \theta_g)\nabla^3\overline{\ell}(\theta^*) \xrightarrow{P} 0$ *as* $n \to \infty$.

☐ For $\varepsilon > 0$, $B_n = \{|R_n| > \varepsilon\}$, $A_n = \{|\widehat{\theta} - \theta_g| > \delta\}$ and $\delta > 0$ small enough that $\mathcal{N}$ contains a ball of radius δ around θ_g, we have

$$\mathrm{P}(|R_n| > \varepsilon) = \mathrm{P}(B_n \cap A_n) + \mathrm{P}(B_n \cap A_n^c) \leq \mathrm{P}(A_n) + \mathrm{P}(B_n \cap A_n^c),$$

where the first term tends to zero because the sequence $\widehat{\theta}$ is consistent.

☐ If $|\widehat{\theta} - \theta_g| < \delta$, then (C3) implies that

$$|R_n| \leq \delta n^{-1} \sum_{j=1}^{n} |\partial^3 \log f(Y_j; \theta^*)/\partial\theta^3| \leq \delta n^{-1} \sum_{j=1}^{n} m(Y_j) = \delta\overline{M}_n,$$

say, and clearly $\overline{M}_n \xrightarrow{P} M$, say. Therefore

$$\mathrm{P}(B_n \cap A_n^c) = \mathrm{P}(B_n \cap |\widehat{\theta} - \theta_g| > \delta) \leq \mathrm{P}(B_n \cap |R_n| \leq \delta\overline{M}_n)$$

and for $\eta > 0$ this equals

$$\mathrm{P}(B_n \cap |R_n| \leq \delta\overline{M}_n \cap \overline{M}_n \leq M + \eta) + \mathrm{P}(B_n \cap |R_n| \leq \delta\overline{M}_n \cap \overline{M}_n > M + \eta),$$

which is bounded by

$$\mathrm{P}\{|R_n| > \varepsilon \cap |R_n| \leq \delta(M + \eta)\} + \mathrm{P}(|\overline{M}_n - M| > \eta).$$

The last term here tends to zero, because $\overline{M}_n \xrightarrow{P} M$, and the first can be made equal to zero by choosing δ such that $\delta(M + \eta) < \varepsilon$. This proves the lemma.

Classical asymptotics

- ☐ The true model is supposed to lie in the candidate family, i.e., $g \in \mathcal{F}$, so $\theta_g \in \Theta$.
- ☐ We can differentiate under the integral sign and get the **Bartlett identities**:

$$\begin{aligned} 1 &= \int f(y;\theta)\,\mathrm{d}y, \\ 0 &= \int \nabla \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y, \\ 0 &= \int \nabla^2 \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y + \int \nabla \log f(y;\theta)\, \nabla^{\mathrm{T}} \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y, \\ 0 &= \cdots \end{aligned}$$

  giving the moments of the $d \times 1$ **score vector** $U(\theta) = \nabla \ell(\theta)$, viz

$$\mathrm{E}\{U(\theta)\} = 0, \quad \mathrm{var}\{U(\theta)\} = \mathrm{E}\left\{\nabla \ell(\theta) \nabla^{\mathrm{T}} \ell(\theta)\right\} = \mathrm{E}\left\{-\nabla^2 \ell(\theta)\right\}, \quad \ldots$$

- ☐ Hence $I(\theta) = K(\theta)$, and $I(\theta) = nI_1(\theta) = nK_1(\theta)$ when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$.
- ☐ The assumption that $g \in \mathcal{F}$ is technically always false, but this is irrelevant if model-checking suggests that $\mathcal{F}$ is 'close enough' to g.
- ☐ Crucially, the interest parameter ψ should have a stable interpretation for candidates likely to be close to g (i.e., within $n^{-1/2}$) , so $\mathcal{F}$ is 'robustly specified'.

Note to the Bartlett identities

- ☐ The first is true for any θ, and provided we can exchange the order of integration and differentiation we have

$$0 = \nabla \int f(y;\theta)\,\mathrm{d}y = \int \nabla f(y;\theta)\,\mathrm{d}y = \int \nabla f(y;\theta) \frac{f(y;\theta)}{f(y;\theta)}\,\mathrm{d}y = \int \nabla \log f(y;\theta)\, f(y;\theta)\,\mathrm{d}y.$$

- ☐ The second stems from a second differentiation and applying the chain rule to the terms in the final integral here; likewise for the third and higher-order ones, which give higher-order moments of $U(\theta)$.
- ☐ For independent data $Y_1, \ldots, Y_n$ we have $U(\theta) = \sum_{j=1}^n U_j(\theta)$, where the $U_j = \nabla \log f(Y_j;\theta)$ are independent, so using the Bartlett identities for the individual densities $f_j(y_j;\theta)$ we have

$$\mathrm{var}\{U(\theta)\} = \sum_{j=1}^n \mathrm{var}\{U_j(\theta)\} = \sum_{j=1}^n \mathrm{E}\{U_j(\theta) U_j^{\mathrm{T}}(\theta)\} = \sum_{j=1}^n -\mathrm{E}\{\nabla^{\mathrm{T}} U_j(\theta)\} = -\mathrm{E}\left\{\nabla^{\mathrm{T}} U(\theta)\right\}$$

  and this equals $\mathrm{E}\left\{-\nabla^2 \ell(\theta)\right\} = I(\theta)$, and this in turn equals $nI_1(\theta)$ if $Y_j \overset{\text{iid}}{\sim} f_{\theta_g}$.

In practice ...

- We usually assume classical asymptotics and replace the sandwich matrix

$$I(\theta_g)^{-1}K(\theta_g)I(\theta_g)^{-1} \quad \text{by the observed information matrix} \quad \widehat{\jmath} = -\nabla^2\ell(\widehat{\theta}),$$

  which
  - can be computed numerically without (possibly awkward) expectations,
  - will (helpfully!) misbehave if the maximisation is questionable,
  - has been found to give generally good results in applications,
  - has the heuristic justification that $(\widehat{\theta}, \widehat{\jmath})$ are approximately sufficient for θ_g, as

$$\ell(\theta_g) \doteq \ell(\widehat{\theta}) - \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}}\widehat{\jmath}(\widehat{\theta} - \theta_g).$$

- Standard errors for $\widehat{\theta}$ are the square roots of the diagonal elements of $\widehat{\jmath}^{-1}$.
- To make the sandwich we can replace $I(\theta_g)$ by $\widehat{\jmath}$ and $K(\theta_g)$ by (some version of)

$$\widehat{K} = \sum_{j=1}^{n} \nabla \log f(Y_j; \widehat{\theta}) \nabla^{\mathrm{T}} \log f(Y_j; \widehat{\theta}),$$

  though $\widehat{\jmath}^{-1}\widehat{K}\,\widehat{\jmath}^{-1}$ can unstable because of numerical problems with $\widehat{K}$.

Related statistics

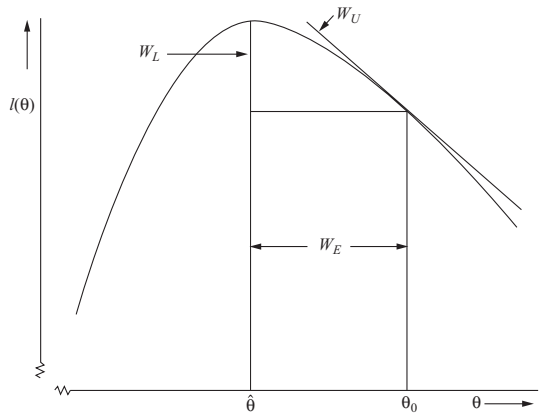


Figure 6.2. Three asymptotically equivalent ways, all based on the log likelihood function of testing null hypothesis $\theta = \theta_0$: W_E, horizontal distance; W_L vertical distance; W_U slope at null point.

From Cox (2006, *Principles of Statistical Inference*)

Related statistics

☐ When θ is scalar the asymptotic arguments support inference based on any of the pivots

$$T = t(\theta_g) = \widehat{\jmath}^{1/2}(\widehat{\theta} - \theta_g) \dot{\sim} \mathcal{N}(0,1), \qquad \textbf{Wald statistic},$$
$$S = s(\theta_g) = \widehat{\jmath}^{-1/2} U(\theta_g) \dot{\sim} \mathcal{N}(0,1), \qquad \textbf{score statistic},$$
$$W = w(\theta_g) = 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} \dot{\sim} \chi^2_1, \qquad \textbf{likelihood ratio statistic},$$
$$R = r(\theta_g) = \mathrm{sign}(\widehat{\theta} - \theta_g) w(\theta_g)^{1/2} \dot{\sim} \mathcal{N}(0,1), \qquad \textbf{likelihood root}.$$

The likelihood root has other names (e.g., directed likelihood ratio statistic).

☐ The distribution of W follows from the expansion on the previous slide.

☐ If $\widehat{\theta}^{\mathrm{o}}$ and $\jmath(\widehat{\theta}^{\mathrm{o}})$ have been obtained for observed data y^{o}, then the approximation

$$\mathrm{P}_g\{T(\theta_g) \leq t^{\mathrm{o}}(\theta_g)\} \doteq \Phi\{t^{\mathrm{o}}(\theta_g)\}$$

leads to $(1-\alpha)$ confidence interval $\widehat{\theta}^{\mathrm{o}} \pm \jmath(\widehat{\theta}^{\mathrm{o}})^{-1/2} z_{1-\alpha/2}$ based on T, while that based on W is

$$\{\theta : W^{\mathrm{o}}(\theta) \leq \chi^2_1(1-\alpha)\} = \{\theta : \ell^{\mathrm{o}}(\theta) \geq \ell^{\mathrm{o}}(\widehat{\theta}^{\mathrm{o}}) - \tfrac{1}{2}\chi^2_1(1-\alpha)\},$$

where z_p and $\chi^2_\nu(p)$ are respectively the p quantiles of the $N(0,1)$ and χ^2_ν distributions.

Comments

☐ Comparative comments:

- confidence intervals based on T are symmetric, but those based on W or R take the shape of ℓ into account and are parametrisation-invariant;
- in small samples the distributional approximations for W and R are better than that for T, and that for W can be improved by **Bartlett correction**, using $W_{\mathrm{B}} = W/(1 + b/n)$;
- confidence sets based on W may not be connected (and if so those based on T or R are unreliable);
- the main use of S is for testing in situations where maximisation of ℓ is awkward, and then $\widehat{\jmath}$ is often replaced by $I(\theta_g)$;
- a variant of R, the **modified likelihood root**

$$R^* = r^*(\theta_g) = r(\theta_g) + \frac{1}{r(\theta_g)} \log \frac{q(\theta_g)}{r(\theta_g)},$$

often gives almost perfect inferences even in small samples (more later ...).

Example 45 *Compute the above statistics when* $y_1, \ldots, y_n \overset{\mathrm{iid}}{\sim} \exp(\theta)$ *and compare the resulting inferences with those from an exact pivot.*

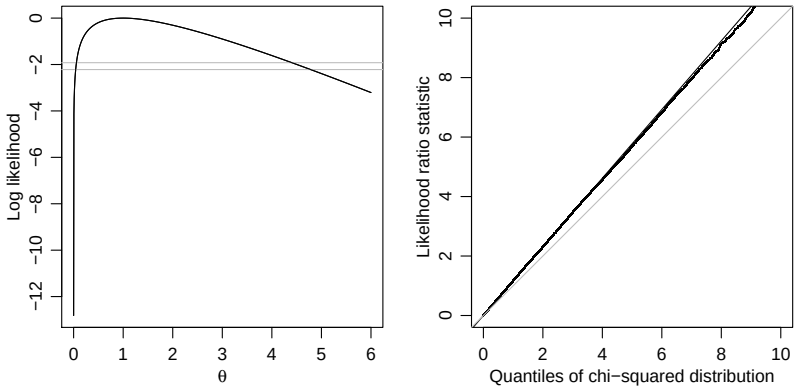
Note to Example 45

- ☐ The log likelihood is $\ell(\theta) = n(\log\theta - \theta\overline{y})$, for $\theta > 0$, which is clearly unimodal with $\widehat{\theta} = 1/\overline{y}$ and $\jmath(\theta) = n/\theta^2$.
- ☐ Hence

$$\begin{aligned} t(\theta) &= n^{1/2}(1-\theta\,\overline{y}), \\ s(\theta) &= n^{1/2}\{1/(\theta\,\overline{y}) - 1\}, \\ w(\theta) &= 2n\,\{\theta\,\overline{y} - \log(\theta\,\overline{y}) - 1\}\,, \\ r(\theta) &= \mathrm{sign}(1-\theta\,\overline{y})\,[2n\,\{\theta\,\overline{y} - \log(\theta\,\overline{y}) - 1\}]^{1/2}\,. \end{aligned}$$

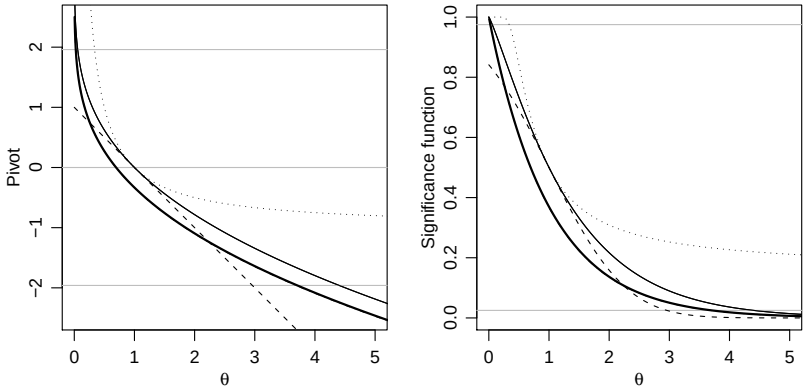
- ☐ The exact pivot is $\theta\sum Y_j$ whose distribution is gamma with unit scale and shape parameter n.
- ☐ Consider an exponential sample with $n = 1$ and $\overline{y} = 1$; then $\widehat{\jmath} = 1$. The log likelihood $\ell(\theta)$, shown in the left-hand panel of the figure, is unimodal but strikingly asymmetric, suggesting that confidence intervals based on an approximating normal distribution for $\widehat{\theta}$ will be poor. The right-hand panel is a chi-squared probability plot in which the ordered values of simulated $w(\theta)$ are graphed against quantiles of the χ^2_1 distribution—if the simulations lay along the diagonal line $x = y$, then this distribution would be a perfect fit. The simulations do follow a straight line rather closely, but with slope $(1 + b/n)\chi^2_1$, where $b = 0.1544$. This indicates that the distribution of the Bartlett-adjusted likelihood ratio statistic $w(\theta)/(1 + b/n)$ would be essentially χ^2_1. The 95% confidence intervals for θ based on the unadjusted and adjusted likelihood ratio statistics are $(0.058, 4.403)$ and $(0.042, 4.782)$ respectively.

Exponential example



Likelihood inference for exponential sample of size $n = 1$. Left: log likelihood $\ell(\theta)$. Intersection of the function with the two horizontal lines gives two 95% confidence intervals for θ: the upper line is based on the χ^2_1 approximation to the distribution of $w(\theta)$, and the lower line is based on the Bartlett-corrected statistic. Right: comparison of simulated values of likelihood ratio statistic $w(\theta)$ with χ^2_1 quantiles. The χ^2_1 approximation is shown by the line of unit slope, while the $(1 + b/n)\chi^2_1$ approximation is shown by the upper straight line.

Exponential example



Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_j$ (dot-dash). The modified likelihood root is indistinguishable from the exact pivot. The horizontal lines are at $0, \pm 1.96$. Right: corresponding significance functions, with horizontal lines at 0.025 and 0.975.

Vector case

☐ When θ is a vector and under classical asymptotics we base inference on the approximations

$$\widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_d(\theta_g, \widehat{\jmath}^{-1}), \quad w(\theta_g) = 2\left\{\ell(\widehat{\theta}) - \ell(\theta_g)\right\} \mathrel{\dot\sim} \chi^2_d, \quad s(\theta_g) = \widehat{\jmath}^{-1/2} U(\theta_g) \mathrel{\dot\sim} \mathcal{N}_d(0, I_d),$$

with

- the first very commonly used for inferences on parameters;
- the second used to test whether $\theta = \theta_g$;
- the third much less used than the others, generally in the form $s(\theta_g)^{\mathrm{T}} s(\theta_g) \mathrel{\dot\sim} \chi^2_d$.

☐ If θ divides into a $p \times 1$ **interest parameter** ψ and a $q \times 1$ **nuisance parameter** λ, then

$$\widehat{\theta} = \begin{pmatrix} \widehat{\psi} \\ \widehat{\lambda} \end{pmatrix} \mathrel{\dot\sim} \mathcal{N}_{p+q}\left\{ \begin{pmatrix} \psi_g \\ \lambda_g \end{pmatrix}, \begin{pmatrix} \widehat{\jmath}_{\psi\psi} & \widehat{\jmath}_{\psi\lambda} \\ \widehat{\jmath}_{\lambda\psi} & \widehat{\jmath}_{\lambda\lambda} \end{pmatrix}^{-1} \right\},$$

where for brevity we now write $\widehat{\lambda}_\psi = \max_\lambda \ell(\psi, \lambda)$, $\tilde{\theta} = \widehat{\theta}_\psi = (\psi, \widehat{\theta}_\psi)$,

$$\ell_\psi = \left.\frac{\partial \ell(\theta)}{\partial \psi}\right|_{\theta=\theta_g}, \quad \widehat{\jmath}_{\psi\psi} = -\widehat{\ell}_{\psi\psi} = -\left.\frac{\partial^2 \ell(\theta)}{\partial \psi \partial \psi^{\mathrm{T}}}\right|_{\theta=\widehat{\theta}}, \quad \tilde{\ell}_{\psi\psi} = \left.\frac{\partial^2 \ell(\theta)}{\partial \psi \partial \psi^{\mathrm{T}}}\right|_{\theta=\tilde{\theta}}, \quad \text{etc.}$$

Inference on ψ

- Under classical asymptotics and setting $\widehat{\jmath}^{\psi\psi} = (\widehat{\jmath}_{\psi\psi} - \widehat{\jmath}_{\psi\lambda}\widehat{\jmath}_{\lambda\lambda}^{-1}\widehat{\jmath}_{\lambda\psi})^{-1}$ we have

$$\widehat{\psi} \dot{\sim} \mathcal{N}_p\left(\psi_g, \widehat{\jmath}^{\psi\psi}\right) \qquad \text{maximum likelihood estimator},$$
$$w_{\mathrm{p}}(\psi_g) = 2\left\{\ell_{\mathrm{p}}(\widehat{\psi}) - \ell_{\mathrm{p}}(\psi_g)\right\} \dot{\sim} \chi^2_p \qquad \text{(generalized) likelihood ratio statistic},$$
$$s(\psi_g) = \tilde{\ell}_{\psi}^{\mathrm{T}}\tilde{\jmath}^{\psi\psi}\tilde{\ell}_{\psi} \dot{\sim} \chi^2_p \qquad \text{score statistic},$$

  where we defined w_{p} using the **profile log likelihood** $\ell_{\mathrm{p}}(\psi) = \ell(\psi, \widehat{\lambda}_{\psi}) = \max_{\lambda} \ell(\psi, \lambda)$.
- If ψ is scalar ($p = 1$, the usual situation), the **likelihood root**

$$r(\psi_g) = \mathrm{sign}\left(\widehat{\psi} - \psi_g\right)\sqrt{w(\psi_g)} \dot{\sim} \mathcal{N}(0, 1).$$

- Properties:
  - inferences using $w(\psi_g)$ and $r(\psi_g)$ are invariant to interest-respecting reparametrisation, so are preferable but more computationally burdensome;
  - $s(\psi_g)$ is mainly used for tests, since only λ must be estimated (as $\psi = \psi_g$ is known).
- A $(1-\alpha)$ confidence set based on $w_{\mathrm{p}}(\psi_g)$ (or equivalently on $\ell_{\mathrm{p}}(\psi)$) is

$$\left\{\psi : w_{\mathrm{p}}(\psi) \leq \chi^2_p(1-\alpha)\right\} = \left\{\psi : \ell(\psi, \widehat{\lambda}_{\psi}) \geq \ell(\widehat{\psi}, \widehat{\lambda}) - \tfrac{1}{2}\chi^2_p(1-\alpha)\right\}.$$

Note: Large-sample distribution of the likelihood ratio statistic $w_{\mathrm{p}}(\psi_g)$

☐ We write

$$w_{\mathrm{p}}(\psi_g) = 2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\} = 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} - 2\{\ell(\widehat{\theta}_\psi) - \ell(\theta_g)\}$$

and shall use Taylor series to approximate both terms by quadratic forms in $\widehat{\theta} - \theta_g$ and $\widehat{\lambda}_\psi - \lambda_g$.

☐ We shall need to express ℓ_θ, ℓ_λ and $\widehat{\lambda}_\psi - \lambda_g$ in terms of $\widehat{\theta} - \theta_g$. Taylor expansion gives

$$0 = \widehat{\ell}_\theta = \ell_\theta + \ell_{\theta\theta}(\widehat{\theta} - \theta_g) + \cdots = \ell_\theta - \imath_{\theta\theta}(\widehat{\theta} - \theta_g) + \cdots,$$

where $\imath_{\theta\theta}$ denotes the expected information matrix evaluated at θ_g and $\cdots$ denotes terms of smaller order containing third derivatives. Likewise

$$0 = \tilde{\ell}_\lambda = \ell_\lambda + \ell_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) + \cdots = \ell_\lambda - \imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) + \cdots.$$

This implies that

$$\ell_\lambda \doteq \imath_{\lambda\psi}(\widehat{\psi} - \psi_g) + \imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g) = \imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g),$$

so the necessary approximations are

$$\ell_\theta \doteq \imath_{\theta\theta}(\widehat{\theta} - \theta_g), \quad \ell_\lambda \doteq \imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g), \quad \widehat{\lambda}_\psi - \lambda_g \doteq \widehat{\lambda} - \lambda_g + \imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g).$$

☐ To obtain the quadratic forms we write

$$\begin{aligned}\ell(\widehat{\theta}) &= \ell(\theta_g) + (\widehat{\theta} - \theta_g)^{\mathrm{T}}\ell_\theta + \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}}\ell_{\theta\theta}(\widehat{\theta} - \theta_g) + \cdots \\ &\doteq \ell(\theta_g) + (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g) - \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g),\end{aligned}$$

resulting in

$$2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} \doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g),$$

and with a similar expression for $2\{\ell(\widehat{\theta}_\psi) - \ell(\theta_g)\}$ we obtain

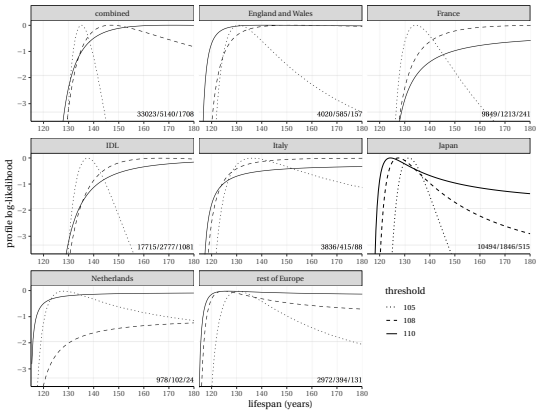
$$\begin{aligned}w_{\mathrm{p}}(\psi_g) &\doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g) - (\widehat{\lambda}_\psi - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) \\ &\doteq (\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\psi}(\widehat{\psi} - \psi_g) + 2(\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\lambda}(\widehat{\lambda} - \lambda_g) + (\widehat{\lambda} - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g) \\ &\quad - \left\{(\widehat{\lambda} - \lambda_g) + \imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g)\right\}^{\mathrm{T}}\imath_{\lambda\lambda}\left\{(\widehat{\lambda} - \lambda_g) + \imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi}(\widehat{\psi} - \psi_g)\right\} \\ &= (\widehat{\psi} - \psi_g)^{\mathrm{T}}(\imath_{\psi\psi} - \imath_{\psi\lambda}\imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi})(\widehat{\psi} - \psi_g),\end{aligned}$$

and as $\widehat{\psi} \mathrel{\dot\sim} \mathcal{N}\{\psi_g, (\imath_{\psi\psi} - \imath_{\psi\lambda}\imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi})^{-1}\}$, we see that $w_{\mathrm{p}}(\psi_g) \mathrel{\dot\sim} \chi^2_p$, as claimed.

☐ Arguments along the lines of Lemma 44 show that the terms dropped above all tend in probability to zero, and thus do not affect the approximation.

Example: Human lifespan

Example 46 *The figure below shows profile log likelihoods for the endpoint ψ of a generalized Pareto distribution fitted to data on lifetimes of semi-supercentenarians from different databases, with thresholds at 105, 108, 110 years. Here λ is scalar, so $p = q = 1$, and the horizontal line at $-\frac{1}{2}\chi^2_1(0.95) = -1.92$ indicates 95% confidence regions.*



From Belzile et al. (2022, *Annual Reviews of Statistics and its Application*).

Model selection

- ☐ The fact that the Kullback–Leibler divergence

$$\mathrm{KL}(g, f) = \mathrm{E}_g\{\log g(Y) - \log f(Y)\} = \mathrm{E}_g\left[-\log\left\{\frac{f(Y)}{g(Y)}\right\}\right] \geq 0,$$

  is minimised when $f = g$ suggested that we compare competing models $\mathcal{F}_1, \ldots, \mathcal{F}_M$ in terms of their maximised log likelihoods $\log f_m(y; \widehat{\theta}_m) = \widehat{\ell}_m$.
- ☐ But $\widehat{\ell}_m$ should be penalized, because
  - $\widehat{\ell}_m \geq \log f_m(y; \theta_m)$ even if $\mathcal{F}_m$ is the true model class, and
  - enlarging θ_m will increase $\widehat{\ell}_m$ even if further parameters are unnecessary.
- ☐ Akaike proposed minimising $2\mathrm{E}_g\mathrm{E}_g^+\left[-\log\{f(Y^+; \widehat{\theta})/g(Y^+)\}\right]$, where $Y^+, Y \overset{\text{iid}}{\sim} g$ are independent datasets. The idea is that if $\widehat{\theta} = \widehat{\theta}(Y)$ is estimated separately from Y^+, there will be a penalty due to 'missing θ_g' which will grow with $\dim(\theta)$ (picture ...)
- ☐ This leads to choosing m to minimise the **Akaike** or the **network** information criteria

$$\mathrm{AIC}_m = 2\left(d_m - \widehat{\ell}_m\right), \quad \mathrm{NIC}_m = 2\left\{\mathrm{tr}(\widehat{K}_m\widehat{J}_m^{-1}) - \widehat{\ell}_m\right\},$$

  where the first takes $\mathrm{tr}(\widehat{K}_m\widehat{J}_m^{-1}) \approx d_m = \dim(\theta_m)$.

Note: Derivation of AIC/NIC

☐ Now

$$2\mathrm{E}_g\mathrm{E}_g^+\left[-\log\{f(Y^+;\widehat{\theta})/g(Y^+)\}\right] = 2\mathrm{E}_g^+\left\{\log g(Y^+)\right\} - 2\mathrm{E}_g\mathrm{E}_g^+\left\{\log f(Y^+;\widehat{\theta})\right\},$$

so we can ignore the first term in the minimisation over f. An unbiased estimator of the second term would be $2\ell^+(\widehat{\theta})$, where ℓ^+ is the log likelihood based on Y^+ and $\widehat{\theta}$ is based on Y, but the estimator we have available is $2\ell(\widehat{\theta})$, in which the log likelihood and $\widehat{\theta}$ are both based on Y. Clearly $\ell(\widehat{\theta})$ is upwardly biased, but by how much?

☐ To find out we consider the expectation over Y^+ and Y of

$$2\left\{\ell(\widehat{\theta}) - \ell^+(\widehat{\theta})\right\} = 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} + 2\{\ell(\theta_g) - \ell^+(\theta_g)\} + 2\{\ \ell^+(\theta_g) - \ell^+(\widehat{\theta})\}, \tag{5}$$

where as before θ_g is the best candidate parameter value under f.

☐ As $\widehat{\theta}$ maximises the log likelihood, $\ell_\theta(\widehat{\theta}) = 0$, so the first term on the right-hand side of (5) is

$$\begin{aligned} 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} &\doteq 2\left\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}) - \ell_\theta(\widehat{\theta})(\theta_g - \widehat{\theta}) - \tfrac{1}{2}(\theta_g - \widehat{\theta})^{\mathrm{T}}\ell_{\theta\theta}(\widehat{\theta})(\theta_g - \widehat{\theta})\right\} \\ &\doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)(\widehat{\theta} - \theta_g), \end{aligned}$$

where we have neglected terms that are $o_p(1)$. The expectation of this scalar equals that of its trace, and the large-sample normal distribution of $\widehat{\theta}$ gives

$$\begin{aligned} \mathrm{E}_g\left[\mathrm{tr}\left\{(\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)(\widehat{\theta} - \theta_g)\right\}\right] &= \mathrm{E}_g\left[\mathrm{tr}\left\{(\widehat{\theta} - \theta_g)(\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)\right\}\right] \\ &\doteq \mathrm{tr}\left\{\imath_{\theta\theta}^{-1}(\theta_g)K(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)\imath_{\theta\theta}(\theta_g)\right\} \\ &= \mathrm{tr}\left\{K(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)\right\}. \end{aligned}$$

☐ The second term on the right-hand side of (5) has expectation zero.

☐ The third term on the right-hand side of (5) can be written as

$$2\{\ell^+(\theta_g) - \ell^+(\widehat{\theta})\} \doteq 2\left\{\ell^+(\theta_g) - \ell^+(\theta_g) - \ell_\theta^+(\theta_g)(\widehat{\theta} - \theta_g) - \tfrac{1}{2}(\widehat{\theta} - \theta_g)^{\mathrm{T}}\ell_{\theta\theta}^+(\theta_g)(\widehat{\theta} - \theta_g)\right\},$$

plus $o_p(1)$ terms. Now $\mathrm{E}_g^+\left\{\ell_\theta^+(\theta_g)\right\} = 0$ and $\mathrm{E}_g^+\left\{\ell_{\theta\theta}^+(\theta_g)\right\} = -\imath_{\theta\theta}(\theta_g)$, so

$$2\mathrm{E}_g\mathrm{E}_g^+\left\{\ell^+(\theta_g) - \ell^+(\widehat{\theta})\right\} \doteq \mathrm{E}_g\left\{(\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)(\widehat{\theta} - \theta_g)\right\} \doteq \mathrm{tr}\left\{K(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)\right\}.$$

☐ Hence

$$2\mathrm{E}_g\mathrm{E}_g^+\left[-\log f(Y^+;\widehat{\theta})\right] \doteq 2\mathrm{E}_g\mathrm{E}_g^+\left[-\log f(Y;\widehat{\theta})\right] + 2\mathrm{tr}\left\{K(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)\right\}.$$

If $K(\theta_g) \doteq \imath_{\theta\theta}(\theta_g)$, then this final expression can be estimated by $\mathrm{AIC} = 2\{d - \ell(\widehat{\theta})\}$, where $d = \dim(\theta)$, or by the *network information criterion* $\mathrm{NIC} = 2\{\mathrm{tr}(\widehat{K}\widehat{\jmath}^{-1}) - \ell(\widehat{\theta})\}$, though neither gives consistent estimation of the true model, which would require the penalty to grow with n. The calculations above rely on generic large-sample likelihood results, and could be improved in specific cases (e.g., with normal errors).

Dealing with nuisance parameters

- Profiling removes nuisance parameters, but the bias of $\widehat{\psi}$ is $O(d^3/n)$ in general, and then we require $d = o(n^{1/3})$ for consistency. Hence accuracy may be low if $\dim(\lambda)$ is high.
- Other approaches to dealing with λ include:
  - basing inference on a **marginal likelihood** or a **conditional likelihood**,

$$f(y;\psi,\lambda) = f(w;\psi) \times f(y \mid w;\psi,\lambda) = f(y \mid w_\psi;\psi) \times f(w_\psi;\psi,\lambda),$$

    where w_ψ may not depend on ψ (recall Lemmas 16 and 17) — OK for any configuration of λs, but may lose information on ψ;
  - constructing a **partial likelihood** (like the above, but harder to build);
  - **higher-order inference** such as using a **modified profile likelihood**, which can approximate both conditional and marginal likelihoods;
  - taking $\lambda \sim h(\cdot)$ and using the **integrated likelihood** $\int f(y;\psi,\lambda)h(\lambda)\,\mathrm{d}\lambda$ — depends on h, like Bayesian inference;
  - using **orthogonal parameters**, i.e., mapping $\lambda \mapsto \zeta(\lambda,\psi)$ which is orthogonal to ψ; or
  - using a **composite likelihood** in which λ does not appear.
- Below we sketch some of these.

Modified profile likelihood

- Replace profile likelihood $\exp\{\ell_{\mathrm{p}}(\psi)\}$ by the **modified profile likelihood**

$$L_{\mathrm{mp}}(\psi) = \exp\{\ell_{\mathrm{mp}}(\psi)\} = M(\psi)L_{\mathrm{p}}(\psi),$$

  with $M(\psi)$ chosen to mimic properties of marginal or conditional likelihood.
- Taking

$$M(\psi) = \left|j_{\lambda\lambda}(\psi,\widehat{\lambda}_\psi)\right|^{-1/2} \left|\frac{\partial\widehat{\lambda}}{\partial\widehat{\lambda}_\psi^{\mathrm{T}}}\right|$$

  does this in some generality.
- The
  - first term can be obtained numerically if need be, but
  - the second term is hard to compute in general.
- Simpler to base a likelihood on the normal distribution of the modified likelihood root $r^*(\psi)$ (next).

Higher-order inference . . .

- Classical theory gives first-order accuracy. With ψ scalar using the likelihood root gives

$$\mathrm{P}\left\{r(\psi_g) \leq r^{\mathrm{o}}(\psi_g)\right\} = \Phi\{r^{\mathrm{o}}(\psi)\} + O(n^{-1/2}),$$

  so tests and confidence sets based on data y^{o} have error $n^{-1/2}$.
- If we use the **modified likelihood root**,

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\left\{\frac{q(\psi)}{r(\psi)}\right\},$$

  where $q(\psi)$ depends on the model, then the error drops to $O(n^{-3/2})$ for continuous responses and to $O(n^{-1})$ for discrete responses, so

$$\mathrm{P}\left\{r^*(\psi_g) \leq r^{*\mathrm{o}}(\psi_g)\right\} = \Phi\{r^{*\mathrm{o}}(\psi_g)\} + O(n^{-3/2}),$$

  for continuous data (often almost exact even for tiny n; see Example 45).
- Highly accurate even into the distribution tails, because the relative error is bounded.
- A $1 - 2\alpha$ confidence interval,

$$\{\psi : z_\alpha \leq r^{*\mathrm{o}}(\psi) \leq 1 - \alpha\},$$

  has error of order $n^{-3/2}$ (often effectively perfect).

. . . with nuisance parameters

- With nuisance parameters, $r(\psi) = \mathrm{sign}(\widehat{\psi} - \psi)\sqrt{w_{\mathrm{p}}(\psi)}$, and

$$q(\psi) = \frac{|\, \varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \;\; \varphi_\lambda(\widehat{\theta}_\psi) \,|}{|\, \varphi_\theta(\widehat{\theta}) \,|} \left\{\frac{|\, \widehat{\jmath} \,|}{|\, \jmath_{\lambda\lambda}(\widehat{\theta}_\psi) \,|}\right\}^{1/2}$$

  where φ is the $d \times 1$ canonical parameter of a local exponential family approximation to the model at the observed data y^{o}, with $\varphi_\theta(\theta) = \partial\varphi(\theta)/\partial\theta^{\mathrm{T}}$, etc.
- In a general exponential family $\varphi(\theta)$ is the canonical parameter, and in a linear exponential family,

$$q(\psi) = (\widehat{\psi} - \psi)\left\{\frac{|\widehat{\jmath}|}{|\jmath_{\lambda\lambda}(\widehat{\theta}_\psi)|}\right\}^{1/2}.$$

- In general for independent continuous observations we write

$$\varphi(\theta)_{d\times 1} = V^{\mathrm{T}}_{d\times n} \left.\frac{\partial \ell(\theta; y)}{\partial y}\right|_{y=y^{\mathrm{o}}} = \sum_{j=1}^{n} V_j^{\mathrm{T}} \left.\frac{\partial \log f(y_j; \psi, \lambda)}{\partial y_j}\right|_{y=y^{\mathrm{o}}},$$

  where the $1 \times d$ $V_j = \partial y_j/\partial\theta^{\mathrm{T}}$ are evaluated at y^{o} and $\widehat{\theta}^{\mathrm{o}}$.

Properties of higher order approximations

- ☐ Invariant to interest-respecting reparameterization.
- ☐ Computation almost as easy as first order versions.
- ☐ Error $O(n^{-3/2})$ in continuous response models, $O(n^{-1})$ in discrete response models.
- ☐ Relative (not absolute) error, so highly accurate in tails.
- ☐ Bayesian version is also available (and easier to derive).

Example 47 (Location-scale model) *Compute $\varphi(\theta)$ for a location-scale model, in which independent observations Y_j have density $\tau^{-1}h\{(y-\eta)/\tau\}$. What about the normal density?*

Note to Example 47

- ☐ In this case the overall log likelihood is

$$\ell(\eta,\tau) = -n\log\tau + \sum_{j=1}^{n} \log h\{(y_j-\eta)/\tau\},$$

  so the vector $\partial\ell(\eta,\tau)/\partial y$ has components $\tau^{-1}(\log h)'\{(y_j-\eta)/\tau\}$, evaluated at the maximum likelihood estimates $\widehat{\eta}^{\mathrm{o}}$ and $\widehat{\tau}^{\mathrm{o}}$ and observed data vector $y_1^{\mathrm{o}},\ldots,y_n^{\mathrm{o}}$.
- ☐ To compute the V_j we use the structural expression $y = \eta + \tau\varepsilon$, where $\varepsilon \sim h$. This represents y as a function of $\theta^{\mathrm{T}} = (\eta,\tau)$, and yields $\partial y_j/\partial\theta^{\mathrm{T}} = (1,\varepsilon_j)$. This has to be evaluated at the observed data point y^{o}, and at that point the parameters are replaced by their maximum likelihood estimates, giving $V_j^{\mathrm{T}} = (1,(y_j^{\mathrm{o}}-\widehat{\eta}^{\mathrm{o}})/\widehat{\tau}^{\mathrm{o}})$.
- ☐ This yields

$$\varphi(\theta) = \sum_{j=1}^{n} \tau^{-1}(\log h)'\{(y_j^{\mathrm{o}}-\eta)/\tau\}(1,e_j)^{\mathrm{T}},$$

  where we have set $e_j = (y_j^{\mathrm{o}}-\widehat{\eta}^{\mathrm{o}})/\widehat{\tau}^{\mathrm{o}}$.
- ☐ If h is normal, then $\log h(u) \equiv -u^2/2$, so $(\log h)'\{(y_j^{\mathrm{o}}-\eta)/\tau\} = -(y_j^{\mathrm{o}}-\eta)/\tau^2$, leading to

$$\varphi(\theta)^{\mathrm{T}} = \left(\sum_{j=1}^{n}(\eta-y_j^{\mathrm{o}})/\tau^2, \sum_{j=1}^{n}(\eta-y_j^{\mathrm{o}})/\tau^2 \times e_j\right) \equiv (\eta/\tau^2, 1/\tau^2),$$

  because it turns out that inferences are invariant under non-singular affine transformations of $\varphi(\theta)$ (exercise).

Orthogonal parameters

- If the expected information matrix is block diagonal, with $\imath_{\psi,\lambda}(\theta) = 0$ for all θ, then $\widehat{\psi}$ is asymptotically independent of $\widehat{\lambda}$, and we can hope that the effect on $\widehat{\psi}$ of estimating λ will be limited. If so, we say that ψ and λ are **orthogonal**.
- This suggests mapping (ψ, γ) to (ψ, λ), where $\lambda = \lambda(\psi, \gamma)$ is orthogonal to ψ.
- Writing $\gamma = \gamma(\psi, \lambda)$ gives

$$\ell(\psi, \lambda) = \ell^* \left\{\psi, \gamma(\psi, \lambda)\right\},$$

and differentiation with respect to ψ and λ leads to

$$\frac{\partial^2 \ell}{\partial \lambda \partial \psi} = \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} \frac{\partial^2 \ell^*}{\partial \gamma \partial \psi} + \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} \frac{\partial^2 \ell^*}{\partial \gamma \partial \gamma^{\mathrm{T}}} \frac{\partial \gamma}{\partial \psi} + \frac{\partial^2 \gamma^{\mathrm{T}}}{\partial \lambda \partial \psi} \frac{\partial \ell^*}{\partial \gamma}.$$

- For orthogonality this must have expectation zero, so

$$0 = \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} i^*_{\gamma\psi} + \frac{\partial \gamma^{\mathrm{T}}}{\partial \lambda} i^*_{\gamma\gamma} \frac{\partial \gamma}{\partial \psi},$$

where $i^*_{\gamma\psi}$ and $i^*_{\gamma\gamma}$ are components of the expected information matrix in the non-orthogonal parametrization, so λ solves the system of q PDEs

$$\frac{\partial \gamma}{\partial \psi} = -i^{*-1}_{\gamma\gamma}(\psi, \gamma) i^*_{\gamma\psi}(\psi, \gamma).$$

Orthogonal parameters II

- A (possibly numerical) solution always exists when $\dim(\psi) = 1$, but need not exist when ψ is vector, because then we must simultaneously solve

$$\frac{\partial \gamma}{\partial \psi_1} = -i^{*-1}_{\gamma\gamma}(\psi, \gamma) i^*_{\gamma\psi_1}(\psi, \gamma), \quad \frac{\partial \gamma}{\partial \psi_2} = -i^{*-1}_{\gamma\gamma}(\psi, \gamma) i^*_{\gamma\psi_2}(\psi, \gamma),$$

for all γ, ψ_1 and ψ_2, but the compatibility condition

$$\frac{\partial^2 \gamma}{\partial \psi_1 \partial \psi_2} = \frac{\partial^2 \gamma}{\partial \psi_2 \partial \psi_1}$$

may fail.

Example 48 (Linear exponential family) *What parameter is orthogonal to ψ in the linear exponential family with log likelihood*

$$\ell^*(\psi, \gamma) \equiv s_1^{\mathrm{T}} \psi + s_2^{\mathrm{T}} \gamma - k(\psi, \gamma)?$$

Consider normal and Poisson likelihoods in particular.

Note to Example 48

☐ The parameters $\lambda = \lambda(\psi, \gamma)$ orthogonal to ψ are determined by

$$\frac{\partial \gamma}{\partial \psi^{\mathrm{T}}} = -k_{\gamma\gamma}^{-1}(\psi, \gamma) k_{\gamma\psi}(\psi, \gamma). \tag{6}$$

If we reparametrize in terms of ψ and $\lambda = k_\gamma(\psi, \gamma) = \partial k(\psi, \gamma)/\partial\gamma$, then in this new parametrization, γ is a function of ψ and λ, and

$$0 = \frac{\partial \lambda^{\mathrm{T}}}{\partial \psi} = \frac{\partial \gamma^{\mathrm{T}}}{\partial \psi} k_{\gamma\gamma}(\psi, \gamma) + k_{\psi\gamma}(\psi, \gamma),$$

so $\lambda = k_\gamma(\psi, \gamma)$ is a solution to (6). That is, the parameter orthogonal to ψ is the so-called complementary mean parameter $\lambda(\psi, \gamma) = \mathrm{E}(S_2; \psi, \gamma)$. By symmetry, $\mathrm{E}(S_1; \psi, \gamma)$ is orthogonal to γ.

☐ The normal distribution with mean μ and variance σ^2 has canonical parameter $(\mu/\sigma^2, -1/(2\sigma^2))$. The canonical statistic (Y, Y^2) has expectation $(\mu, \mu^2 + \sigma^2)$, so μ is orthogonal to $-1/(2\sigma^2)$, and hence to σ^2, while μ/σ^2 is orthogonal to $\mu^2 + \sigma^2$.

☐ Independent Poisson variables Y_1 and Y_2 with means $\exp(\gamma)$ and $\exp(\gamma + \psi)$ have log likelihood

$$\ell^*(\psi, \gamma) \equiv (y_1 + y_2)\gamma + y_2\psi - e^\gamma - e^{\gamma+\psi}.$$

The discussion above suggests that

$$\lambda = \mathrm{E}(Y_1 + Y_2) = \exp(\gamma) + \exp(\gamma + \psi) = e^\gamma(1 + e^\psi)$$

is orthogonal to ψ, so $\gamma = \log\lambda - \log(1 + e^\psi)$ and

$$\ell(\psi, \lambda) \equiv y_2\psi - (y_1 + y_2)\log(1 + e^\psi) + (y_1 + y_2)\log\lambda - \lambda.$$

The separation of ψ and λ implies that the profile and modified profile likelihoods for ψ are proportional. They correspond to the conditional likelihood obtained from the density of Y_2 given $Y_1 + Y_2$.

Composite likelihood

- Used when full likelihood can't be computed but densities for distinct subsets of the observations, $y_{\mathcal{S}_1}, \ldots, y_{\mathcal{S}_C}$, are available, can use a **composite (log) likelihood**

$$\ell_{\mathrm{C}}(\theta) = \sum_{c=1}^{C} \log f(y_{\mathcal{S}_c}; \theta).$$

- The choice of subsets $\mathcal{S}_1, \ldots, \mathcal{S}_C$ determines what parameters can be estimated.
- Special cases:
  - **independence likelihood** takes $\mathcal{S}_j = \{y_j\}$ and treats (possibly dependent) y_j as independent;
  - **pairwise likelihood** uses subsets of distinct pairs $\{y_j, y_{j'}\}$.
- May be useful with spatial data, and then contributions from distant pairs may be downweighted or dropped entirely.
- $\ell_{\mathrm{C}}(\theta)$ satisfies the first Bartlett identity, so can give consistent estimators $\tilde{\theta}$, but requires a sandwich variance matrix (or some other approach) to estimate $\mathrm{var}(\tilde{\theta})$.
- Model comparisons use the **composite likelihood information criterion**

$$\mathrm{CLIC} = 2\left[\mathrm{tr}\{K(\tilde{\theta})\jmath(\tilde{\theta})^{-1}\} - \ell_{\mathrm{C}}(\tilde{\theta})\right].$$

Empirical likelihood

- Empirical likelihood allows nonparametric estimation of constrained distributions.
- If $y_1, \ldots, y_n \overset{\text{iid}}{\sim} G$, then the $\widehat{G}$ that maximises the 'nonparametric likelihood'

$$L(G) = \prod_{j=1}^{n} G(\mathrm{d}y_j) = \prod_{j=1}^{n} p_j, \quad \text{subject to} \quad p_j \geq 0, \quad \sum_{j=1}^{n} p_j \leq 1,$$

  sets $G(\mathrm{d}y_j) = \widehat{p}_j \equiv n^{-1}$: $\widehat{G}$ is the empirical distribution function of $y_1, \ldots, y_n$.
- Adding the constraint $\mathrm{E}\{c(Y;\theta)\} = 0$ leads to maximising

$$\sum_{j=1}^{n} \log p_j \quad \text{subject to} \quad p_j \geq 0, \quad \sum_{j=1}^{n} p_j \leq 1, \quad \sum_{j=1}^{n} p_j c(y_j;\theta) = 0,$$

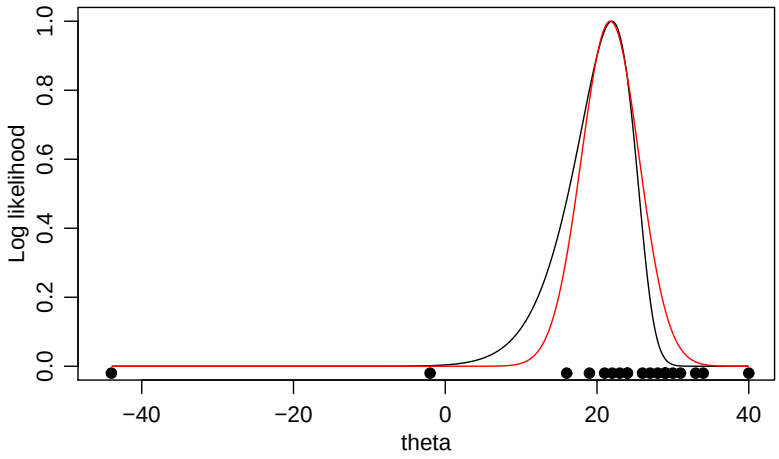
  and a use of Lagrange multipliers shows that we must find $a = a_\theta$ to solve

$$\sum_{j=1}^{n} \frac{c(y_j;\theta)}{n\{1 + ac(y_j;\theta)\}} = 0$$

  giving $\widehat{a} = \widehat{a}_\theta$, $\widehat{p}_j(\theta) = n^{-1}/\{1 + \widehat{a}c(y_j;\theta)\}$ and empirical likelihood ratio statistic $w(\theta) = 2\sum \log\{1 + \widehat{a}c(y_j;\theta)\}$
- The usual χ^2 result applies to $w(\theta)$. Can be widely generalised ...

Example: Newcomb data

$n = 20$ observations on the speed of light (made by Simon Newcomb), empirical likelihood (black) and normal likelihood (red) for the mean θ. Note how the empirical likelihood adapts to the 'shape' of the data.



Non-regular models

- The regularity conditions (C1)–(C4) apply in many settings met in practice, but not universally. The most common failures arise when
  - some of the parameters are discrete (e..g, change point problems),
  - the model is not identifiable (distinct θ values give the same model),
  - θ_g is on the boundary of the parameter space (e.g., testing for a zero variance),
  - $d = \dim(\theta)$ grows (too fast) with n, or
  - the support of $f(y; \theta)$ depends on θ (so the Bartlettt identities fail).
- Even when the conditions are satisfied there can be datasets for which maximum likelihood estimation fails, e.g.,
  - there is no unique maximum to the likelihood, or
  - the maximum is on the edge of the parameter space,

  and then penalisation (equivalent to using a prior) is often used.

Example 49 *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0, \theta)$*, find the likelihood and* $\widehat{\theta}$*, and show that the limit distribution of* $n(\theta - \widehat{\theta})/\theta$ *when* $n \to \infty$ *is* $\exp(1)$*. Discuss.*

Note to Example 49

Owing to the independence,

$$L(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta) = \prod_{j=1}^{n} \left\{\theta^{-1} I(0 < y_j < \theta)\right\} = \theta^{-n} I(\max y_j < \theta), \quad \theta > 0,$$

and therefore $\widehat{\theta} = M = \max Y_j$, whose distribution is

$$\mathrm{P}(M \leq x) = (x/\theta)^n, \quad 0 < x < \theta.$$

Now

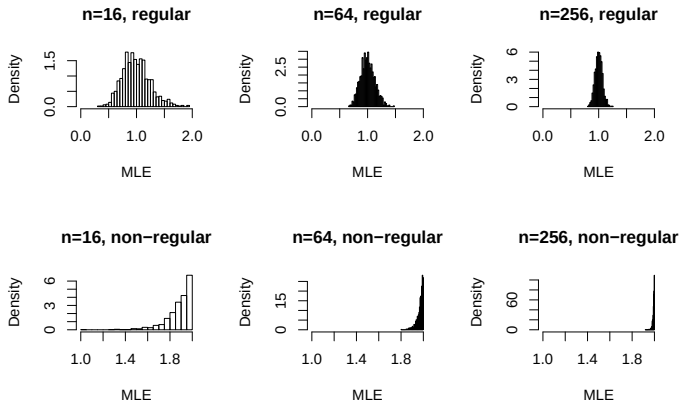
$$\mathrm{P}\left\{n(\theta - \widehat{\theta})/\theta \leq x\right\} = \mathrm{P}(\widehat{\theta} \geq \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n \to 1 - \exp(-x),$$

as required. Note that:

- ☐ the scaling needed to get a limiting distribution is much faster here than in the regular case (we have to multiply by n to get a non-degenerate limit);
- ☐ the limit is not normal.

Uniform example

Comparison of the distributions of $\widehat{\theta}$ in a regular case (panels above, with standard deviation $\propto n^{-1/2}$) and in a nonregular case (Example 49, panels below, with standard deviation $\propto n^{-1}$). In other nonregular cases it might happen that the distribution is nasty (unlike here) and/or that the convergence is slower than in regular cases.



Comments

- Other likelihoods and/or likelihood-like functions are widely used, especially
  - **partial likelihood**, used to eliminate nuisance functions for inference (survival data),
  - **quasi-likelihood**, used to model over-dispersion in exponential family models,
  - **pseudo-likelihood**, treats data as Gaussian even when they are not (econometrics).
- Strengths of likelihood approach:
  - heuristic as plausibility of a model as explanation of data;
  - we 'just' have to write down the density of the observed data;
  - invariance to data and parameter transformations;
  - simple and (fairly) general approximate theory for inference under regularity conditions, also easily implemented numerically;
  - large-sample optimality properties in regular cases;
  - close links to Bayesian inference (next).
- Weaknesses of likelihood approach:
  - requires 'parametric' model for data;
  - can have trouble in high-dimensional settings;
  - not all models are regular.

Parameters and functionals

- ☐ Parametric models are determined by a finite vector $\theta \in \Theta$. Does this generalise?
- ☐ If $Y \sim G$, then we can define a parameter in terms of a **statistical functional**, e.g.,

$$\mu = t_1(G) = \int y \, \mathrm{d}G(y), \quad \sigma^2 = t_2(G) = \int y^2 \, \mathrm{d}G(y) - \left\{ \int y \, \mathrm{d}G(y) \right\}^2.$$

- ☐ Below we always assume that such functionals are well-defined.
- ☐ We apply the '**plug-in principle**' and replace G by an estimator $\widehat{G}$, giving

$$\widehat{\mu} = t_1(\widehat{G}) = \int y \, \mathrm{d}\widehat{G}(y), \quad \widehat{\sigma}^2 = t_2(\widehat{G}) = \int y^2 \, \mathrm{d}\widehat{G}(y) - \left\{ \int y \, \mathrm{d}\widehat{G}(y) \right\}^2.$$

- ☐ With a parametric model we can write $G \equiv G_\theta$ and $\widehat{G} \equiv G_{\widehat{\theta}}$, but a general estimator of G based on $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$ is the **empirical distribution function (EDF)**

$$\widehat{G}(y) = \frac{1}{n} \sum_{j=1}^{n} H(y - Y_j), \quad H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}$$

  where $H(\cdot)$ is the **Heaviside function**.

Algorithmic approach

Example 50 *Give general definitions of the median and the parameter obtained from a maximum likelihood fit of a density $f(y;\theta)$. What are the corresponding estimators (a) under a fitted exponential model, and (b) a nonparametric model?*

- ☐ This approach is essentially algorithmic: $t(\cdot)$ is an algorithm that
  - when applied to the distribution G gives the parameter $t(G)$;
  - when applied to an estimator $\widehat{G}$ based on data $Y_1, \ldots, Y_n$ gives the estimator $t(\widehat{G})$.
- ☐ The algorithm $t(\cdot)$ can be (almost) arbitrarily complex.
- ☐ This point of view suggests a sampling approach to frequentist inference:
  - if we knew G, we could assess the properties of $t(\widehat{G})$ by generating many samples $\widehat{G} \equiv \{Y_1, \ldots, Y_n\}$ from G and looking at the corresponding values of $t(\widehat{G})$;
  - since G is unknown, we replace it by $\widehat{G}$, generate samples $\widehat{G}^* \equiv \{Y_1^*, \ldots, Y_n^*\}$ from $\widehat{G}$, and use the corresponding values of $t(\widehat{G}^*)$ to estimate the distribution of $t(\widehat{G})$.
- ☐ The samples $\widehat{G}^* \equiv \{Y_1^*, \ldots, Y_n^*\}$ are known as **bootstrap samples**, and the overall procedure is known as a **bootstrap**, one of many possible **resampling** procedures.

Example 50

- The usual definition of the p quantile is

$$t_1(G) = \inf\{x : G(x) \geq p\},$$

for $p \in (0, 1)$. For the median we set $p = 1/2$.

- The maximum likelihood estimator is defined as

$$t_2(G) = \arg\max_\theta \mathrm{E}_G\{\log f(Y;\theta)\} = \arg\max_\theta \int \log f(y;\theta)\,\mathrm{d}\widehat{G}(y),$$

which we earlier called θ_g.

- Under an exponential model

$$t_1(G) = \inf\{x : 1 - \exp(-\lambda x) \geq p\} = -\lambda^1 \log(1-p) = \lambda^{-1}\log 2,$$

so if the fitted model has parameter $\widehat{\lambda}$, then $t_1(\widehat{G}) = \widehat{\lambda}^{-1}\log 2$.
Likewise θ_g is estimated by

$$\arg\max_\theta \int \log f(y;\theta)\,\widehat{\lambda}e^{-\widehat{\lambda}y}\,\mathrm{d}y;$$

note that f is not necessarily exponential.

- Under the general model and with order statistics $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$,

$$t_1(\widehat{G}) = \inf\{x : \widehat{G}(x) \geq p\} = Y_{(m)},$$

where $m = \lfloor (n+1)/2 \rfloor$, and as $\mathrm{d}H(u)$ puts a unit mass at $u = 0$,

$$\begin{aligned}
t_2(\widehat{G}) &= \arg\max_\theta \int \log f(y;\theta)\,\mathrm{d}\widehat{G}(y) \\
&= \arg\max_\theta \int \log f(y;\theta)\,\mathrm{d}\left\{ n^{-1}\sum_{j=1}^n H(y - Y_j)\right\} \\
&= \arg\max_\theta n^{-1}\sum_{j=1}^n \int \log f(y;\theta)\,\mathrm{d}H(y - Y_j) \\
&= \arg\max_\theta n^{-1}\sum_{j=1}^n \log f(Y_j;\theta),
\end{aligned}$$

i.e., the maximum likelihood estimator of θ based on the sample.

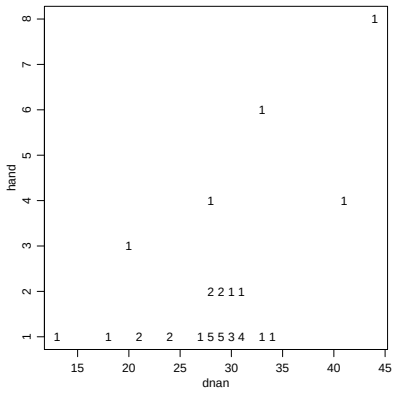
Example: Handedness data

Table 1: Data from a study of handedness; `hand` is an integer measure of handedness, and `dnan` a genetic measure. Data due to Dr Gordon Claridge, University of Oxford.

	dnan	hand		dnan	hand		dnan	hand		dnan	hand
1	13	1	11	28	1	21	29	2	31	31	1
2	18	1	12	28	2	22	29	1	32	31	2
3	20	3	13	28	1	23	29	1	33	33	6
4	21	1	14	28	4	24	30	1	34	33	1
5	21	1	15	28	1	25	30	1	35	34	1
6	24	1	16	28	1	26	30	2	36	41	4
7	24	1	17	29	1	27	30	1	37	44	8
8	27	1	18	29	1	28	31	1			
9	28	1	19	29	1	29	31	1			
10	28	2	20	29	2	30	31	1			

Example: Handedness data

Scatter plot of handedness data. The numbers show the multiplicities of the observations.



Example: Handedness data

- How do we quantify dependence between `dnan` and `hand` for these $n = 37$ individuals?
- A standard measure is the **product-moment (Pearson) correlation** for $G(u, v)$, i.e.,

$$\theta = t(G) = \frac{\int \{u - \int u \, \mathrm{d}G(u,v)\} \{v - \int v \, \mathrm{d}G(u,v)\} \, \mathrm{d}G(u,v)}{\left[\int \{u - \int u \, \mathrm{d}G(u,v)\}^2 \, \mathrm{d}G(u,v) \int \{v - \int v \, \mathrm{d}G(u,v)\}^2 \, \mathrm{d}G(u,v)\right]^{1/2}}.$$

- With $(u, v) = (\texttt{dnan}, \texttt{hand})$, the sample version is

$$\begin{aligned} \widehat{\theta} = t(\widehat{G}) &= \frac{\sum_{j=1}^n (\texttt{dnan}_j - \overline{\texttt{dnan}})(\texttt{hand}_j - \overline{\texttt{hand}})}{\left\{\sum_{j=1}^n (\texttt{dnan}_j - \overline{\texttt{dnan}})^2 \sum_{j=1}^n (\texttt{hand}_j - \overline{\texttt{hand}})^2\right\}^{1/2}} \\ &= 0.509. \end{aligned}$$

- Standard (bivariate normal) 95% confidence interval is $(0.221, \ 0.715)$, but this is obviously inappropriate (the data look highly non-normal).
- Try simulation approach ...

Bootstrap simulation

- Whether $\widehat{G}$ is parametric or non-parametric, we simulate as follows:
  - For $r = 1, \ldots, R$:
    - ▷ generate a bootstrap sample $y_1^*, \ldots, y_n^* \overset{\text{iid}}{\sim} \widehat{G}$,
    - ▷ compute $\widehat{\theta}_r^*$ using $y_1^*, \ldots, y_n^*$,

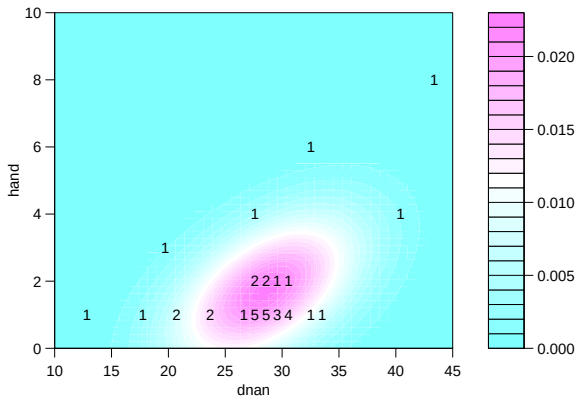
    so the output is a set of **bootstrap replicates**,

$$\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*.$$

- We then use $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ to estimate properties of $\widehat{\theta}$ (histogram, ...).
- If $R \to \infty$, then get perfect match to theoretical calculation based on $\widehat{G}$ (if this is available): Monte Carlo error disappears completely.
- In practice R is finite, so some Monte Carlo error remains.
- If $\widehat{G}$ is the EDF, then $y_1^*, \ldots, y_n^* \overset{\text{iid}}{\sim} \widehat{G}$ are sampled with replacement and equal probabilities from $y_1, \ldots, y_n$, so if $f_i^* = \#\{y_j^* = y_i\}$, then $(f_1^*, \ldots, f_n^*)$ has the multinomial distribution with denominator n and probability vector $(n^{-1}, \ldots, n^{-1})$.
- Although $\mathrm{E}^*(f_j^*) = 1$, y_j can appear $0, 1, \ldots, n$ times in the bootstrap sample.

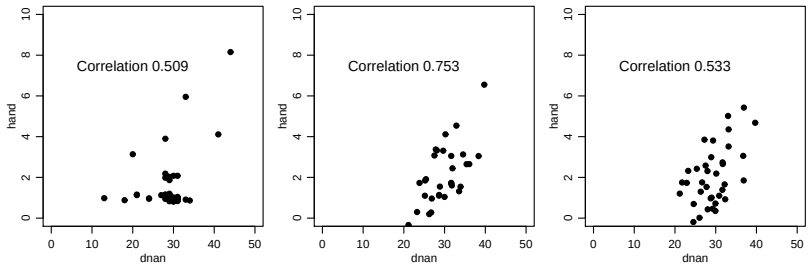
Handedness data: Fitted bivariate normal model

Contours of bivariate normal distribution fitted to handedness data; parameter estimates are $\widehat{\mu}_1 = 28.5$, $\widehat{\mu}_2 = 1.7$, $\widehat{\sigma}_1 = 5.4$, $\widehat{\sigma}_2 = 1.5$, $\widehat{\rho} = 0.509$. The data are also shown.



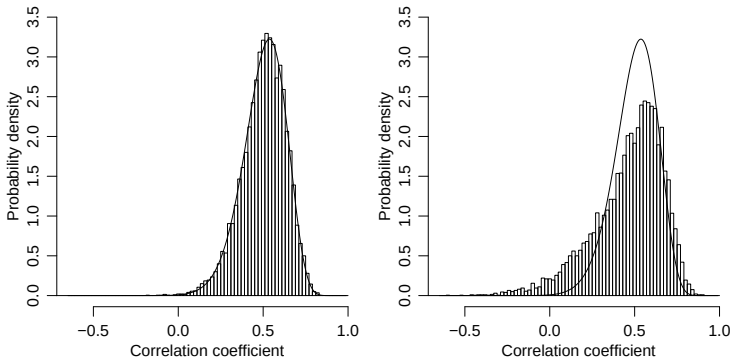
Handedness data: Parametric bootstrap samples

Left: original data, with jittered vertical values. Centre and right: two samples generated from the fitted bivariate normal distribution.



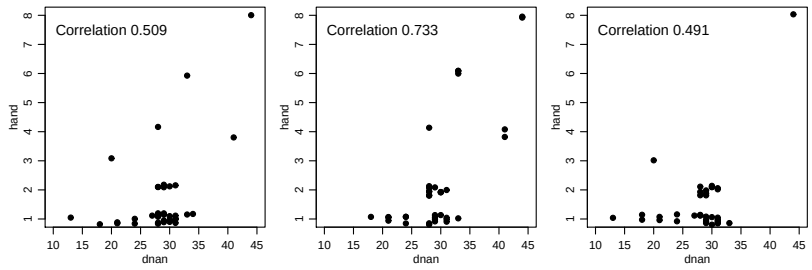
Handedness data: Correlation coefficient

Bootstrap distributions with $R = 10000$. Left: simulation from fitted bivariate normal distribution. Right: nonparametric sampling from the EDF. The lines show the theoretical probability density function of the correlation coefficient under sampling from a fitted bivariate normal distribution.



Handedness data: Bootstrap samples

Left: original data, with jittered vertical values. Centre and right: two bootstrap samples, with jittered vertical values.



Using the $\widehat{\theta}^*$

- The **bias** and **variance** of $\widehat{\theta}$ as an estimator of $\theta = t(G)$,

$$\beta(G) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} G) - t(G), \quad \nu(G) = \mathrm{var}(\widehat{\theta} \mid G),$$

  are estimated by replacing the unknown G by its known estimate $\widehat{G}$:

$$\beta(\widehat{G}) = \mathrm{E}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} \widehat{G}) - t(\widehat{G}), \quad \nu(\widehat{G}) = \mathrm{var}(\widehat{\theta} \mid y_1, \ldots, y_n \overset{\text{iid}}{\sim} \widehat{G}).$$

- The Monte Carlo approximations to $\beta(\widehat{G})$ and $\nu(\widehat{G})$ are

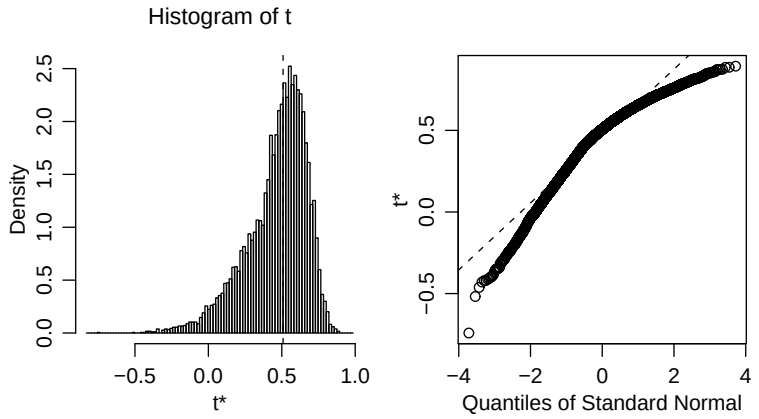
$$b = \overline{\widehat{\theta}^*} - \widehat{\theta} = R^{-1} \sum_{r=1}^{R} \widehat{\theta}^*_r - \widehat{\theta}, \quad v = \frac{1}{R-1} \sum_{r=1}^{R} \left(\widehat{\theta}^*_r - \overline{\widehat{\theta}^*} \right)^2.$$

  For the handedness data, $R = 10^4$ and $b = -0.046$, $v = 0.043 = 0.205^2$.

- We estimate the p **quantile** of $\widehat{\theta}$ using the p quantile of $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$, i.e., $\widehat{\theta}^*_{((R+1)p)}$.

Handedness data

Summaries of the $\widehat{\theta}^*$. Left: histogram, with vertical line showing $\widehat{\theta}$. Right: normal Q–Q plot of $\widehat{\theta}^*$.



Common questions

- **How big should n be?** — depends on the context
- **What if the sample is unrepresentative?** — this is always a potential problem in statistics, not specific to resampling methods.
- **How big should R be?** — at least 1000 for most purposes
- **Why take resamples of size n?**
  - We usually want to mimic the sampling properties of samples like the original one, so take resamples of size n,
  - but sometimes we take resamples of size $m \ll n$ in order to achieve validity of the bootstrap—e.g., for extreme quantiles.
- **Why resample from the EDF?**
  - The EDF is the nonparametric MLE of G, so is a natural choice, but
  - sometimes (e.g., testing) we resample from a constrained version of $\widehat{G}$,
  - sometimes it may be useful to smooth $\widehat{G}$;
  - sometimes it may be useful to simulate from (several) parametric fits.

How big should n be?

- For the **average** $\widehat{\theta} = \overline{y}$, the number of distinct samples is

$$m_n = \binom{2n-1}{n},$$

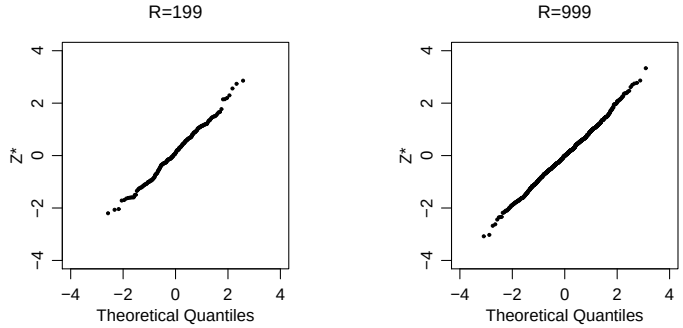
  the most probable of which has probability $p_n = n!/n^n$.
  For $n > 12$, we have $m_n > 10^6$ and $p_n < 6 \times 10^{-5}$.
- Bootstrapping of smooth statistics like the average will often work OK provided $n > 20$.
- For the **median** of a sample of size $n = 2m + 1$, the possible distinct values of $\widehat{\theta}^*$ are $y_{(1)} < \cdots < y_{(n)}$, and

$$\mathrm{P}^*(\widehat{\theta}^* > y_{(l)}) = \sum_{r=0}^{m} \binom{n}{r} \left(\frac{l}{n}\right)^r \left(1 - \frac{l}{n}\right)^{n-r},$$

  so exact calculations of the variance etc. are possible.
- However the median is very vulnerable to bad sample values, so for the median (and other 'non-smooth' statistics) much larger n is needed for reliable inference.

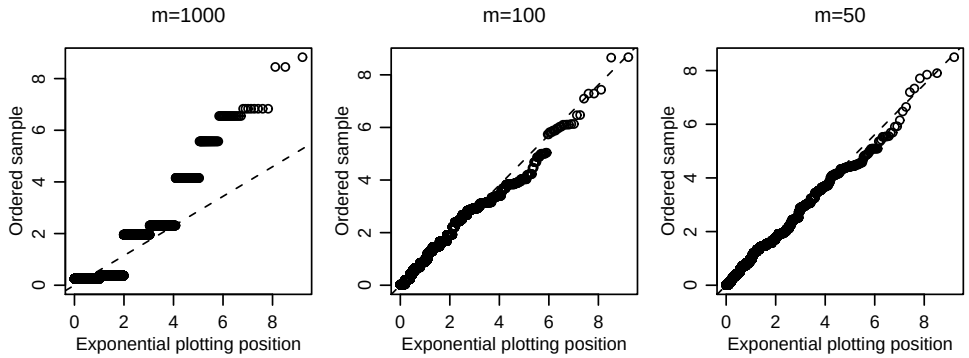
How many bootstraps?

- ☐ Must estimate moments and quantiles of $\widehat{\theta}$ and derived quantities. Often feasible to take $R \gg 1000$
- ☐ Need $R \geq 200$ to estimate bias, variance, etc.
- ☐ Need $R \gg 100$, preferably $R \geq 2500$ to estimate quantiles needed for 95% confidence intervals



Resamples of size n?

- ☐ Exponential sample of size $n = 1000$
- ☐ Distribution of $n \min(Y_1, \ldots, Y_n)$ is $\exp(1)$
- ☐ Resampling distribution $m \min(Y_1^*, \ldots, Y_m^*)$ using resamples of size $m = 1000$, 100, 50
- ☐ To avoid discreteness must choose $m \ll n$, but how?



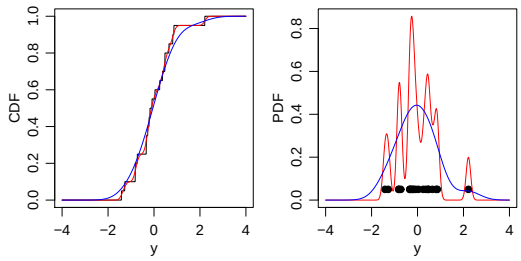
Variants of $\widehat{G}$?

- Can be useful to simulate from a smoothed EDF, given by

$$Y^* = y_{j^*} + h\varepsilon^*, \quad \varepsilon^* \sim \mathcal{N}(0,1) \perp\!\!\!\perp j^* \sim U\{1,\ldots,n\},$$

  equivalent to simulating from a kernel density estimate. Below, with $h = 0.1$ (red) and $h = 0.5$ (blue).
- Since $\mathrm{var}^*(Y^*) = \widehat{\sigma}^2 + h^2$, may prefer a shrunk smoothed estimate, given by

$$Y^* = \overline{y} + \frac{(y_{j^*} - \overline{y}) + h\varepsilon^*}{(1 + h^2/\widehat{\sigma}^2)^{1/2}}.$$



When does the bootstrap work?

- 'Work' might mean the bootstrap gives
  - **reliable** answers when used in practice, or
  - **mathematically correct** answers under 'suitable' regularity conditions.
- For the second of these, suppose we seek to estimate properties of a standardized quantity $Q = q(Y_1,\ldots,Y_n;G)$, maybe $Q = n^{1/2}(\overline{Y} - \theta)$. Let $n \to \infty$ to get limiting results for the distribution function

$$H_{G,n}(q) = \mathrm{P}_G\{Q(Y_1,\ldots,Y_n;G) \le q\},$$

  where subscript G indicates that $Y_1,\ldots,Y_n$ is a random sample from G.
- Bootstrap estimate of this is

$$H_{\widehat{G},n}(q) = \mathrm{P}_{\widehat{G}}\left\{Q(Y_1^*,\ldots,Y_n^*;\widehat{G}) \le q\right\}$$

  where $Q(Y_1^*,\ldots,Y_n^*;\widehat{G}) = n^{1/2}(\overline{Y}^* - \overline{y})$.
- We need conditions under which $H_{\widehat{G},n} \xrightarrow{D} H_{G,n}$ as $n \to \infty$.

Regularity conditions

- ☐ The true distribution G is surrounded by a neighbourhood $\mathcal{N}$ in a suitable space of distributions, and as $n \to \infty$, $\widehat{G}$ eventually falls into $\mathcal{N}$ with probability one. Also:
  1. for any $F \in \mathcal{N}$, $H_{F,n}$ converges weakly to a limit $H_{F,\infty}$;
  2. this convergence must be uniform on $\mathcal{N}$; and
  3. the function mapping F to $H_{F,\infty}$ must be continuous.
- ☐ Weak convergence of $H_{F,n}$ to $H_{F,\infty}$ means that for all integrable $b(\cdot)$,

$$\int b(u)\,\mathrm{d}H_{F,n}(u) \quad \to \quad \int b(u)\,\mathrm{d}H_{F,\infty}(u), \qquad n \to \infty.$$

- ☐ Under these conditions the bootstrap is **consistent**: for any q and $\varepsilon > 0$,

$$\mathrm{P}\{|H_{\widehat{G},n}(q) - H_{G,\infty}(q)| > \varepsilon\} \to 0, \quad n \to \infty.$$

- ☐ The first condition ensures that there is a limit for $H_{G,n}$ to converge to.
- ☐ As n increases, $\widehat{G}$ changes, so the second and third conditions are needed to ensure that $H_{\widehat{G},n}$ approaches $H_{G,\infty}$ along every possible sequence of $\widehat{G}$s.
- ☐ If any one of these conditions fails, the bootstrap can fail. For the minimum (for example) the convergence is not uniform on suitable neighbourhoods of G.

Summary

- ☐ **Estimator is algorithm**:
  - applied to original data $y_1, \ldots, y_n$ gives original $\widehat{\theta}$;
  - applied to simulated data $y_1^*, \ldots, y_n^*$ gives $\widehat{\theta}^*$;
  - $\widehat{\theta}$ can be of (almost) any complexity; but
  - for more sophisticated ideas to work, $\widehat{\theta}$ must often be smooth function of data.
- ☐ **Sample is used to estimate G**:
  - $\widehat{G} \approx G$ — heroic assumption
- ☐ **Simulation replaces theoretical calculation:**
  - removes need for mathematical skill;
  - does not remove need for thought; and in particular,
  - check code **very** carefully — garbage in, garbage out!
- ☐ **Two sources of error**:
  - statistical ($\widehat{G} \neq G$) — reduce by thought; and
  - simulation ($R \neq \infty$) — reduce by taking R large (enough).

Bootstrap confidence Intervals: Desiderata

- A $(1-\alpha)$ **upper confidence limit** for a scalar parameter θ based on data Y is a random variable $\theta_\alpha = \theta_\alpha(Y)$ for which

$$\mathrm{P}\left(\theta \leq \theta_\alpha\right) = \alpha, \quad 0 < \alpha < 1, \theta \in \Theta. \tag{7}$$

- We may seek invariance to monotone transformations $\psi = \psi(\theta)$, that is

$$\mathrm{P}\left\{\psi(\theta) \leq \psi_\alpha\right\} = \alpha, \quad 0 < \alpha < 1, \theta \in \Theta.$$

- In practice exact intervals are rarely available, and we seek intervals such that (7) is satisfied as closely as possible. If $Y \equiv Y_1, \ldots, Y_n$, then we typically have

$$\mathrm{P}\left(\theta \leq \theta_\alpha\right) = \alpha + \mathcal{O}(n^{-1/2}), \quad 0 < \alpha < 1, \theta \in \Theta,$$

and the corresponding two-sided interval satisfies

$$\mathrm{P}\left(\theta_\alpha < \theta \leq \theta_{1-\alpha}\right) = (1-2\alpha) + \mathcal{O}(n^{-1}), \quad 0 < \alpha < 1/2, \theta \in \Theta.$$

Normal confidence intervals

- If $\widehat{\theta} \dot{\sim} \mathcal{N}(\theta + \beta, \nu)$ with known bias $\beta = \beta(G)$ and variance $\nu = \nu(G)$, then a $(1-2\alpha)$ confidence interval is based on the equation

$$\mathrm{P}\left(z_\alpha < \frac{\widehat{\theta} - \theta - \beta}{\nu^{1/2}} \leq z_{1-\alpha}\right) = 1 - 2\alpha,$$

and has limits $\widehat{\theta} - \beta \pm z_\alpha \nu^{1/2}$, where $\Phi(z_\alpha) = \alpha$.

- We replace β, ν by the bootstrap estimates

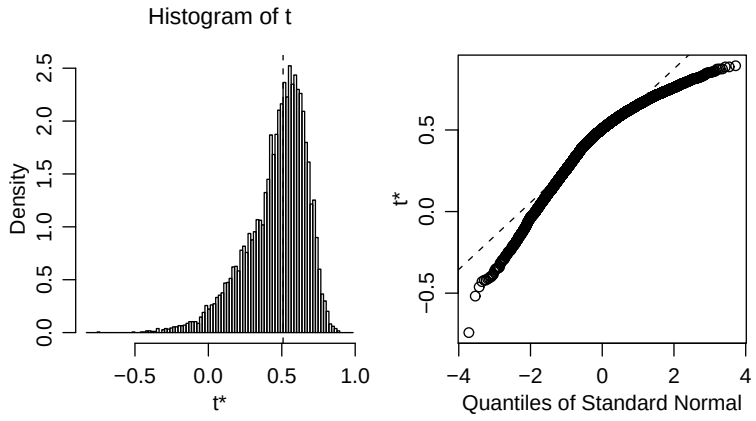
$$\begin{aligned} \beta(G) &\doteq \beta(\widehat{G}) \doteq b = \overline{\widehat{\theta}^*} - \widehat{\theta}, \\ \nu(G) &\doteq \nu(\widehat{G}) \doteq v = (R-1)^{-1} \sum_r (\widehat{\theta}_r^* - \overline{\widehat{\theta}^*})^2, \end{aligned}$$

to get the $(1-2\alpha)$ interval with limits $\widehat{\theta} - b \pm z_\alpha v^{1/2}$.

- For the handedness data we have $R = 10,000$, $b = -0.046$, $v = 0.205^2$, $\alpha = 0.025$, $z_\alpha = -1.96$, so 95% CI is $(0.147, 0.963)$
- We can use the $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ to check the quality of the normal approximation, and perhaps to suggest transformations.

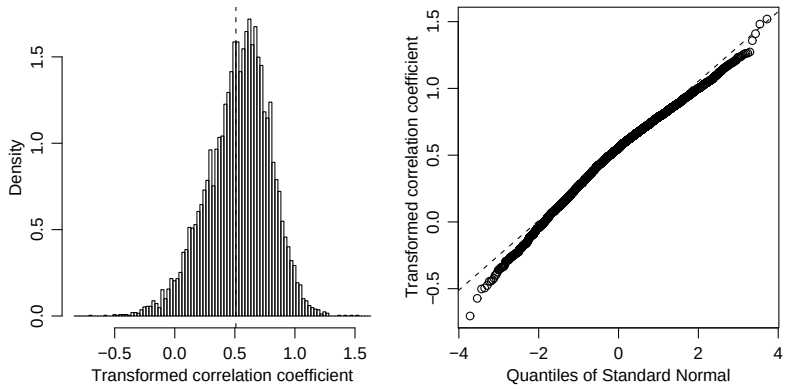
Handedness data

Summaries of the $\widehat{\theta}^*$. Left: histogram, with vertical line showing $\widehat{\theta}$. Right: normal Q–Q plot of $\widehat{\theta}^*$.



Handedness data: Transformed scale?

Plots for $\widehat{\psi}^* = \frac{1}{2}\log\{(1+\widehat{\theta}^*)/(1-\widehat{\theta}^*)\}$:



Normal confidence intervals

- Correlation coefficient: try Fisher's z transformation:

$$\widehat{\psi}^* = \psi(\widehat{\theta}^*) = \tfrac{1}{2}\log\{(1+\widehat{\theta}^*)/(1-\widehat{\theta}^*)\}$$

with bias and variance estimates

$$b_\psi = R^{-1}\sum_{r=1}^{R}\widehat{\psi}_r^* - \widehat{\psi}, \quad v_\psi = \frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{\psi}_r^* - \overline{\widehat{\psi}^*}\right)^2,$$

- Then the $(1-2\alpha)$ confidence interval for θ is

$$\psi^{-1}\left\{\widehat{\psi} - b_\psi - z_{1-\alpha}v_\psi^{1/2}\right\}, \quad \psi^{-1}\left\{\widehat{\psi} - b_\psi - z_\alpha v_\psi^{1/2}\right\}$$

- For handedness data, get $(0.074, 0.804)$ ... but how do we choose a transformation in general?

Pivots

- Assume properties of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ mimic effect of sampling from original model (plug-in principle) — false in general, but more nearly true for pivots.
- **Pivot** is combination of data and parameter whose distribution is independent of underlying model, such as t statistic

$$Z = \frac{\overline{Y} - \mu}{(S^2/n)^{1/2}} \quad \sim \quad t_{n-1},$$

when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

- Exact pivot generally unavailable in nonparametric case, but if we can estimate the variance of $\widehat{\theta}^*$ using V, we use

$$Z = \frac{\widehat{\theta} - \theta}{V^{1/2}}$$

- If the quantiles z_α of Z known, then

$$\mathrm{P}\left(z_\alpha \leq Z \leq z_{1-\alpha}\right) = \mathrm{P}\left(z_\alpha \leq \frac{\widehat{\theta} - \theta}{V^{1/2}} \leq z_{1-\alpha}\right) = 1 - 2\alpha$$

(z_α no longer denotes a normal quantile!) gives $(1-2\alpha)$ CI $(\widehat{\theta} - V^{1/2}z_{1-\alpha}, \widehat{\theta} - V^{1/2}z_\alpha)$

Studentized statistic

- Bootstrap sample gives $(\widehat{\theta}^*, V^*)$ and hence

$$Z^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{V^{*1/2}}.$$

- We bootstrap to get R copies of $(\widehat{\theta}, V)$, i.e.,

$$(\widehat{\theta}^*_1, V^*_1), \quad (\widehat{\theta}^*_2, V^*_2), \quad \ldots, \quad (\widehat{\theta}^*_R, V^*_R),$$

  and the corresponding

$$z^*_1 = \frac{\widehat{\theta}^*_1 - \widehat{\theta}}{V^{*1/2}_1}, \quad z^*_2 = \frac{\widehat{\theta}^*_2 - \widehat{\theta}}{V^{*1/2}_2}, \quad \ldots, \quad z^*_R = \frac{\widehat{\theta}^*_R - \widehat{\theta}}{V^{*1/2}_R},$$

  then order these to estimate quantiles of Z, with z_p estimated by $z^*_{(p(R+1))}$.
- Get $(1-2\alpha)$ **Studentized bootstrap confidence interval**

$$\widehat{\theta} - V^{1/2} z^*_{((1-\alpha)(R+1))}, \quad \widehat{\theta} - V^{1/2} z^*_{(\alpha(R+1))}.$$

- This is not invariant to transformation and needs an estimated variance V^*_r for each $\widehat{\theta}^*_r$.

Why Studentize?

- If we Studentize, then $Z \xrightarrow{D} N(0,1)$ as $n \to \infty$, and we can use Edgeworth series to write

$$\mathrm{P}_G(Z \le z) = \Phi(z) + n^{-1/2} a(z)\phi(z) + O(n^{-1}),$$

  where $a(\cdot)$ is an even quadratic polynomial.
- For example, if we use $\widehat{\theta} = \overline{Y}$ and $V = n^{-1}S^2$ to compute Z for data with skewness γ, then $a(x) = \gamma(2x^2+1)/6$ and (next slide) $a'(x) = -\gamma(x^2-1)/6$.
- The corresponding expansion for Z^* is

$$\mathrm{P}_{\widehat{G}}(Z^* \le z) = \Phi(z) + n^{-1/2} \widehat{a}(z)\phi(z) + O_p(n^{-1}).$$

- Typically $\widehat{a}(z) = a(z) + O_p(n^{-1/2})$, so

$$\mathrm{P}_{\widehat{G}}(Z^* \le z) - \mathrm{P}_G(Z \le z) = O_p(n^{-1}),$$

  so the order of error is n^{-1}.

Why Studentize? II

- Without Studentization, $Z = n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, \nu')$, and then

$$\mathrm{P}_G(Z \leq z) = \Phi\left(\frac{z}{\nu'^{1/2}}\right) + n^{-1/2} a' \left(\frac{z}{\nu'^{1/2}}\right) \phi\left(\frac{z}{\nu'^{1/2}}\right) + O(n^{-1})$$

  and

$$\mathrm{P}_{\widehat{G}}(Z^* \leq z) = \Phi\left(\frac{z}{\widehat{\nu}'^{1/2}}\right) + n^{-1/2} \widehat{a}' \left(\frac{z}{\widehat{\nu}'^{1/2}}\right) \phi\left(\frac{z}{\widehat{\nu}'^{1/2}}\right) + O_p(n^{-1}).$$

- Typically $\widehat{\nu}' = \nu' + O_p(n^{-1/2})$, giving

$$\mathrm{P}_{\widehat{G}}(Z^* \leq z) - \mathrm{P}_G(Z \leq z) = O_p(n^{-1/2}),$$

  and the difference in the leading terms means that the overall error is of order $n^{-1/2}$.
- Thus Studentizing reduces error from $O_p(n^{-1/2})$ to $O_p(n^{-1})$: better than using large-sample asymptotics, for which error is usually $O_p(n^{-1/2})$.

Other confidence intervals

- Simpler approaches:
  - **Basic bootstrap** interval: treat $\widehat{\theta} - \theta$ as pivot, get

$$\widehat{\theta} - (\widehat{\theta}^*_{((R+1)(1-\alpha))} - \widehat{\theta}), \quad \widehat{\theta} - (\widehat{\theta}^*_{((R+1)\alpha)} - \widehat{\theta}).$$

  - **Percentile interval**: use empirical quantiles of $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$:

$$\widehat{\theta}^*_{((R+1)\alpha)}, \quad \widehat{\theta}^*_{((R+1)(1-\alpha))}.$$

- The percentile interval is transformation-invariant, not the basic bootstrap interval.
- **Bias-corrected and accelerated (BC_a)** intervals replace percentile interval with $(\widehat{\theta}^*_{((R+1)\alpha')}, \widehat{\theta}^*_{((R+1)(1-\alpha''))})$, where

$$\alpha' = \Phi\left\{w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right\}, \quad w = \Phi^{-1}\left\{\widehat{G}^*(\widehat{\theta})\right\}, \quad a = \frac{1}{6}\frac{\sum_{j=1}^n l_j^3}{\left(\sum_{j=1}^n l_j^2\right)^{3/2}},$$

  with $\widehat{G}^*$ the EDF of the $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$, and $l_1, \ldots, l_n$ the empirical influence values (soon).
- If the **bias** $w = 0$, then $\widehat{G}^*(\widehat{\theta}) = \frac{1}{2}$, so $\widehat{\theta}$ is at the median of the EDF of $\widehat{\theta}^*$
- If the **acceleration** $a = 0$, then the effect of the data $y_1, \ldots, y_n$ on $\widehat{\theta}$ is symmetric.

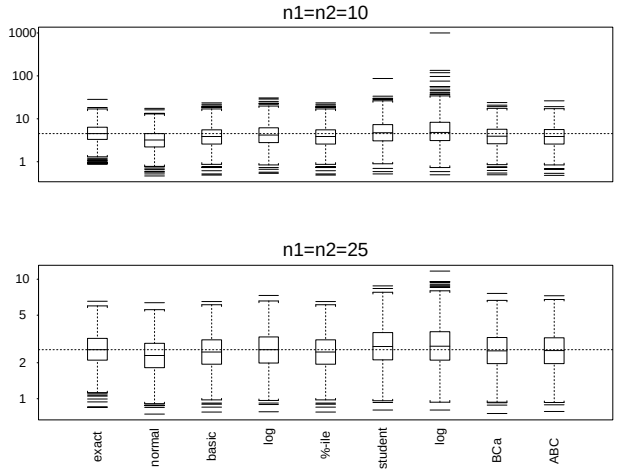
Comparisons

Table 2: Empirical error rates (%) for nonparametric bootstrap confidence limits in ratio estimation: rates for sample sizes $n_1 = n_2 = 10$ are given above those for sample sizes $n_1 = n_2 = 25$. $R = 999$ for all bootstrap methods. 10,000 data sets generated from Gamma distributions.

Method	Nominal error rate							
	Lower limit				Upper limit			
	1	2.5	5	10	10	5	2.5	1
Exact	1.0	2.8	5.5	10.5	9.8	4.8	2.6	1.0
	1.0	2.3	4.8	9.9	10.2	4.9	2.5	1.1
Normal approximation	0.1	0.5	1.7	6.3	20.6	15.7	12.5	9.6
	0.1	0.5	2.1	6.4	16.3	11.5	8.2	5.5
Basic bootstrap	0.0	0.0	0.2	1.8	24.4	21.0	18.6	16.4
	0.0	0.1	0.4	3.0	19.2	15.0	12.5	10.3
Basic bootstrap, log scale	2.6	4.9	8.1	12.9	13.1	7.5	4.8	2.5
	1.6	3.2	6.0	11.4	11.5	6.3	3.3	1.7
Studentized bootstrap	0.6	2.1	4.6	9.9	11.9	6.7	4.0	2.0
	0.8	2.3	4.6	9.9	10.9	5.9	3.0	1.4
Studentized bootstrap, log scale	1.1	2.8	5.6	10.7	11.6	6.3	3.5	1.7
	1.1	2.5	5.0	10.1	10.8	5.7	2.9	1.3
Bootstrap percentile	1.8	3.6	6.5	11.6	14.6	8.9	5.9	3.3
	1.2	2.6	5.1	10.1	12.6	7.1	4.2	2.1
BC_a	1.9	4.0	6.9	12.3	14.0	8.3	5.3	3.0
	1.4	3.0	5.6	10.9	11.8	6.8	3.8	1.9
ABC	1.9	4.2	7.4	12.7	14.6	8.7	5.5	3.1
	1.3	3.0	5.7	11.0	12.1	6.8	3.7	1.9

Confidence interval lengths

Lengths of 95% confidence intervals for the first 1000 simulated samples in the numerical experiment with Gamma data.



Discussion

- ☐ Bootstrap confidence intervals usually under-cover (i.e., are too short).
- ☐ Normal, basic, and studentized intervals depend on scale.
- ☐ Percentile interval often too short but is transformation-invariant.
- ☐ Studentized intervals give best coverage overall, but
  - they depend on scale, can be sensitive to V;
  - their lengths can be very variable;
  - they are best when V is approximately constant.
- ☐ Improved percentile intervals have same asymptotic error as Studentized intervals, but often are shorter, so give lower coverage probabilities.
- ☐ Caution: Edgeworth theory OK for smooth statistics, but beware rough statistics: must check output.
- ☐ Typically need $R > 1000$ for reliable estimation of quantiles.

Nonparametric delta method

- ☐ The **delta method** (Theorem 40) gives variance formulae for functions of averages.
- ☐ More generally we use the **nonparametric delta method**, which is based on the linear functional expansion

$$t(F) \doteq t(G) + \int L_t(x; G)\, \mathrm{d}F(x),$$

where L_t, the first derivative of $t(\cdot)$ at G, is defined by

$$L_t(y; G) = \lim_{\varepsilon \to 0} \frac{t\{(1-\varepsilon)G + \varepsilon H_y\} - t(G)}{\varepsilon} = \frac{\partial t\,\{(1-\varepsilon)G + \varepsilon H_y\}}{\partial \varepsilon}\bigg|_{\varepsilon=0},$$

with $H_y(u) \equiv H(u - y)$ the Heaviside function jumping from 0 to 1 at $u = y$.

- ☐ The **influence function value** $L_t(y; G)$ for the statistical functional t for an observation at y when the background distribution is G, satisfies $\mathrm{E}_G\{L_t(Y; G)\} = 0$.
- ☐ If $\widehat{G}$ is based on a random sample $y_1, \ldots, y_n$, then the jth **empirical influence value** is

$$l_j = L_t(y_j; \widehat{G}),$$

and $\mathrm{E}_{\widehat{G}}\{L_t(Y; \widehat{G})\} = n^{-1} \sum_j l_j = 0$.

- ☐ The influence function also plays an important role in robust statistics.

Nonparametric delta method II

☐ If we replace F by the EDF $\widehat{G}$ for a random sample $Y_1, \ldots, Y_n$, then

$$t(\widehat{G}) \doteq t(G) + \int L_t(x; G)\, \mathrm{d}\widehat{G}(x) = t(G) + \frac{1}{n}\sum_{j=1}^{n} L_t(Y_j; G),$$

has variance

$$\mathrm{var}\{t(\widehat{G})\} \doteq \frac{1}{n^2}\sum_{j=1}^{n} L_t^2(Y_j; G) = V_L,$$

say, which we estimate based on a sample $y_1, \ldots, y_n$ by $v_L = n^{-2}\sum l_j^2$.

Example 51 *Apply the nonparametric delta method to the average $\overline{Y}$.*

Example 52 *Apply the nonparametric delta method to a statistic defined by an estimating equation, and hence find the variance of the ratio $\overline{V}/\overline{U}$ for data pairs $Y = (U, V)$.*

Example 51

☐ The population mean and its empirical version are

$$\theta = t(G) = \int x\, dG(x), \quad \widehat{\theta} = t(\widehat{G}) = \int x\, d\widehat{G}(x) = n^{-1}\sum_{j=1}^{n} Y_j = \overline{Y}.$$

☐ If H_y puts unit mass at y, its 'density' is a Dirac delta function $\delta_y(x)$, and

$$\begin{aligned} \theta\{(1-\varepsilon)G + \varepsilon H_y\} &= \int x\, \mathrm{d}\{(1-\varepsilon)G + \varepsilon H_y\}(x) \\ &= (1-\varepsilon)\int x\, \mathrm{d}G(x) + \varepsilon \int x\, \mathrm{d}H_y(x) = (1-\varepsilon)\theta(G) + \varepsilon y \end{aligned}$$

and therefore

$$L(y; G) = \lim_{\varepsilon \to 0} \frac{\theta\{(1-\varepsilon)G + \varepsilon H_y\} - \theta(G)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\theta(G) + \varepsilon y - \theta(G)}{\varepsilon} = y - \theta(G),$$

☐ Hence the empirical influence values and variance estimate are

$$l_j = L(y_j; \widehat{G}) = y_j - \overline{y}, \qquad v_L = \frac{1}{n^2}\sum (y_j - \overline{y})^2 = \frac{n-1}{n} n^{-1} s^2.$$

Example 52

☐ The scalar parameter $\theta = t(G)$ is determined implicitly through the estimating equation

$$\int a(x;\theta)\,\mathrm{d}G(x) = \int a\{x;t(G)\}\,\mathrm{d}G(x) = 0.$$

We replace G by $G_\varepsilon = (1-\varepsilon)G + \varepsilon H_y$ and see that

$$\begin{aligned} 0 &= \int a\{x;t(G_\varepsilon)\}\ \mathrm{d}G_\varepsilon(x) \\ &= (1-\varepsilon)\int a\{x;t(G_\varepsilon)\}\ \mathrm{d}G(x) + \varepsilon\int a\{x;t(G_\varepsilon)\}\ \mathrm{d}H_y(x) \\ &= (1-\varepsilon)\int a\{x;t(G_\varepsilon)\}\ \mathrm{d}G(x) + \varepsilon a\{y;t(G_\varepsilon)\}, \end{aligned}$$

and differentiation using the chain rule gives

$$0 = a\{y;t(G_\varepsilon)\} - \int a\{x;t(G_\varepsilon)\}\ \mathrm{d}G(x) + \varepsilon a_\theta\{y;t(G_\varepsilon)\}\frac{\partial t(G_\varepsilon)}{\partial\varepsilon} + (1-\varepsilon)\int a_\theta\{x;t(G_\varepsilon)\}\frac{\partial t(G_\varepsilon)}{\partial\varepsilon}\,\mathrm{d}G(x),$$

which reduces to

$$0 = a\{y;t(G)\} + \int a_\theta\{x;t(G)\}\ \mathrm{d}G(x)\ \frac{\partial t(G)}{\partial\varepsilon}\bigg|_{\varepsilon=0}$$

on setting $\varepsilon = 0$. Hence

$$L_t(y;G) = \frac{\partial t(G_\varepsilon)}{\partial\varepsilon}\bigg|_{\varepsilon=0} = \frac{a(y;\theta)}{-\int a_\theta(x;\theta)\,\mathrm{d}G(x)}, \quad \text{where} \quad a_\theta(x;\theta) = \frac{\partial a(x;\theta)}{\partial\theta}.$$

☐ In the case of the ratio and with $y = (u,v)$, we take $a(y;\theta) = v - \theta u$, so

$$\theta = \theta(G) = \int v\,\mathrm{d}G(u,v) \Big/ \int u\,\mathrm{d}G(u,v), \quad \widehat{\theta} = \overline{v}/\overline{u},$$

and $a_\theta = -u$, so $l_j = (x_j - \widehat{\theta}u_j)/\overline{u}$, giving

$$v_L = \frac{1}{n^2}\sum\left(\frac{x_j - \widehat{\theta}u_j}{\overline{u}}\right)^2.$$

Comments

- For statistics involving only averages (ratio, correlation coefficient, ...), the nonparametric delta method retrieves the delta method.
- For example, the correlation coefficient may be written as a function of $\overline{xu} = n^{-1}\sum x_j u_j$, etc.:

$$\widehat{\theta} = \frac{\overline{xu} - \overline{x}\,\overline{u}}{\left\{(\overline{x^2} - \overline{x}^2)(\overline{u^2} - \overline{u}^2)\right\}^{1/2}},$$

  from which empirical influence values l_j can be derived, giving $v_L = 0.029$ for the handedness data, to be compared with $v = 0.043$ obtained by bootstrapping.
- v_L typically underestimates $\mathrm{var}(\widehat{\theta})$!
- The l_j can also be obtained by numerical differentiation if $t(\widehat{G})$ is coded appropriately, or approximated using a jackknife method.

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 vs frequentist inference

Observed data y^{o} assumed to be realisation of $Y \sim f(y;\theta) \equiv f(y \mid \theta)$, where $\theta \in \Theta$.

- **Frequentist viewpoint**:
  - some 'true value' of θ generated the data;
  - this 'true value' of θ is treated as an unknown constant;
  - probability statements compare y^{o} with outcomes in a suitable reference set $\mathcal{S}$.
- **Bayesian viewpoint**:
  - degrees of belief should (and can) be expressed using probability distributions;
  - knowledge about θ prior to seeing y^{o} is expressed as a **prior density** $\pi(\theta)$;
  - Bayes' theorem
$$\pi(\theta \mid y^{\text{o}}) = \frac{\pi(\theta) f(y^{\text{o}} \mid \theta)}{\int \pi(\theta) f(y^{\text{o}} \mid \theta)\, \mathrm{d}\theta}$$
    should be used to convert $\pi(\theta)$ into a **posterior density** $\pi(\theta \mid y^{\text{o}})$;
  - probability statements are based on $\pi(\theta \mid y^{\text{o}})$ and thus are conditioned on all observed quantities.
- The benefit is that statistics reduces to calculations of probabilities, at the cost of expressing prior information in distributional terms.

Example

Example 53 *(a) Find the posterior density for the success probability θ based on a series of independent Bernoulli trials $y_1, \ldots, y_n$, when the prior density is the* **Beta density**

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \quad 0 < \theta < 1, \quad a, b > 0,$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the **beta function**, *and*

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} \, \mathrm{d}u$$

is the gamma function.
(b) Show how the mean and variance of θ are updated.
(c) Find the posterior density for predicting the result Z of the next trial.

Example 53

☐ Suppose that conditional on θ, the data $y_1, \ldots, y_n$ are a random sample from the Bernoulli distribution, for which $\mathrm{P}(Y_j = 1) = \theta$ and $\mathrm{P}(Y_j = 0) = 1 - \theta$, where $0 < \theta < 1$. The likelihood is

$$L(\theta) = f(y \mid \theta) = \prod_{j=1}^{n} \theta^{y_j}(1-\theta)^{1-y_j} = \theta^s(1-\theta)^{n-s}, \quad 0 < \theta < 1,$$

where $s = \sum y_j$.

☐ A natural prior here is the beta density with parameters a and b,

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1, \quad a, b > 0, \tag{8}$$

where $B(a,b)$ is the beta function $\Gamma(a)\Gamma(b)/\Gamma(a+b)$.

☐ The posterior density of θ conditional on the data is

$$\begin{aligned} \pi(\theta \mid y) &= \frac{\theta^{s+a-1}(1-\theta)^{n-s+b-1}/B(a,b)}{\int_0^1 \theta^{s+a-1}(1-\theta)^{n-s+b-1}\,\mathrm{d}\theta/B(a,b)} \\ &\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1. \end{aligned} \tag{9}$$

As (8) has unit integral for all positive a and b, the constant normalizing (??) must be $B(a+s, b+n-s)$. Therefore

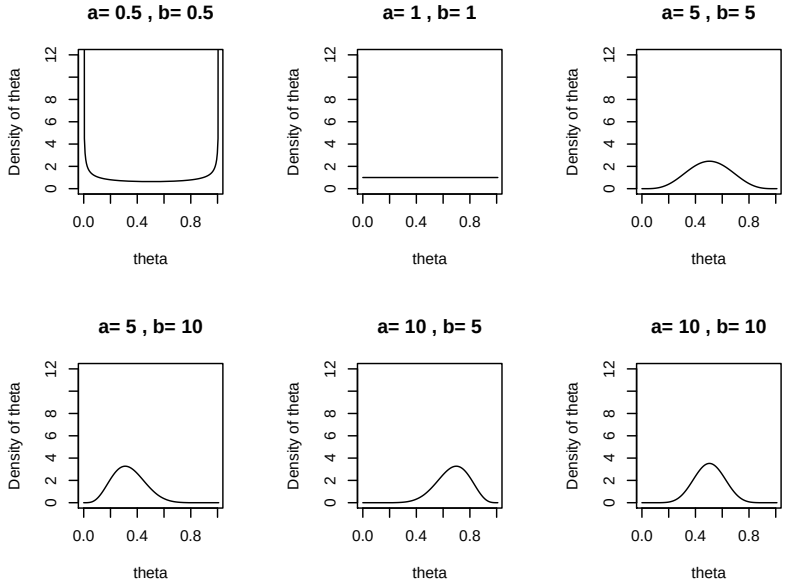
$$\pi(\theta \mid y) = \frac{1}{B(a+s, b+n-s)}\theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1.$$

☐ Thus the posterior density of θ has the same form as the prior: acquiring data has the effect of updating (a,b) to $(a+s, b+n-s)$. As the mean of the $B(a,b)$ density is $a/(a+b)$, the posterior mean is $(s+a)/(n+a+b)$, and this is roughly s/n in large samples. Hence the prior density inserts information equivalent to having seen a sample of $a+b$ observations, of which a were successes. If we were very sure that $\theta \doteq 1/2$, for example, we might take $a = b$ very large, giving a prior density tightly concentrated around $\theta = 1/2$, whereas taking smaller values of a and b would increase the prior uncertainty.

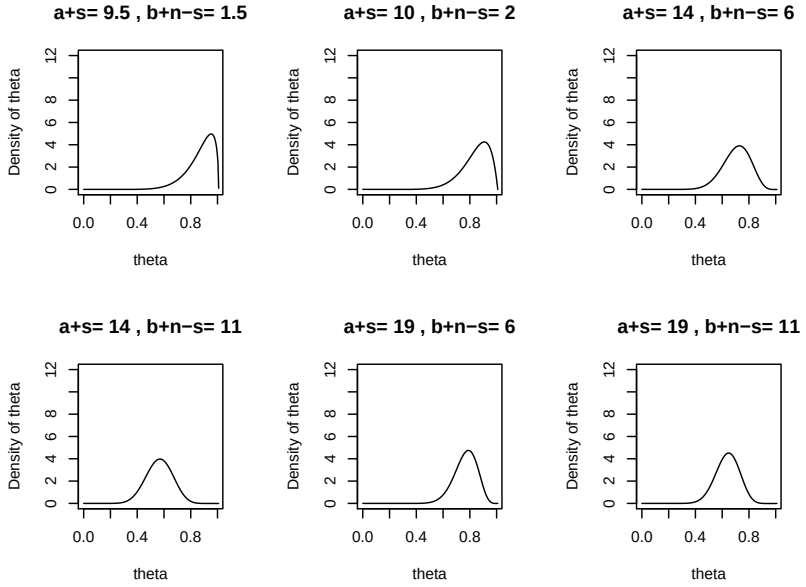
100 spins of a 5Fr coin

```
1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 0 1 1
1 1 1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 1 0 1
1 1 1 0 0 1 0 1 1 1 1 1 0 0 1 1 1 1 1 1
1 0 1 0 1 1 0 1 1 1 0 0 1 1 1 0 1 1 1 1
1 0 0 0 0 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0
```

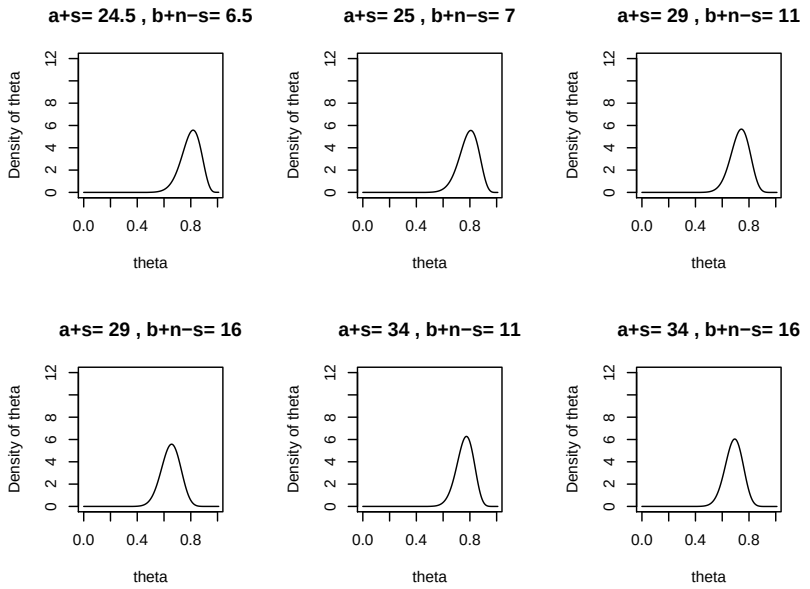

Beta prior densities



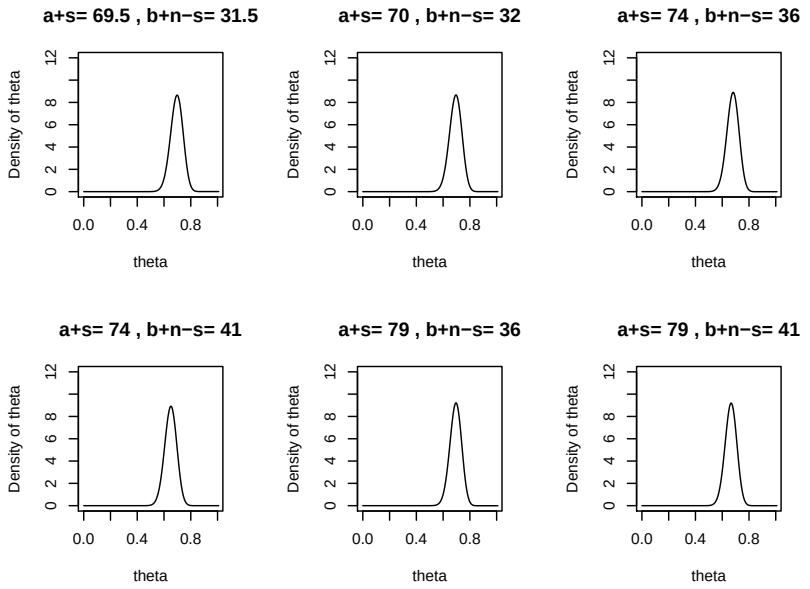
$n = 10$, $s = 9$



$n = 30,\ s = 24$



$n = 100,\ s = 69$



Link to likelihood

- In large samples the prior has less influence, because

$$\log \pi(\theta \mid y) = \log \pi(\theta) + \ell(\theta) - \log f(y),$$

  where the terms on the right are successively $O(1)$, $O(n)$ and $O(n)$.
- Later we shall see that

$$f(y) \doteq \left(\frac{2\pi}{\widehat{\jmath}}\right)^{1/2} \pi(\widehat{\theta}) e^{\ell(\widehat{\theta})}$$

  in terms of the MLE $\widehat{\theta}$ and observed information $\widehat{\jmath}$, so

$$\pi(\theta \mid y) \doteq \frac{\pi(\theta)}{\pi(\widehat{\theta})} \times \left(\frac{\widehat{\jmath}}{2\pi}\right)^{1/2} e^{\ell(\theta)-\ell(\widehat{\theta})} \doteq \frac{\pi(\theta)}{\pi(\widehat{\theta})} \times \left(\frac{\widehat{\jmath}}{2\pi}\right)^{1/2} e^{-\widehat{\jmath}(\widehat{\theta}-\theta)^2/2},$$

  giving the distributional approximation

$$\theta \mid y \dot{\sim} \mathcal{N}(\widehat{\theta}, \widehat{\jmath}^{-1}).$$

- Formal versions of this result, known as **Bernstein–von Mises theorems**, suggest that large-sample Bayesian and likelihood-based inferences will be similar.
- Hence we need to consider situations in which the prior may be appreciable relative to the information in the data, or in which standard likelihood approaches are unsuitable.

Conjugate priors

- Certain combinations of data model $f(y \mid \theta)$ and prior $\pi(\theta)$ 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.
- Conjugate priors greatly simplify computation and are widely used in modelling.
- Mixtures of conjugate priors are also conjugate (problem in Week 1).

Lemma 54 *An exponential family density*

$$f(y \mid \theta) = m(y) \exp[s(y)\varphi(\theta) - k\{\varphi(\theta)\}], \quad y \in \mathcal{Y}, \theta \in \Theta,$$

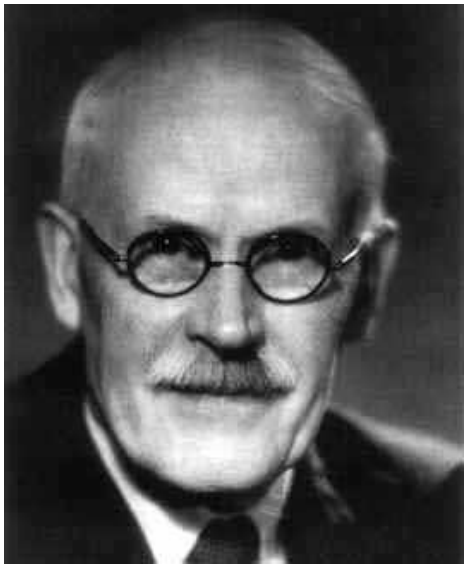
has conjugate prior

$$f(\theta; a, b) = h(a, b) \exp[a\varphi(\theta) - bk\{\varphi(\theta)\}], \quad \theta \in \Theta,$$

that depends on **hyperparameters** a, b.

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' about what?

Definition 55

- ☐ *A* **uniform prior** *satisfies* $\pi(\theta) \propto 1$ *for* $\theta \in \Theta$.
- ☐ *An* **improper prior** *cannot be renormalised to have finite integral.*
- ☐ *The* **Jeffreys prior** *for a statistical model with Fisher information* $\imath(\theta)$ *is* $\pi(\theta) \propto |\imath(\theta)|^{1/2}$.

Example 56 *What does a uniform prior for* $\theta \in (0, 1)$ *imply for* $\psi = \log\{\theta/(1-\theta)\} \in \mathbb{R}$*?*

Lemma 57 *The Jeffreys prior is invariant to smooth reparametrizations* $\theta = \theta(\psi)$.

- ☐ Jeffreys priors were introduced to give 'objective' expressions of ignorance, and give uniform priors for location parameters, $1/\theta$ for scale parameters, etc.
- ☐ Jeffreys priors for the same θ based on different experiments might differ!
- ☐ Many other attempts to represent 'ignorance' have been made (e.g., by providing priors with minimal information), but none is seen as fully satisfactory.
- ☐ In practice 'uninformative' (i.e., flat but proper) priors are usually chosen and then sensitivity analyses performed.

Example 56

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: $\pi(\theta) = 1$, $0 \leq \theta \leq 1$. 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

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

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 57

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

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

  Consequently $|\imath(\theta)|^{1/2}\mathrm{d}\theta = |\imath(\psi)|^{1/2}\mathrm{d}\psi$: the Jeffreys prior does behave consistently under reparametrization; furthermore such priors give widely-accepted solutions in some standard problems. When θ is vector, $|\imath(\theta)|$ is taken to be the determinant of $\imath(\theta)$.
- 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.

High dimensions

Example 58 (Stein's paradox) *Let $Y_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, 1)$ for $j = 1, \ldots, n$, and set $D = \sum Y_j^2$ and $\theta = \mu_1^2 + \cdots + \mu_n^2$. Show that if the μ_j are independent a priori with flat priors, then*

$$\mathrm{E}(\theta \mid y) = D + n, \quad \textit{but} \quad D \approx \theta + n + O_p(n^{1/2})$$

for any θ, which is absurd.

- Thus although flat priors may be sensible in low dimensions, they can lead to major problems in high dimensions.
- If we seek an uninformative prior for a scalar parameter ψ when nuisance parameters $\lambda_1, \ldots, \lambda_p$ are orthogonal to ψ, we can set

$$\pi(\psi, \lambda) \propto \imath_{\psi\psi}^{1/2}(\psi, \lambda) \times g(\lambda),$$

  where $\imath_{\psi\psi}(\psi, \lambda)$ is the (ψ, ψ) element of the Fisher information matrix and $g(\lambda)$ is an arbitrary function of the nuisance parameter.

Example 58

- If $y \mid \mu \sim \mathcal{N}(\mu, 1)$ and $\pi(\mu) \propto 1$, then symmetry of the normal density ϕ gives

$$\pi(\mu \mid y) = \frac{\phi(y-\mu)}{\int \phi(y-\mu)\,\mathrm{d}\mu} = \frac{\phi(\mu-y)}{\int \phi(\mu-y)\,\mathrm{d}y} = \phi(\mu-y),$$

  so $\mu \mid y \sim \mathcal{N}(y, 1)$. If this is true independently for all the y_j, then

$$\mathrm{E}(\theta \mid y) = \sum_{j=1}^{n} \mathrm{E}(\mu_j^2 \mid y) = \sum_{j=1}^{n} \left\{\mathrm{E}(\mu_j \mid y_j)^2 + \mathrm{var}(\mu_j \mid y_j)\right\} = \sum_{j=1}^{n} (y_j^2 + 1) = D + n,$$

  and its posterior variance is $\mathrm{var}(\theta \mid y) = \sum_{j=1}^{n} \mathrm{var}(\mu_j^2 \mid y_j) = 2n + 4D = O(n)$.
- On the other hand, for large n we have $D = \sum Y_j^2 \approx \mathrm{E}(D) = \sum_{j=1}^{n} (\mu_j^2 + 1) = \theta + n$ and $\mathrm{var}(D) = 2n + 4\theta = O(n)$.
- This implies that the posterior is placing probability in the wrong place asymptotically, i.e., around $D + n$ instead of around $D - n$. Hence the posterior probability that θ lies in any interval $D - n \pm a\sqrt{n}$ tends to zero.

Matching priors

Definition 59 *The* **posterior α quantile** *of a scalar parameter θ satisfies*

$$\mathrm{P}_{\theta|Y}\left\{\theta \le \theta^{\alpha}(y) \mid y\right\} = \int_{-\infty}^{\theta^{\alpha}} \pi(\theta \mid y)\,\mathrm{d}\theta = \alpha, \quad \alpha \in (0,1)$$

- Consider random sample $Y_1, \ldots, Y_n$ with joint density $f(y \mid \theta)$, with prior $\pi(\theta)$ and $\theta \in \mathbb{R}^d$, and let $\widehat{\theta}$ be the MLE and $\widehat{\sigma}^2/n = \widehat{\jmath}^{-1}$ its asymptotic variance.
- Bayes and likelihood inferences will agree as $n \to \infty$, but is (approximate?) agreement achievable for small n?
- If for every $\alpha \in (0,1)$ and $\theta \in \Theta$ we could have

$$\mathrm{P}_{Y|\theta}\left\{\theta^{\alpha}(Y) \ge \theta\right\} = \int I\{\theta^{\alpha}(y) \ge \theta\} f(y \mid \theta)\,\mathrm{d}y = \alpha,$$

  then Bayes and frequentist inference would agree perfectly, and we would have
  - a Bayes/frequentist compromise;
  - default priors for routine Bayesian use; and
  - a basis for assessment of robustness of inference using other priors.

Edgeworth series

We use asymptotic approximations to compare the Bayesian and frequentist solutions.

Definition 60 *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*

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

Matching: scalar θ

- ☐ We now compute Edgeworth series for the Bayesian quantity $n^{1/2}(\theta - \widehat{\theta})/\widehat{\sigma}$, conditional on y (so $\widehat{\theta}(y), \widehat{\sigma}(y)$ are constants), invert it to get the corresponding Cornish–Fisher series

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

  and then insert this expansion into

$$\mathrm{P}_{Y|\theta}\left\{\theta^\alpha(Y) \geq \theta\right\} = \int I\{\theta^\alpha(y) \geq \theta\} f(y \mid \theta)\, dy.$$

- ☐ This gives

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

  where

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

- ☐ Choosing π to knock out T_1 will ensure matching to order n^{-1}, etc.

Discussion

- ☐ Clearly $T_1(\pi, \theta) \equiv 0$ if and only if
$$\pi(\theta) \propto \imath(\theta)^{1/2},$$
so the Jeffreys prior is matching to order n^{-1} for scalar θ.
- ☐ With the Jeffreys prior, T_2 vanishes if and only if
$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left\{\frac{\mathrm{E}_{Y|\theta}(\ell_\theta^3)}{\imath(\theta)^{3/2}}\right\} = 0,$$
so even for scalar θ, higher-order matching is only possible in special cases.
- ☐ In the vector case, inferences for ψ match to order n^{-1} if ψ is orthogonal to the other parameters λ, and
$$\pi(\psi, \lambda) \propto \imath_{\psi\psi}^{1/2}(\psi, \lambda) \times g(\lambda).$$
- ☐ In general impossible to match for all parameters simultaneously—need separate (and incompatible) priors for each parameter.
- ☐ Higher order matching requires data-dependent priors.
- ☐ Kass and Wasserman (1996, JASA) give a general discussion of **reference priors**.

Inference

Once we have a prior, what about

- ☐ confidence sets?
- ☐ prediction?
- ☐ hypothesis tests?
- ☐ model comparison?
- ☐ model checking?

Confidence sets

- ☐ All measures of uncertainty are computed from the relevant posterior density.
- ☐ Posterior confidence bound for θ is quantile of $\pi(\theta \mid y)$:

$$\mathrm{P}\left\{\theta \leq \theta^{\alpha}(y) \mid y\right\} = \int_{-\infty}^{\theta^{\alpha}(y)} \pi(\theta \mid y)\,\mathrm{d}\theta = \alpha, \quad \alpha \in (0,1),$$

  giving $(1-2\alpha)$ posterior **credible set** $(\theta^{\alpha}(y), \theta^{1-\alpha}(y))$.
- ☐ In multiparameter case we use the marginal α quantile of ψ, $\psi^{\alpha} \equiv \psi^{\alpha}(y)$ as

$$\mathrm{P}(\psi \leq \psi^{\alpha} \mid y) = \frac{\int_{-\infty}^{\psi^{\alpha}} \int f(y;\psi,\lambda)\pi(\psi,\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}{\iint f(y;\psi,\lambda)\pi(\psi,\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}\alpha, \quad \alpha \in (0,1),$$

  based on the marginal posterior density of ψ.
- ☐ A **highest posterior density (HPD) credible set** $\mathcal{C}_{1-\alpha}$ satisfies $\mathrm{P}(\theta \in \mathcal{C}_{1-\alpha} \mid y) = 1-\alpha$ and $\sup_{\theta \notin \mathcal{C}_{1-\alpha}} \pi(\theta \mid y) \leq \inf_{\theta \in \mathcal{C}_{1-\alpha}} \pi(\theta \mid y)$.
- ☐ Such intervals/sets are interpreted as probability statements about the the parameter, with y fixed, contrary to frequentist confidence intervals.
- ☐ Likewise prediction intervals are based on the posterior predictive distribution $\mathrm{P}(Z \leq z \mid y)$.

Example

Mortality rates r/m from cardiac surgery in 12 hospitals, showing the numbers of deaths r out of m operations.

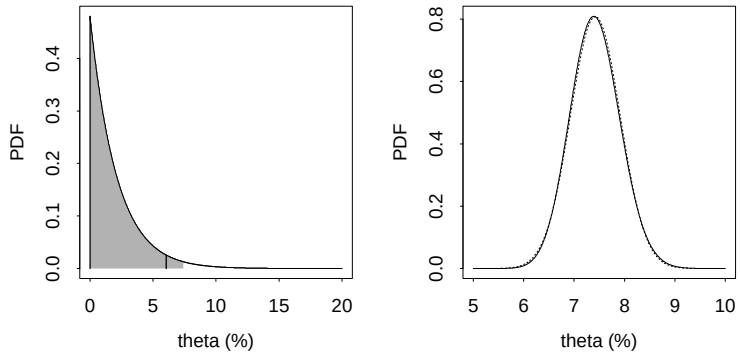
A	0/47	B	18/148	C	8/119	D	46/810	E	8/211	F	13/196
G	9/148	H	31/215	I	14/207	J	8/97	K	29/256	L	24/360

Example 61 (Cardiac surgery data) *A simple model for the data above treats the number of deaths r as binomial with mortality rate θ and denominator m. At hospital A, for example, $m = 47$ and $r = 0$, giving maximum likelihood estimate $\widehat{\theta}_A = 0/47 = 0$, but it seems too optimistic to suppose that θ_A could be so small when the other rates are evidently larger. If we take a beta prior density with $a = b = 1$, the posterior density is beta with parameters $a + r = 1$ and $b + m - r = 48$. The 0.95 HPD credible interval is $(0, 6.05)\%$, while the equitailed credible interval uses the 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ and is $(0.05, 7.40)\%$.*

Show that the maximum posterior density estimator can be regarded as a penalized MLE.

Example

Cardiac surgery data. Left panel: posterior density for θ_A, showing boundaries of 0.95 highest posterior credible interval (vertical lines) and region between posterior 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ (shaded). Right panel: exact posterior beta density for overall mortality rate θ (solid) and normal approximation (dots).



Bayes factors

- Bayes factors compare competing models/hypotheses.
- Given prior probabilities $\mathrm{P}(H_0)$ and $\mathrm{P}(H_1)$ for two hypotheses, we compute

$$\mathrm{P}(H_i \mid y) = \frac{\mathrm{P}(y \mid H_i)\mathrm{P}(H_i)}{\mathrm{P}(y \mid H_0)\mathrm{P}(H_0) + \mathrm{P}(y \mid H_1)\mathrm{P}(H_1)}, \quad i = 0, 1.$$

- Unlike in frequentist testing,
  - prior probabilities for the H_i must be specified, and
  - we compute the probability of each hypothesis given the data.
- To avoid specifying the prior probabilities we write

$$\frac{\mathrm{P}(H_1 \mid y)}{\mathrm{P}(H_0 \mid y)} = \frac{\mathrm{P}(y \mid H_1)}{\mathrm{P}(y \mid H_0)} \times \frac{\mathrm{P}(H_1)}{\mathrm{P}(H_0)} = B_{10} \times \frac{\mathrm{P}(H_1)}{\mathrm{P}(H_0)},$$

  where B_{10} is the **Bayes factor**, and usually

$$\mathrm{P}(y \mid H_i) = \int f(y \mid H_i, \theta_i)\pi(\theta_i \mid H_i)\,\mathrm{d}\theta_i, \quad i = 0, 1.$$

Interpretation

☐ Often $2 \log B_{10}$ is used to summarise the evidence for H_1, using a table like

B_{10}	$2 \log B_{10}$	Evidence for H_1
1–3	0–2	Hardly worth a mention
3–20	2–6	Positive
20–150	6–10	Strong
> 150	> 10	Very strong

☐ As $B_{10} = B_{01}^{-1}$, the evidence for H_0 is $2 \log B_{01} = -2 \log B_{10}$.

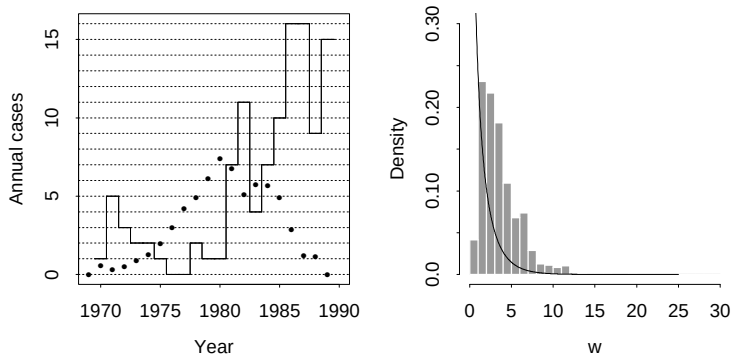
☐ Models $f(y \mid H, \theta)$ for n observations and $d \times 1$ parameter θ often compared using

$$\mathrm{BIC} = -2\ell(\widehat{\theta}) + d \log n,$$

which can be derived by approximating the **model evidence** $\mathrm{P}(y \mid H)$.

Example

Changepoint analysis for data on diarrhoea-associated haemolytic uraemic syndrome (HUS). Left: counts of cases of HUS treated in Birmingham, 1970–1989 (solid), and scaled likelihood ratio statistic $W_{\mathrm{p}}(\tau)/10$ (blobs). Right: density of W, estimated from 10,000 simulations, and χ^2_1 density (solid).



Example

Example 62 (HUS data) *The graph suggests a sharp rise in incidence around 1980. Suppose the annual counts $y_1, \ldots, y_n$ are realizations of independent Poisson variables with means λ_1 for $j = 1, \ldots, \tau$ and λ_2 for $j = \tau + 1, \ldots, n$. Here the changepoint τ can take values $1, \ldots, n-1$. Under H_0, $\lambda_1 = \lambda_2 = \lambda$, that is, no change, and H_τ allows change after year τ. If we suppose that λ_1 and λ_2 have independent gamma prior densities with parameters γ and δ, then B_{10} can be computed for each τ.*

There is very strong evidence for change in any year from 1976 to 1986, with most evidence for a change after 1980.

	1970	*1971*	*1972*	*1973*	*1974*	*1975*	*1976*	*1977*	*1978*	*1979*
y	*1*	*5*	*3*	*2*	*2*	*1*	*0*	*0*	*2*	*1*
$2\log B_{\tau 0}$, $\gamma = \delta = 1$	*4.9*	-0.5	*0.6*	*3.9*	*7.5*	*13*	*24*	*35*	*41*	*51*
$2\log B_{\tau 0}$, $\gamma = \delta = 0.01$	-1.3	-5.9	-4.5	-1.0	*3.0*	*9.7*	*20*	*32*	*39*	*51*
$2\log B_{\tau 0}$, $\gamma = \delta = 0.0001$	-10	-15	-14	-10	-6.1	0.6	11	23	30	42

	1980	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*	*1988*	*1989*
y	*1*	*7*	*11*	*4*	*7*	*10*	*16*	*16*	*9*	*15*
$2\log B_{\tau 0}$, $\gamma = \delta = 1$	*63*	*55*	*38*	*42*	*40*	*31*	*11*	-2.9	-5.3	*0*
$2\log B_{\tau 0}$, $\gamma = \delta = 0.01$	*64*	*57*	*40*	*47*	*46*	*38*	*18*	*1.8*	*1.2*	*0*
$2\log B_{\tau 0}$, $\gamma = \delta = 0.0001$	55	48	31	38	37	29	8.8	-7.1	-7.7	0

Nested models

- ☐ Often $\theta = (\psi, \lambda)$ and we want to compare $H_0 : \psi = \psi_0$ against $H_1 : \psi \neq \psi_0$.
- ☐ A prior density on θ will give

$$\mathrm{P}(H_0) = \iint_{\{(\psi,\lambda):\psi=\psi_0\}} \pi(\psi, \lambda)\, \mathrm{d}\lambda \mathrm{d}\psi = 0,$$

  so the posterior odds in favour of H_1 are infinite for any dataset.
- ☐ To avoid we use prior densities weighted according to prior belief in H_0 and H_1, giving overall prior

$$\pi(\psi, \lambda) = \delta(\psi - \psi_0)\pi(\psi_0, \lambda \mid H_0)\mathrm{P}(H_0) + \pi(\psi, \lambda \mid H_1)\mathrm{P}(H_1),$$

  where

$$\int \pi(\psi_0, \lambda \mid H_0)\, d\lambda = \int \pi(\psi, \lambda \mid H_1)\, d\psi d\lambda = 1.$$

- ☐ Hence Bayes factors are more sensitive to the prior than are posterior densities.
- ☐ Improper priors cannot be used, as B_{10} depends on the ratio of the two arbitrary constants of proportionality in the priors.

Jeffreys–Lindley paradox

- Test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ when $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.
- Frequentist computes P-value $p_{\text{obs}} = \Phi(-n^{1/2}|\overline{y}|/\sigma)$.
- Bayesian writes $\pi_0 = \mathrm{P}(H_0)$, supposes that under H_1, $\mu \sim \mathcal{N}(0, \tau^2)$ and computes

$$B_{01} = \left(1 + n\frac{\tau^2}{\sigma^2}\right)^{1/2} \exp\left\{-\frac{n\overline{y}^2}{2\sigma^2(1 + n^{-1}\sigma^2/\tau^2)}\right\}$$

- If $n\overline{y}^2/\sigma^2 = z^2_{\alpha/2}$, then $p_{\text{obs}} = \alpha$, but B_{01} gives increasingly strong evidence in favour of H_0; see the table, in which $\alpha = 0.01$:

n	1	10	100	1000	10^4	10^6	10^8
B_{01}	0.269	0.163	0.376	1.15	3.63	36.2	362

- The problem is that as $n \to \infty$, $\pi(\mu \mid H_1)$ is increasingly dispersed compared to $|\overline{y} - 0|$.
- To resolve this, note that we use tests when there is doubt about the hypotheses, i.e., sensible alternatives are $O(n^{-1/2})$ from the null, and if we take this account by setting $\tau^2 = \delta\sigma^2/n$, then the paradox dissipates, because (for example) with $\delta = 10$ and $\alpha = 0.05$, 0.01, 0.001, and 0.0001, $B_{10} = 1.73$, 6.2, 41.4, and 293, in broad agreement.

Model criticism

- Use marginal density $f(y)$ to check the model (and degree of agreement between $\pi(\theta)$ and $f(y \mid \theta)$. Simplest if

$$f(y) = f(y \mid s)f(a) \int f(t \mid a, \theta)\pi(\theta)\,\mathrm{d}\theta,$$

  where s is sufficient and a ancillary.
- Often leads to (Bayesian variants of) standard diagnostics (e.g., residuals, ...).
- Another measure of plausibility based on possible new dataset $Y_+ \sim f$ is

$$\mathrm{P}\{f(Y_+) \leq f(y^{\text{o}})\},$$

  and yet another is based on **predictive diagnostics**, comparing a discrepancy measure $D_+ = d(Y_+, \theta)$ with its predictive distribution, i.e.,

$$\mathrm{P}\left\{d(Y_+, \theta) \geq d(y, \theta) \mid y\right\},$$

  where the averaging is over both Y_+ and the posterior distribution of θ.
- We choose $d(Y_+, \theta)$ to measure some key aspects of the data and model.

Prediction and model averaging

- Predict unobserved Z based on observed $Y = y$ from a single model by computing $f(z \mid y)$, but if there are several models, then

$$f(z \mid y) = \sum_{i=1}^{k} f(z \mid y, M_i) \mathrm{P}(M_i \mid y),$$

which averages the posterior distributions of z under the different models, weighted according to their posterior probabilities

$$\mathrm{P}(M_i \mid y) = \frac{f(y \mid M_i)\mathrm{P}(M_i)}{\sum_{l=1}^{k} f(y \mid M_l)\mathrm{P}(M_l)},$$

where

$$\begin{aligned} f(y \mid M_i) &= \int f(y \mid \theta_i, M_i)\pi(\theta_i \mid M_i)\,\mathrm{d}\theta_i, \\ f(z \mid M_i, y) &= \frac{\int f(z \mid y, \theta_i, M_i) f(y \mid \theta_i, M_i)\pi(\theta_i \mid M_i)\,\mathrm{d}\theta_i}{f(y \mid M_i)}. \end{aligned}$$

- If we have all possible models, the main problem is computational ...

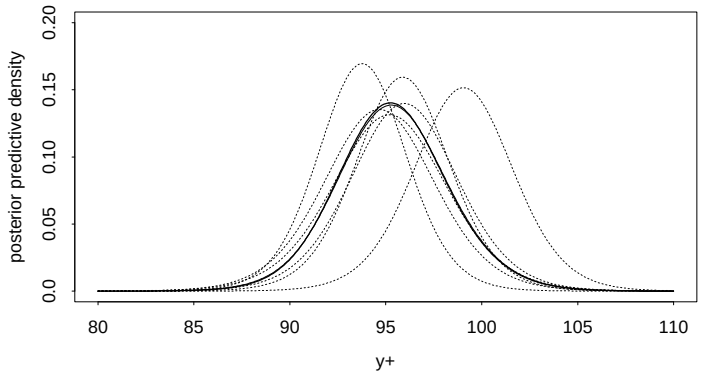
Example

Bayesian prediction using model averaging for the cement data. For each of the 16 possible subsets of covariates, the table shows the log Bayes factor in favour of that subset compared to the model with no covariates and gives the posterior probability of each model. The values of the posterior mean and scale parameters a and b are also shown for the six most plausible models; $(y_+ - a)/b$ has a posterior t density. For comparison, the residual sums of squares are also given.

Model	RSS	$2 \log B_{10}$	$\mathrm{P}(M \mid y)$	a	b
– – – –	2715.8	0.0	0.0000		
1 – – –	1265.7	7.1	0.0000		
– 2 – –	906.3	12.2	0.0000		
– – 3 –	1939.4	0.6	0.0000		
– – – 4	883.9	12.6	0.0000		
1 2 – –	57.9	45.7	0.2027	93.77	2.31
1 – 3 –	1227.1	4.0	0.0000		
1 – – 4	74.8	42.8	0.0480	99.05	2.58
– 2 3 –	415.4	19.3	0.0000		
– 2 – 4	868.9	11.0	0.0000		
– – 3 4	175.7	31.3	0.0002		
1 2 3 –	48.11	43.6	0.0716	95.96	2.80
1 2 – 4	47.97	47.2	0.4344	95.88	2.45
1 – 3 4	50.84	44.2	0.0986	94.66	2.89
– 2 3 4	73.81	33.2	0.0004		
1 2 3 4	47.86	45.0	0.1441	95.20	2.97

Example

Posterior predictive densities for cement data. Predictive densities for y_+ based on individual models are given as dotted curves, and the heavy curve is the averaged prediction from all 16 models.



Arguments for/against Bayes

- ☐ For:
  - provides unified approach to inference—all unknowns, data, parameters, predictands are treated on the same footing;
  - simple recipe — just apply Bayes' theorem and compute ...
  - gives results similar to likelihood inferences (in large samples);
  - argument based on axioms of 'rational behaviour' under uncertainty leads to 'coherent' (i.e., internally consistent) Bayes inference;
- ☐ Against:
  - is it always (ever?) appropriate to treat data (whose model is checkable) on the same basis as the prior?
  - Different priors may give different answers. Which is to be believed by a third party?
  - How do we agree on a prior?
  - External validity (in the frequency sense) with respect to reality is more important than internal consistency (one can be consistently wrong!)
- ☐ In any case, modelling can be flexible and general, provided computation is possible ...

Motivation

☐ We often want to approximate integrals such as those in the marginal posterior density

$$\pi(\psi \mid y) = \frac{\int f(y;\psi,\lambda)\pi(\psi,\lambda)\,\mathrm{d}\lambda}{\iint f(y;\psi,\lambda)\pi(\psi,\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}$$

or the corresponding marginal posterior distribution function

$$\mathrm{P}(\psi \leq \psi^0 \mid y) = \frac{\int_{-\infty}^{\psi^0}\int f(y;\psi,\lambda)\pi(\psi,\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}{\iint f(y;\psi,\lambda)\pi(\psi,\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}.$$

☐ Different approaches exist:

- **deterministic** approximations include
  - ▷ quadrature rules — only work in low dimensions, not much used;
  - ▷ variational Bayes — provides numerical bounds on some integrals;
  - ▷ Laplace approximation — accurate analytical method with wide applications;
- **Monte Carlo** approximations include
  - ▷ importance sampling — uses independent samples, can be unstable;
  - ▷ Markov chain Monte Carlo — widespread use in applications.

Laplace's method

Lemma 63 *Let $h(u)$ be a smooth convex function defined for $u \in \mathbb{R}$, with a minimum at $u = \tilde{u}$, where $h'(\tilde{u}) = 0$ and $h''(\tilde{u}) > 0$, and let*

$$I_n = \int_{-\infty}^{\infty} e^{-nh(u)}\,du.$$

Then

$$I_n = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})} \times \left\{1 + n^{-1}\left(\frac{5h_3^2}{24h_2^3} - \frac{h_4}{8h_2^2}\right) + O\left(n^{-2}\right)\right\},$$

where $h_2 = h''(\tilde{u})$, etc. The leading term $\tilde{I}_n$ is known as the **Laplace approximation** *to I_n.*

Comments:

☐ the error is relative, so the approximation is often very accurate far into the tails;

☐ $\tilde{I}_n$ involves only h and its second derivative at $\tilde{u}$, so can be computed numerically;

☐ the series is asymptotic, so the partial sums may not converge, and including more than the leading term may give no improvements;

☐ most of the normal probability lies within ± 3 SD of the mean, so the limits of the integral don't matter (much) provided they lie outside the interval $\tilde{u} \pm 3(nh_2)^{-1/2}$;

☐ the exponent is written $-nh(u)$ only for formal justification of the approximation; in practice we set $n = 1$.

Note to Lemma 63

Close to $\tilde{u}$ a Taylor series expansion gives $h(u) \doteq h(\tilde{u}) + \frac{1}{2}h_2(u-\tilde{u})^2$, so

$$\begin{aligned} I_n &\doteq e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-nh_2(u-\tilde{u})^2/2}\, du \\ &= e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-z^2/2} \frac{du}{dz}\, dz \\ &= \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})}, \end{aligned}$$

where the first and second equalities use the substitution $z = (nh_2)^{1/2}(u-\tilde{u})$ and the fact that the normal density has unit integral. A more detailed accounting gives the required result.

Laplace's method: General case

Lemma 64 *Let $h(u)$ be a smooth convex function defined for $u \in \mathbb{R}^d$, with a minimum at $u = \tilde{u}$, where $\mathrm{d}h(\tilde{u})/\mathrm{d}u = 0$ and the hessian matrix*

$$h_2 \equiv \frac{\mathrm{d}^2 h(\tilde{u})}{\mathrm{d}u \mathrm{d}u^{\mathrm{T}}}$$

is positive definite, and let

$$I_n = \int_{\mathbb{R}^d} e^{-nh(u)}\, du.$$

Then

$$I_n = \tilde{I}_n \left\{1 + O(n^{-1})\right\} = \left(\frac{2\pi}{n}\right)^{p/2} |h_2|^{-1/2} e^{-nh(\tilde{u})} \left\{1 + O(n^{-1})\right\}.$$

Example 65 *Use Laplace approximation to derive the Bayesian information criterion.*

Note to Example 65

- Laplace approximation to $\log f(y)$ gives

  $$\log \pi(\tilde{\theta}) + \log f(y \mid \tilde{\theta}) + \frac{p}{2}\log(2\pi/n) - \tfrac{1}{2}\log|\tilde{\jmath}| + O(n^{-1}),$$

  where $\tilde{\theta}$ maximises $\log \pi(\theta) + \log f(y \mid \theta)$ and $\tilde{\jmath} = -n^{-1}$ times the hessian matrix of this function, evaluated at $\tilde{\theta}$.
- Now $p\log(2\pi) - \log|\tilde{\jmath}|$ is of order 1 as $n \to \infty$, and so is $\log \pi(\tilde{\theta})$, and $\tilde{\theta} = \widehat{\theta} + O(n^{-1})$, so

  $$-2\log f(y) \doteq -2\log f(y \mid \widehat{\theta}) + p\log n + O(1) \approx \mathrm{BIC}.$$

Integral approximation

Lemma 66 *Let*

$$J_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{u_0} a(u) e^{-ng(u)} \left\{1 + O(n^{-1})\right\}\, du,$$

where $g(u)$ is a smooth convex function defined for $u \in \mathbb{R}$, and in addition to possessing the properties of h in Lemma 1, g satisfies $g(\tilde{u}) = 0$. Also let $a(u) > 0$. Then

$$J_n(u_0) = \Phi(n^{1/2} r_0^*) + O(n^{-1}),$$

where

$$r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right), \quad r_0 = \text{sign}(u_0 - \tilde{u})\{2g(u_0)\}^{1/2}, \quad v_0 = \frac{g'(u_0)}{a(u_0)}.$$

Example 67 *Use the methods above to approximate the posterior conditional distribution*

$$\mathrm{P}(\theta \le \theta_0 \mid y)$$

of a scalar parameter θ based on a random sample $y_1, \ldots, y_n$ from a regular model, and outline how posterior confidence intervals for θ are obtained.

Note to Lemma 66

☐ The first step is to change the variable of integration from u to $r(u) = \text{sign}(u - \tilde{u})\{2g(u)\}^{1/2}$; that is, $r^2/2 = g(u)$. Then $g'(u) = dg(u)/du$ and $r(u)$ have the same sign, and $rdr/du = g'(u)$, so

$$\begin{aligned} J_n(u_0) &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0} a(u) \frac{r}{g'(u)} e^{-nr^2/2} \left\{1 + O(n^{-1})\right\}\, dr \\ &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0} e^{-nr^2/2 + \log b(r)} \left\{1 + O(n^{-1})\right\}\, dr, \end{aligned}$$

where $b(r) = a(u)r/g'(u) > 0$ is regarded as a function of r.

☐ We now change variable again, from r to $r^* = r - (rn)^{-1} \log b(r)$, so

$$-nr^{*2} = -nr^2 + 2\log b(r) - n^{-1} r^{-2} \{\log b(r)\}^2.$$

The Jacobian of the transformation and the third term in $-nr^{*2}$ contribute only to the error of $J_n(u_0)$, so

$$\begin{aligned} J_n(u_0) &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0^*} e^{-nr^{*2}/2} \left\{1 + O(n^{-1})\right\}\, dr^* \\ &= \Phi(n^{1/2} r_0^*) + O(n^{-1}), \end{aligned} \tag{10}$$

where

$$r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right), \quad r_0 = \text{sign}(u_0 - \tilde{u})\{2g(u_0)\}^{1/2}, \quad v_0 = \frac{g'(u_0)}{a(u_0)}.$$

Note to Example 67

☐ We write

$$\mathrm{P}(\theta \leq \theta_0 \mid y) = \frac{\int_{-\infty}^{\theta_0} \pi(\theta) f(y \mid \theta)\,\mathrm{d}\theta}{\int_{-\infty}^{\infty} \pi(\theta) f(y \mid \theta)\,\mathrm{d}\theta}$$

and set $h(\theta) = -n^{-1}\{\ell(\theta) + \log \pi(\theta)\} = -\ell_m(\theta)/n$, say. This (scaled) modified log likelihood is maximised at $\tilde{\theta}$, which is the maximum a posteriori estimate of θ, and $h''(\theta) = -n^{-1}\ell_m''(\theta) = n^{-1}\jmath(\theta) - n^{-1}(\log \pi)''(\theta)$.

☐ Laplace approximation of the denominator integral gives

$$\sqrt{\frac{2\pi}{nh_2}} \exp\{-nh(\tilde{\theta})\} \left\{1 + O(n^{-1})\right\},$$

where $h_2 = h''(\tilde{\theta})$, and inserting this into the expression for the posterior probability gives

$$\mathrm{P}(\theta \leq \theta_0 \mid y) = \sqrt{\frac{nh_2}{2\pi}} \int_{-\infty}^{\theta_0} e^{-n\{h(\theta) - h(\tilde{\theta})\}} \left\{1 + O(n^{-1})\right\}\,\mathrm{d}\theta,$$

to which we can apply Lemma 66 with $g(\theta) = h(\theta) - h(\tilde{\theta}) \geq 0$; this equals zero when $\theta = \tilde{\theta}$, and $a(\theta) = (nh_2)^{1/2}$. We take $u = \theta$, $u^0 = \theta_0$,

$$v_0 = g'(\theta_0)/(nh_2)^{1/2} = -n^{-1}\ell_m'(\theta_0)/\{-\ell_m''(\tilde{\theta})\}^{1/2}, \quad r_0 = \mathrm{sign}(\theta_0 - \tilde{\theta}) \left[2\{\ell_m(\tilde{\theta}) - \ell_m(\theta_0)\}/n\right]^{1/2},$$

and therefore

$$n^{1/2} r_0^* = n^{1/2} r_0 - \frac{1}{n^{1/2} r_0} \log\left\{\frac{-\ell_m'(\theta_0)/\{-\ell_m''(\tilde{\theta})\}^{1/2}}{n^{1/2} r_0}\right\}.$$

Hence we can simply set $n = 1$ and compute $r_0 = \mathrm{sign}(\theta_0 - \tilde{\theta}) \left[2\{\ell_m(\tilde{\theta}) - \ell_m(\theta_0)\}\right]^{1/2}$.

☐ Hence we can write

$$\mathrm{P}(\theta \leq \theta_0 \mid y) = \Phi\{r_B^*(\theta_0)\} \left\{1 + O(n^{-1})\right\},$$

where $r_B^*(\theta_0)$ is given by the expressions above with $n = 1$. We obtain confidence intervals by solving for θ_0 the equations

$$\alpha, 1 - \alpha = \Phi\{r_B^*(\theta_0)\}, \quad \text{or equivalently} \quad z_\alpha, z_{1-\alpha} = r_B^*(\theta_0).$$

☐ The likelihood root (almost) corresponds to setting $\pi(\theta) \propto 1$, so that $\tilde{\theta} = \widehat{\theta}$ and $nh_2 = \widehat{\jmath}$, and then we get

$$r_0 = -\mathrm{sign}(\widehat{\theta} - \theta_0) \left[2\{\ell(\widehat{\theta}) - \ell(\theta_0)\}\right]^{1/2}, \quad v_0 = -\widehat{\jmath}^{-1/2}\ell'(\theta_0).$$

This makes sense, because

$$\mathrm{P}(\theta \leq \theta_0 \mid y) \doteq \Phi\{r_B^*(\theta_0)\}$$

is increasing in θ_0, but the corresponding expression for a frequentist interval is decreasing in θ_0. So we expect that $r_B^*(\theta_0) \doteq -r^*(\theta_0)$.

Integral approximation: General case

Lemma 68 *Let $u = (u_1, u_2)$, where u_1 is scalar and u_2 a $p \times 1$ vector, and consider*

$$J_n(u_1^0) = (2\pi)^{-(p+1)/2} c \int_{-\infty}^{u_1^0} \int \exp\{-ng(u_1, u_2)\}\, \mathrm{d}u_2 \mathrm{d}u_1, \tag{11}$$

where c is constant, the inner integral being over $\mathbb{R}^p$. Here g is supposed to have its previous smoothness properties, to be maximized at $(\tilde{u}_1, \tilde{u}_2)$, and satisfies $g(\tilde{u}_1, \tilde{u}_2) = 0$. Then

$$J_n(u_1^0) = \Phi(n^{1/2} r_0^*) + O(n^{-1}),$$

where $r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right)$, with

$$r_0 = \mathrm{sign}(u_1^0 - \tilde{u}_1)\left\{2g(u_1^0, \tilde{u}_{20})\right\}^{1/2}, \quad v_0 = c^{-1} \frac{\partial g(u_1^0, \tilde{u}_{20})}{\partial u_1} \left| g_{22}(u_1^0, \tilde{u}_{20}) \right|^{1/2},$$

where $\tilde{u}_{20}$ is the maximizing value of u_2 when $u_1 = u_1^0$.

Multivariate case

- The computations of Example 67 can be extended to the multiparameter case using Lemmas 64 and 68, and give

$$\mathrm{P}(\psi \le \psi_0 \mid y) = \Phi\{r_B^*(\psi_0)\}\left\{1 + O(n^{-1})\right\},$$

where $r_B^*(\psi_0) = r_B(\psi_0) + r_B(\psi_0)^{-1} \log\{v_B(\psi_0)/r_B(\psi_0)\}$, with

$$r_B(\psi_0) = \mathrm{sign}(\psi_0 - \tilde{\psi}) \left[2\left\{\ell_m(\tilde{\psi}, \tilde{\lambda}) - \ell_m(\psi_0, \tilde{\lambda}_{\psi_0})\right\}\right]^{1/2},$$

$$v_B(\psi_0) = -\frac{\partial \ell_m(\psi_0, \tilde{\lambda}_{\psi_0})}{\partial \psi} \left\{ \frac{\left| -\frac{\partial^2 \ell_m(\psi_0, \tilde{\lambda}_{\psi_0})}{\partial \lambda \partial \lambda^{\mathrm{T}}} \right|}{\left| -\frac{\partial^2 \ell_m(\tilde{\psi}, \tilde{\lambda})}{\partial \theta \partial \theta^{\mathrm{T}}} \right|} \right\}^{1/2};$$

here $\tilde{\lambda}_{\psi_0}$ is the maximum *a posteriori* estimate of λ when ψ is fixed at ψ_0.

- Often we find the derivatives numerically.
- There is a close link to maximum likelihood estimation, because $\tilde{\theta} = \widehat{\theta} + O(n^{-1})$, so the order of error is not increased by using the MLEs instead of the MAPs — though the numerical approximations are not so good.

Frequentist aside

- In frequentist inference **saddlepoint approximation** is used to write conditional densities for exponential families as

$$f(t_1 \mid t_2; \psi) \doteq \left\{ \frac{|\jmath_{\lambda\lambda}(\widehat{\theta}_\psi)|}{2\pi |\jmath(\widehat{\theta})|} \right\}^{1/2} \exp\left\{ \ell(\widehat{\theta}_\psi) - \ell(\widehat{\theta}) \right\},$$

  leading to

$$\mathrm{P}(T_1 \leq t_1 \mid T_2 = t_2; \psi) \doteq \Phi\{r^*(\psi)\},$$

  where $r^*(\psi) = r(\psi) + r(\psi)^{-1} \log\{r(\psi)/v(\psi)\}$, with

$$r(\psi) = \mathrm{sign}(\widehat{\psi} - \psi)[2\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\}]^{1/2}, \; v(\psi) = (\widehat{\psi} - \psi) \left\{ \frac{|\jmath(\widehat{\theta})|}{|\jmath_{\lambda\lambda}(\widehat{\theta}_\psi)|} \right\}^{1/2}.$$

- Saddlepoint approximation involves writing the exponential family density as an integral of its Laplace transform (or equivalently its cumulant-generating function), and then approximating the resulting integral.
- The details are somewhat more painful, but the idea is similar to the Bayesian case.
- The approach sketched on slides 117–119 extends this to arbitrary regular models, by approximating them by exponential families.

Comments

- For successful approximation we must be able to write the integrand as

$$\exp\left\{\log f(y; \theta) + \log \pi(\theta)\right\},$$

  where the exponent is $O(n)$ and the integrand has one dominant mode.
- If so the methods can work well in fairly high dimensions, partly because the errors in numerator and denominator can cancel.
- However Monte Carlo methods are more flexible and in more general use ...

Importance sampling

- Seek to estimate
$$\mu = \int m(\theta, y, z)\pi(\theta \mid y)\, d\theta,$$
where taking, for example,
  - $m(\theta, y, z) = I(\theta \leq a)$ will give $\mu = \mathrm{P}(\theta \leq a \mid y)$,
  - $m(\theta, y, z) = f(z \mid y, \theta)$ will give $\mu = f(z \mid y)$.
- If we can sample $\theta_1, \ldots, \theta_R \overset{\text{iid}}{\sim} h(\theta)$, where the support of h includes that of $\pi(\theta \mid y)$, then we have an importance sampling estimator
$$\widehat{\mu} = R^{-1} \sum_{r=1}^{R} m(\theta_r, y, z) \frac{\pi(\theta_r \mid y)}{h(\theta_r)} = R^{-1} \sum_{r=1}^{R} m(\theta_r, y, z) w(\theta_r),$$
where $w(\theta) = \pi(\theta \mid y)/h(\theta)$ is an importance sampling weight.
- Advantage of $\widehat{\mu}$ over MCMC output is that its variance is readily obtained.
- Disadvantage is that choice of h is usually difficult. and especially if $\dim(\theta)$ is large, so huge samples are needed because most of the simulated θ_r receive zero weight and so are wasted.

Markov chain Monte Carlo

- Want to learn about distribution π of random variable $U \in \mathcal{U}$:
  - in Bayesian statistics, U is all unknowns and π is their posterior distribution conditioned on observed data y;
  - in frequentist statistics U may be functions of the data y, and we seek to condition on other functions, e.g., to perform a conditional test.
- Construct a Markov chain $\{U^t\}$ with state space $\mathcal{U}$ and transition kernel P, whose limiting distribution is π, i.e.,
$$\mathrm{P}(U^t \in \mathcal{A} \mid u^0) \to \pi(\mathcal{A}) \quad t \to \infty, \quad u^0 \in \mathcal{U}, \mathcal{A} \subset \mathcal{U}.$$
- We then use P to simulate a realisation $u^0, u^1, \ldots, u^R$ of the chain, and hence get estimates such as
$$\mathrm{E}_\pi\{g(U) \mid y\} = \int g(u)\pi(u \mid y)\, \mathrm{d}u \approx \frac{1}{R} \sum_{r=1}^{R} g(u^r), \quad \pi(\mathcal{A} \mid y) \approx \frac{1}{R} \sum_{r=1}^{R} I(u^r \in \mathcal{A}).$$
- Must choose P and u^0 so that
  - the distribution of U^t converges quickly to π (so minimise simulation effort);
  - $u^0, u^1, \ldots, u^R$ are as independent as possible (so have efficient estimation).

Markov chains

Definition 69

(a) A sequence $U^0, U^1, U^2, \ldots$ of elements of a set $\mathcal{U}$ is a **Markov chain** *if the conditional distribution of U^{t+1} given $U^1, \ldots, U^t$ depends only on U^t:*

$$\mathrm{P}(U^{t+1} \in \mathcal{A} \mid U^1, \ldots, U^t) = \mathrm{P}(U^{t+1} \in \mathcal{A} \mid U^t), \quad \mathcal{A} \subset \mathcal{U}.$$

We call $\mathcal{U}$ the **state space** *of the Markov chain.*

(b) A Markov chain has **stationary transition probabilities** *if the conditional distribution of U^{t+1} given U^t does not depend on t.*

(c) The distribution of U^0 is called the **initial distribution**, *and the conditional distribution*

$$P(u, \mathcal{A}) = \mathrm{P}(U^{t+1} \in \mathcal{A} \mid U^t = u)$$

is called the **transition probability distribution** *(or* **transition kernel***); this does not depend on t if the chain has stationary transition probabilities, and then we denote it by P.*

(d) The **stationary** *or* **invariant** *or* **equilibrium** *distribution of a Markov chain with transition kernel P satisfies*

$$\pi(\mathcal{A}) = \int P(u, \mathcal{A})\, \pi(\mathrm{d}u), \quad \mathcal{A} \subset \mathcal{U}.$$

Ergodicity and convergence

For the distribution of U^t to converge to a stationary distribution, the chain must satisfy three important properties:

- ☐ **irreducibility** — $\mathcal{U}$ does not split into separate parts when we run the chain on it, so the kernel P allows us to reach any point of $\mathcal{U}$ starting from anywhere else;
- ☐ **aperiodicity** — precludes the possibility of the 'limiting' distribution depending on the iteration number, i.e., eliminates possibilities like $a_n = (-1)^n$, which equals 1 if n is even and otherwise is odd;
- ☐ **positive recurrence** — every state is visited infinitely often, if the chain is run forever. This enables estimation of properties of that state.
- ☐ An irreducible, aperiodic, positive recurrent chain is called **ergodic**.
- ☐ The **ergodic theorem** states that an ergodic Markov chain has a unique stationary distribution π,

$$\mathrm{P}(U^t \in \mathcal{A} \mid U_0 = u) \to \pi(\mathcal{A}), \quad t \to \infty, \quad u \in \mathcal{U}, \mathcal{A} \subset \mathcal{U},$$

and if g is a real-valued function with $\int |g(u)| \pi(\mathrm{d}u) < \infty$, then

$$\frac{1}{R} \sum_{t=1}^{R} g(U^t) \xrightarrow{\text{a.s.}} \int g(u) \pi(\mathrm{d}u), \quad R \to \infty.$$

Detailed balance

- Modulo technical details (skipped here), the implication is that if we can find a transition kernel P with invariant distribution π, then we can generate samples (almost) from π.
- Why 'almost'? Because we run the chain for a finite number of steps, so in general our samples are not exactly from π.
- We now describe some standard recipes for building MCMC algorithms.
- For simplicity of exposition we take $\mathcal{U}$ to be countable, so $P \equiv P(u, v)$ for $u, v \in \mathcal{U}$.
- A sufficient condition for invariance is **detailed balance**:

$$\pi(u)P(u, v) = \pi(v)P(v, u), \quad u, v \in \mathcal{U}.$$

- This guarantees invariance because

$$\begin{aligned} \int P(u, \mathcal{A})\pi(\mathrm{d}u) &= \sum_{v \in \mathcal{A}} \sum_{u \in \mathcal{U}} \pi(u)P(u, v) \\ &= \sum_{v \in \mathcal{A}} \sum_{u \in \mathcal{U}} \pi(v)P(v, u) \\ &= \sum_{v \in \mathcal{A}} \pi(v) \sum_{u \in \mathcal{U}} P(v, u) = \pi(\mathcal{A}) \times 1 = \pi(\mathcal{A}). \end{aligned}$$

Metropolis–Hastings algorithm

- A very general algorithm to estimate a target density π, with many variants.
- Hastings (1970) generalised an idea of Metropolis et al. (1953):
  - given a current value u of the chain, construct a candidate new value (a 'proposal') v by drawing from an arbitrary density $q(v \mid u)$;
  - accept the proposal as the next state of the chain with probability

$$a(u, v) = \min\left\{1, \frac{\pi(v)q(u \mid v)}{\pi(u)q(v \mid u)}\right\}$$

    and otherwise leave u unchanged.
- The target density π is needed only up to the constant of proportionality, and only at u and the proposal v, so in particular the normalising constant is not needed.
- An important special case, the **Gibbs sampler**, updates each component u_i of u by successively writing $u = (u_i, u_{-i})$ and then replacing u_i with $v_i \sim \pi(u_i \mid u_{-i})$, where $\pi(u_i \mid u_{-i})$ is called the **full conditional density**.

Example 70 (Toy) *Construct a Metropolis–Hastings algorithm with $\mathcal{N}(0, 1)$ target density and proposal distribution $q(v \mid u) = \sigma^{-1}\phi\{(v - u)/\sigma\}$.*

Note: Detailed balance for the M-H algorithm

☐ First we note that

$$P(u,v) = q(v \mid u)a(u,v) + r(u)I(u=v),$$

where

$$r(u) = 1 - \int q(v \mid u)a(u,v)\,dv.$$

The first and second terms of $P(u,v)$ are the probability density for a move from u to v being proposed and accepted, and the probability that a move away from u is rejected.

☐ The Metropolis–Hastings update step satisfies detailed balance because

$$\begin{aligned}
\pi(u)P(u,v) &= \pi(u)q(v \mid u)\min\left\{1, \frac{\pi(v)q(u \mid v)}{\pi(u)q(v \mid u)}\right\} + \pi(u)r(u)I(u=v) \\
&= \pi(v)q(u \mid v)\min\left\{\frac{\pi(u)q(v \mid u)}{\pi(v)q(u \mid v)}, 1\right\} + \pi(v)r(v)I(v=u) \\
&= \pi(v)P(v,u).
\end{aligned}$$

Hence the corresponding Markov chain is reversible with equilibrium distribution π, provided it is irreducible and aperiodic.

Note to Example 70

☐ We need to work out the acceptance ratio

$$\frac{\pi(v)q(u \mid v)}{\pi(u)q(v \mid u)}$$

where

$$\pi(u) \propto e^{-u^2/2}, \quad q(u \mid v) = (2\pi\sigma^2)^{-1/2}e^{-(u-v)^2/2\sigma^2},$$

and this is

$$\frac{e^{-v^2/2} \times (2\pi\sigma^2)^{-1/2}e^{-(u-v)^2/2\sigma^2}}{e^{-u^2/2} \times (2\pi\sigma^2)^{-1/2}e^{-(v-u)^2/2\sigma^2}} = \exp\{\tfrac{1}{2}(u^2 - v^2)\},$$

so the move $u \mapsto v$ is accepted with probability $\min[1, \exp\{\frac{1}{2}(u^2 - v^2)\}]$.

☐ If $v^2 \leq u^2$ the acceptance ratio is greater than unity and the move is always accepted, whereas if $v^2 > u^2$ the move may not be accepted, and if $v^2 \gg u^2$ the move is very unlikely to be accepted.

☐ Note that

- we did not need the normalising constant for π to run the algorithm;
- the acceptance ratio does not depend on σ;
- the acceptance probability does depend on σ. With $W \sim U(0,1)$, it is

$$\begin{aligned}
\mathrm{P}(u \mapsto V \mid u) &= \mathrm{P}\left(W \leq \min[1, \exp\{\tfrac{1}{2}(u^2 - v^2)\}] \mid u\right) \\
&= \mathrm{P}(|V| \leq u \mid u) + \int_{\{v:|v|>|u|\}} e^{(u^2-v^2)/2} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(v-u)^2/2\sigma^2}\,du,
\end{aligned}$$

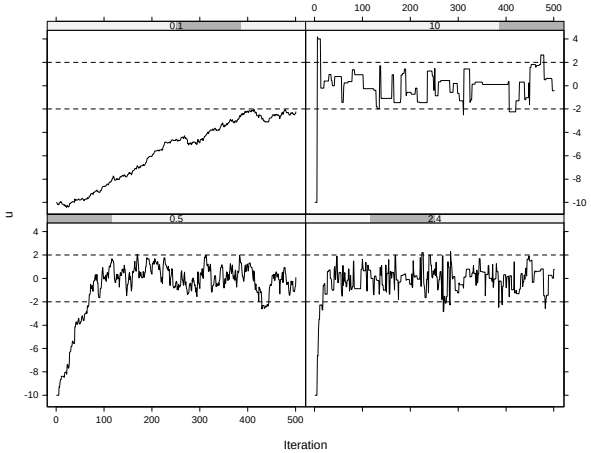
which clearly depends on u and on σ.

Toy MH example: Code

```
toy.MH <- function(R=5000, sig=1, u0=-10, seed)
{
  set.seed(seed)
  u <- rep(u0,R)
  for (r in 2:R)
   {
     v <- rnorm(1, u[r-1], sig)
     log.ratio <- dnorm(v, log=T) + dnorm(v, mean=u[r-1], sd=sig, log=T) -
                  dnorm(u[r-1], log=T) - dnorm(u[r-1], mean=v, sd=sig, log=T)
     a <- min( 1, exp(log.ratio) )
     u[r] <- u[r-1]
     if (runif(1)<=a) u[r] <- v
      }
  u
}

save.seed <- .Random.seed  # use the same seed for each simulation
out1 <- toy.MH(sig=0.1, seed=save.seed)
out2 <- toy.MH(sig=0.5, seed=save.seed)
plot.ts(out1, ylim=c(-10,3), xlab="Iteration", ylab="u")
plot.ts(out2, ylim=c(-10,3), xlab="Iteration", ylab="u")
```


Toy MH example

Simulations from a Metropolis–Hastings algorithm with $\mathcal{N}(0,1)$ target density, with $u^0 = -10$ and random walk proposal $v \sim \mathcal{N}(u, \sigma^2)$ with $\sigma = 0.1, 0.5, 2.4, 10$.



With $\sigma = 0.1, 0.5$, proposals often accepted but chain moves too slowly. With $\sigma = 10$ chain gets stuck for too long. Here $\sigma = 2.4$ seems best.

Proposal distributions

- In principle there is an (almost) completely free choice for the proposal distributions q_i, but just a few possibilities are typically used:
  - **Independence Metropolis–Hastings**, in which $q(v)$ is unrelated to u. Not much use in practice, but helpful for theoretical analysis.
  - **Random walk Metropolis**, in which $q(u, v) = q(v - u)$ and $q(\cdot)$ is a density symmetric about 0, giving
  $$a(u, v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}$$
  because $q(u, v) = q(v, u)$. This amounts to setting $v = u + \varepsilon$, where $\varepsilon \sim q$.
  - **Random walk Metropolis on the log scale**, applied when $u > 0$, in which random walk Metropolis is applied to $\log u$; then $q(v, u)/q(u, v) = v/u$ and so
  $$a(u, v) = \min\left\{1, \frac{\pi(v)v}{\pi(u)u}\right\}.$$
  Similar random walks can be applied to other transformations.

Toy Gibbs sampler

Example 71 *Find the joint posterior density for the mean and standard deviation of a normal random sample of size n with prior distributions $\mu \sim \mathcal{N}(\xi, \kappa^{-1})$ and $\sigma^{-2} \sim \Gamma(\alpha, \beta)$.*

Note to Example 71

The joint posterior is

$$\pi(\mu, \sigma^{-2} \mid y) \propto (\sigma^{-2})^{\alpha+n/2-1} \exp\left\{-\frac{\beta}{\sigma^2} - \frac{\kappa(\mu - \xi)^2}{2} - \frac{\sum(y_j - \mu)^2}{2\sigma^2}\right\}$$

so the parameters are dependent *a posteriori* although they were independent *a priori*. The full conditional densities are

$$\begin{aligned}
\mu \mid \sigma, y &\sim \mathcal{N}\left(\frac{\sum y_j + \sigma^2\kappa\xi}{n + \kappa\sigma^2}, \frac{1}{n\sigma^{-2} + \kappa}\right), \\
\frac{1}{\sigma^2} \mid \mu, y &\sim \Gamma\left(\alpha + n/2, \beta + \sum(y_j - \mu)^2/2\right),
\end{aligned}$$

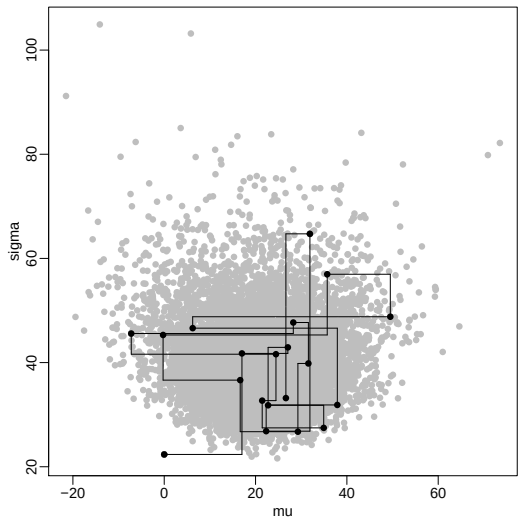
and the Gibbs sampler alternates updates of μ and of σ^{-2} using these two equations.

Toy Gibbs example: Code

```
# Darwin's maize data in eighths of an inch
n <- 15
y <- c(49,-67,8,16,6,23,28,41,14,29,56,24,75,60,-48)

# Set (improper) prior parameters and number of iterations R
xi <- kappa <- alpha <- beta <- 0
R <- 10000

# Gibbs sampler with initial values mu=0, 1/sig^2=0.002
out <- matrix(NA,R,2)
out[1,] <- c(0, 0.002)
for (r in 2:R)
{
new.mean <- (sum(y) + kappa*xi/out[r-1,2])/(n+kappa/out[r-1,2])
new.var <- 1/(n*out[r-1,2] + kappa)
out[r,1] <- rnorm(1, mean=new.mean, sd=sqrt(new.var))
out[r,2] <- rgamma(1, rate=beta+sum((y-out[r,1])^2)/2, shape=alpha+n/2)
}

# posterior values of sigma
out[,2] <- sqrt(1/out[,2])
```

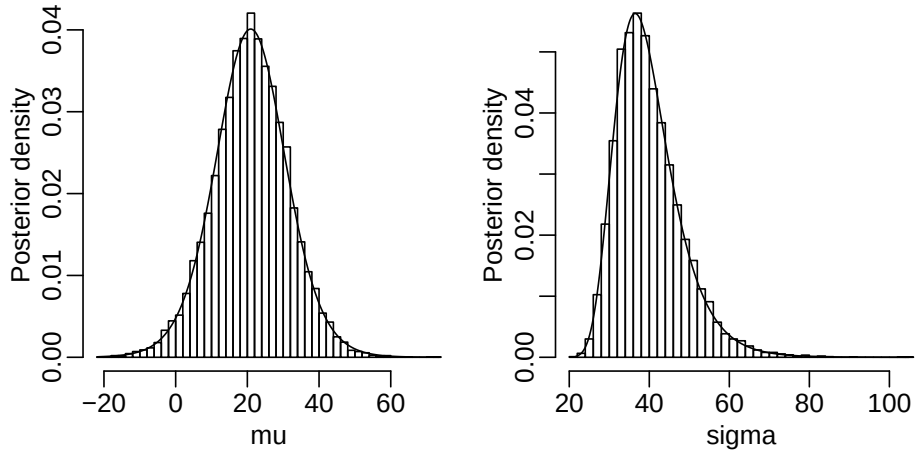

Toy Gibbs example

10,000 iterations of Gibbs sampler for (μ, σ), with initial value $\mu = 0$; the (μ update, σ update) steps are shown for the first 20 iterations:



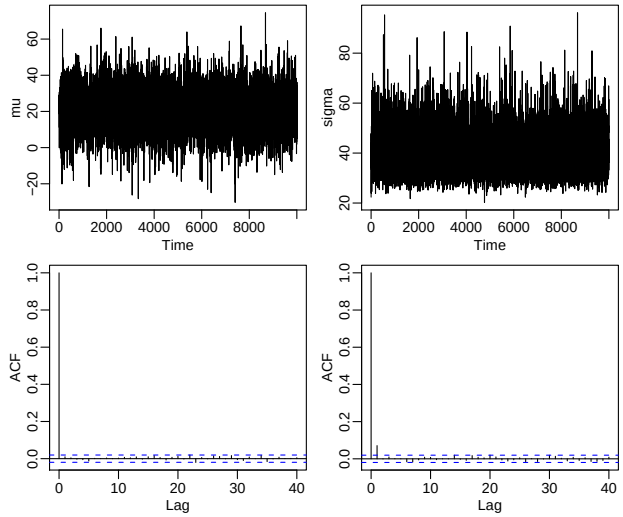
Toy Gibbs example

Marginal histograms and density estimates for μ and σ, based on 10,000 simulations:



Toy Gibbs example

Time series of μ and σ, and correlograms. The series appear to be stationary with (very) low autocorrelation:



Toy Gibbs example: Estimation

- Any function of (μ, σ) can be estimated using the successive pairs $(\mu, \sigma)_1, \ldots, (\mu, \sigma)_R$.
- For example, to compute $\psi = \mathrm{P}(Y_+ \leq -50 \mid y)$ we can either add simulation of a new observation Y_+ to each iteration, giving $(\mu, \sigma, Y_+)_1, \ldots, (\mu, \sigma, Y_+)_R$, or we can use conditioning to obtain the estimators

$$\widehat{\psi}_1 = \frac{1}{R}\sum_{r=1}^{R} I(Y_{+,r} \leq -50), \quad \widehat{\psi}_2 = \frac{1}{R}\sum_{r=1}^{R} \Phi\left(\frac{-50-\mu_r}{\sigma_r}\right).$$

  The maximum likelihood estimator of ψ is $\widehat{\psi} = \Phi\{(-50-\widehat{\mu})/\widehat{\sigma}\} = 0.030$, where $\widehat{\mu}$, $\widehat{\sigma}$ are the MLEs, but the Bayes estimator is $\widehat{\psi}_2 = 0.045$, which is larger because it allows for the variability of the parameters (though it depends on the prior).
- Similar arguments apply to estimation of marginal densities, using either by a kernel density estimator or an unbiased estimator based on the full conditionals. For example,

$$\pi(\mu \mid y) \doteq R^{-1}\sum_{r=1}^{R} \frac{1}{h} K\left(\frac{\mu-\mu_r}{h}\right), \quad \pi(\mu \mid y) \doteq R^{-1}\sum_{r=1}^{R} \pi(\mu \mid \sigma_r, y),$$

  where K is a kernel function with bandwidth h.

Discussion

- Update several variables at once by taking vector u_i — most useful if the components of u_i are conditionally independent given u_{-i}, which allows parallel updates.
- All the methods use the full conditionals $\pi(u_i \mid u_{-i})$: the Gibbs sampler draws from them, but the M–H algorithm only evaluates them at u and v.
- To ensure that the overall chain is ergodic we must make the chain reversible as a whole. In some cases this is obvious, but if not, and the kernels for updating different variables are $P_1, \ldots, P_m$, then we might take

$$P = P_1 \cdots P_{m-1} P_m P_{m-1} \cdots P_1, \quad \text{or} \quad P = m^{-1}\sum_{i=1}^{m} P_i, \quad \text{or} \quad \frac{1}{m!}\sum_{\xi}\prod_{i=1}^{m} P_{\xi(i)},$$

  where ξ is a random permutation of $\{1, \ldots, m\}$.
- **Convergence diagnostics** are needed to check 'stationarity' of output — simple time series plots are helpful, but more sophisticated methods exist, often based on comparing multiple chains.
- There is a huge (and still growing) literature on all aspects of these methods.

Motivation

- ☐ Many types of data have layers of variation, which must be modelled:
  - disease incidence varies between regions of a country, and within regions it may vary due to effects of poverty, pollution, ...
  - success of surgical interventions may depend on patients (age/state of health) within surgeons (different experience/skill) within hospitals (different environments/skill of nursing staff)
- ☐ We think of populations from which patients, doctors, hospitals, ... are drawn, and this suggests modelling them using layers of randomness.
- ☐ This sort of construction is very common in modelling complex data, in both classical and Bayesian frameworks.
- ☐ Some theoretical justification is provided by the notion of exchangeability.
- ☐ First, an aside on graphical representations of complex models.

Graphical models

- ☐ Complex dependencies are often represented using graphs:
  - helps understanding;
  - transforming the type of graph can simplify certain computations.
- ☐ Graph language for generic variables $Y_1, \ldots, Y_n$:
  - Y_j is represented by a **node** of the graph, so the node set is $\mathcal{J} = \{1, \ldots, n\}$;
  - we define a **neighbourhood system** $\mathcal{N} = \{\mathcal{N}_j, j \in \mathcal{J}\}$ such that
    - ▷ the **neighbours** of j are the elements of $\mathcal{N}_j \subset \mathcal{J}$, where for each j the **neighbourhood** $\mathcal{N}_j$ satisfies

$$\text{(i)} \quad j \notin \mathcal{N}_j, \qquad \text{(ii)} \quad i \in \mathcal{N}_j \quad \Leftrightarrow \quad j \in \mathcal{N}_i,$$

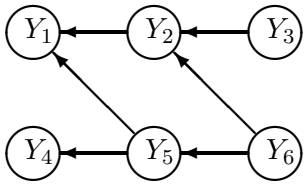
      and let $\tilde{\mathcal{N}}_j = \mathcal{N}_j \cup \{j\}$;
  - the set of nodes and the neighbourhood structure $(\mathcal{J}, \mathcal{N})$ define the graph.

Directed acyclic graphs

Definition 72 *A* **directed acyclic graph (DAG)** *is a graphical model that represents a hierarchical dependence structure:*

- ☐ *conditional dependence of* Y_1 *on* Y_2 *is represented by an arrow from the* **parent** *node* Y_2 *to the* **child** *node* Y_1*;*
- ☐ Y_1 *is a* **descendent** *of* Y_3 *if there is a chain of arrows from* Y_3 *to* Y_1*;*
- ☐ *it is* **directed** *because each arc is an arrow; and*
- ☐ *it is* **acyclic** *because it is impossible to start from a node, traverse a path by following arrows, and end up at the starting-point.*

The decomposition $f(y) = f(y_1 \mid y_2, y_5) f(y_2 \mid y_3, y_6) f(y_3) f(y_4 \mid y_5) f(y_5 \mid y_6) f(y_6)$ gives:



Conditional independence graph

- ☐ Construct a **conditional independence graph** from a DAG, by adding edges between any parents that share a child and dropping the arrowheads.
- ☐ The conditional distribution of Y_j given Y_{-j} depends only on the variables $Y_{\mathcal{N}_j}$ directly linked to Y_j in the conditional independence graph:

$$f(y_j \mid y_{-j}) = f(y_j \mid y_{\mathcal{N}_j}).$$

- ☐ Why? For any DAG,

$$f(y) = \prod_{j \in \mathcal{J}} f(y_j \mid \text{parents of } y_j)$$

so

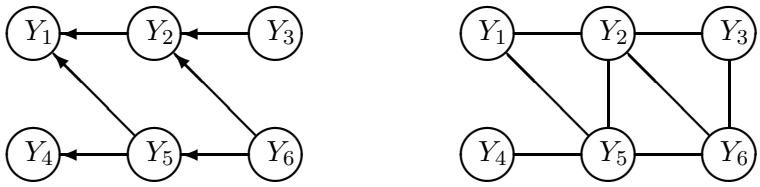
$$\begin{aligned} f(y_j \mid y_{-j}) &= \frac{f(y)}{\int f(y)\,\mathrm{d}y_j} = \frac{\prod_{i \in \mathcal{J}} f(y_i \mid \text{parents of } y_i)}{\int \prod_{i \in \mathcal{J}} f(y_i \mid \text{parents of } y_i)\,\mathrm{d}y_j} \\ &\propto f(y_j \mid \text{parents of } y_j) \prod_{\{i:\, y_i \text{ is child of } y_j\}} f(y_i \mid \text{parents of } y_i) \\ &\propto f(y_j \mid y_{\mathcal{N}_j}), \end{aligned}$$

because terms without y_j cancel from the ratio.

Simplifying full conditional distributions

- The DAG and conditional independence graph help in constructing an MCMC sampler:
  - we use the model definition to write down the DAG;
  - we convert the DAG into a conditional independence graph;
  - we read the required conditional dependencies off from the conditional independence graph.
- The conditional independence graph (right) implies that

$$f(y_1 \mid y_{-1}) = f(y_1 \mid y_2, y_5), \quad f(y_2 \mid y_{-2}) = f(y_2 \mid y_1, y_3, y_5, y_6),$$
$$f(y_3 \mid y_{-3}) = f(y_3 \mid y_2, y_6), \quad f(y_4 \mid y_{-4}) = f(y_4 \mid y_5), \quad \ldots$$



Exchangeability

Back to hierarchical models:

Definition 73 *The random variables $U_1, \ldots, U_n$ are called* **finitely exchangeable** *if their density has the property*

$$f(u_1, \ldots, u_n) = f\left(u_{\xi(1)}, \ldots, u_{\xi(n)}\right)$$

for any permutation ξ of the set $\{1, \ldots, n\}$. An infinite sequence $U_1, U_2, \ldots,$ is called **infinitely exchangeable** *if every finite subset of it is finitely exchangeable.*

- Variables are exchangeable if there is no reason to distinguish them.
- f is completely symmetric in its arguments and in probabilistic terms the $U_1, \ldots, U_n$ are indistinguishable (but not necessarily independent).

De Finetti's theorem

Theorem 74 (de Finetti) *If $U_1, U_2, \ldots$, is an infinitely exchangeable sequence of binary variables, taking values $u_j = 0, 1$, then for any n there is a distribution G such that*

$$f(u_1, \ldots, u_n) = \int_0^1 \prod_{j=1}^{n} \theta^{u_j} (1-\theta)^{1-u_j} \, G(\mathrm{d}\theta) \tag{12}$$

where

$$G(\theta) = \lim_{m\to\infty} \mathrm{P}\left\{m^{-1}(U_1 + \cdots + U_m) \leq \theta\right\}, \quad \theta = \lim_{m\to\infty} m^{-1}(U_1 + \cdots + U_m).$$

- ☐ Hence any set of exchangeable binary variables $U_1, \ldots, U_n$ that may be embedded within an infinite sequence may be modelled as if they were independent Bernoulli variables, conditional on their success probability θ, this having distribution G and being interpretable as the long-run proportion of successes.
- ☐ Similar theorems apply to other types of variables (continuous, ...).
- ☐ Thus a judgement that certain quantities are exchangeable implies that they may be represented as a random sample conditional on some θ — equivalent to using a prior distribution for θ.

Normal example

The following example illustrates properties of all hierarchical models.

Example 75 *Suppose that $v_1, \ldots, v_n$, σ^2, μ_0 and τ^2 are known and*

$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0, \tau^2), \\ \theta_1, \ldots, \theta_n \mid \mu &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \\ y_j \mid \theta_j &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_j, v_j), \quad j = 1, \ldots, n. \end{aligned}$$

Here the **hyperparameters** *μ_0 and τ^2 control the uncertainty at the top level of the hierarchy. Give the DAG and conditional independence graph for this model, and show that*

$$\mathrm{E}(\mu \mid y) = \frac{\mu_0/\tau^2 + \sum y_j/(\sigma^2 + v_j)}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)}, \quad \mathrm{var}(\mu \mid y) = \frac{1}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)},$$

$$\mathrm{E}(\theta_j \mid y) = \frac{\sigma^2 y_j + v_j \mathrm{E}(\mu \mid y)}{\sigma^2 + v_j}, \quad \mathrm{var}(\theta_j \mid y) = \frac{1 + \mathrm{var}(\mu \mid y)/\sigma^2}{1/v_j + 1/\sigma^2}.$$

Discuss.

Note to Example 75

☐ The y_j have different variances, but their means θ_j are supposed indistinguishable and hence are modelled as exchangeable, being normal with unknown mean μ, and we can write

$$y_j = \mu_0 + (\mu - \mu_0) + (\theta_j - \mu) + (y_j - \theta_j),$$

where μ_0 is known, and as the y_j and θ_j are linear combinations of normal variables it is straightforward to check that

$$\begin{pmatrix}\mu\\ \theta\\ y\end{pmatrix} \sim \mathcal{N}_{2n+1}\left\{\mu_0 1_{2n+1}, \quad \begin{pmatrix}\tau^2 & \tau^2 1_n^{\mathrm{T}} & \tau^2 1_n^{\mathrm{T}}\\ \tau^2 1_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n\\ \tau^2 1_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n & V + \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n\end{pmatrix}\right\}, \tag{13}$$

where 1_n denotes the $n \times 1$ vector of ones and $V = \mathrm{diag}(v_1, \ldots, v_n)$.

☐ The most direct approach to computing the posterior distributions μ and θ given y is to write

$$\begin{pmatrix}\mu\\ \theta\\ y\end{pmatrix} \sim \mathcal{N}_{2n+1}\left\{\mu_0 1_{2n+1}, \quad \begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{pmatrix}\right\},$$

where $\mathrm{var}(y) = \Omega_{22}$. Then the posterior density of the parameters given y is also normal, with

$$\begin{pmatrix}\mu\\ \theta\end{pmatrix} \mid y \sim \mathcal{N}_{n+1}\left\{\mu_0 1_{n+1} + \Omega_{12}\Omega_{22}^{-1}(y - \mu_0 1_n), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right\}. \tag{14}$$

We shall take a less messy and maybe more enlightening route,, first computing the posterior distribution of μ, then that of θ given both μ and y, and then marginalising the latter over μ.

☐ Expression (13) shows that the joint density of μ and y is normal with covariance matrix

$$\begin{pmatrix}A & B^{\mathrm{T}}\\ B & C\end{pmatrix}, \quad A = \tau^2, \quad B = \tau^2 1_n, \quad C = \tau^2 1_n 1_n^{\mathrm{T}} + D, \quad D = \mathrm{diag}(\sigma^2 + v_1, \ldots, \sigma^2 + v_n).$$

The Woodbury formula gives

$$(D + \tau^2 1_n 1_n^{\mathrm{T}})^{-1} = D^{-1} - D^{-1} 1_n (\tau^{-2} + 1_n^{\mathrm{T}} D^{-1} 1_n)^{-1} 1_n D^{-1}$$

so with $a = 1_n^{\mathrm{T}} D^{-1} 1_n$ we have

$$\begin{aligned} A - BC^{-1}B^{\mathrm{T}} &= \tau^2 - \tau^2 1_n^{\mathrm{T}}\{D^{-1} - D^{-1}1_n(\tau^{-2} + 1_n^{\mathrm{T}}D^{-1}1_n)^{-1}1_n D^{-1}\}\tau^2 1_n \\ &= \tau^2 - \tau^4\left\{a - \frac{a^2}{\tau^{-2} + a}\right\} \\ &= (\tau^{-2} + a)^{-1}, \end{aligned}$$

which gives $\mathrm{var}(\mu \mid y)$, and a simpler calculation using (14) with μ only gives the mean, resulting in

$$\mathrm{E}(\mu \mid y) = \frac{\mu_0/\tau^2 + \sum y_j/(\sigma^2 + v_j)}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)}, \quad \mathrm{var}(\mu \mid y) = \frac{1}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)}.$$

The posterior mean of μ is a weighted average of its prior mean μ_0 and of the y_j, weighted according to their precisions. Typically τ^2 is taken to be very large, and then $\mathrm{E}(\mu \mid y)$ is essentially a weighted average of the data. Even when $v_j \to 0$ for all j there is still posterior uncertainty about μ, whose variance is σ^2/n because $y_1, \ldots, y_n$ is then a random sample from $N(\mu, \sigma^2)$.

Note 2 to Example 75

- To compute the posterior mean and variance of θ_j we note that the graph structure gives $f(\theta_j \mid \mu, y) = f(\theta_j \mid \mu, y_j)$. This simplifies the computation because we need only compute the joint distribution of (μ, θ_j, y_j), and this is

$$\mathcal{N}_3 \left\{ 1_3\mu_0, \begin{pmatrix} \tau^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2+\sigma^2 & \tau^2+\sigma^2 \\ \tau^2 & \tau^2+\sigma^2 & \tau^2+\sigma^2+v_j \end{pmatrix} \right\}$$

  from which we obtain $\theta_j \mid \mu, y_j \sim \mathcal{N}\{(y_j/v_j + \mu/\sigma^2)/(1/v_j + 1/\sigma^2), (1/v_j + 1/\sigma^2)^{-1}\}$. As

$$\mathrm{E}(\theta_j \mid y) = \mathrm{E}\left\{\mathrm{E}(\theta_j \mid \mu, y_j)\right\}, \quad \mathrm{var}(\theta_j \mid y) = \mathrm{E}\left\{\mathrm{var}(\theta_j \mid \mu, y_j)\right\} + \mathrm{var}\left\{\mathrm{E}(\theta_j \mid \mu, y_j)\right\},$$

  where the outer expectation and variance are over the distribution of μ given y, we finally obtain

$$\mathrm{E}(\theta_j \mid y) = \frac{\sigma^2 y_j + v_j \mathrm{E}(\mu \mid y)}{\sigma^2 + v_j}, \quad \mathrm{var}(\theta_j \mid y) = \frac{1 + \mathrm{var}(\mu \mid y)/\sigma^2}{1/v_j + 1/\sigma^2}.$$

- The posterior mean of θ_j is a weighted average of y_j and $\mathrm{E}(\mu \mid y)$, showing shrinkage of y_j towards $\mathrm{E}(\mu \mid y)$ by an amount that depends on v_j. As $v_j \to 0$, $\mathrm{E}(\theta_j \mid y) \to y_j$, while as $v_j \to \infty$, $\mathrm{E}(\theta_j \mid y) \to \mathrm{E}(\mu \mid y)$. This is a characteristic feature of hierarchical models, in which there is a 'borrowing of strength' whereby all the data combine to estimate common parameters such as μ, while estimates of individual parameters such as the θ_j are shrunk towards common values by amounts that depend on the precisions v_j of the corresponding observations.

Example: Cardiac surgery data

A	0/47	B	18/148	C	8/119	D	46/810	E	8/211	F	13/196
G	9/148	H	31/215	I	14/207	J	8/97	K	29/256	L	24/360

Mortality rates r/m from cardiac surgery in 12 hospitals (numbers of deaths r out of m operations).

- Hierarchical model:

$$r_j \mid \theta_j \overset{\text{ind}}{\sim} B(m_j, \theta_j), \quad j = A, \ldots, L, \quad \theta_A, \ldots, \theta_L \mid \zeta \overset{\text{iid}}{\sim} f(\theta \mid \zeta), \quad \zeta \sim \pi(\zeta).$$

  Conditional on θ_j, the number of deaths r_j at hospital j is binomial with probability θ_j and denominator m_j, the number of operations, which plays the same role as v_j^{-1} in the normal example above: when m_j is large then a death rate is relatively precisely known.
- Conditional on ζ, the θ_j are a random sample from a distribution $f(\theta \mid \zeta)$, and the prior distribution for ζ depends on fixed hyperparameters.
- We take $\beta_j = \log\{\theta_j/(1-\theta_j)\} \sim N(\mu, \sigma^2)$, conditional on $\zeta = (\mu, \sigma^2)$, and $\mu \sim N(0, c^2)$ and $\sigma^2 \sim IG(a, b)$, with $a = b = 10^{-3}$, so σ^2 has prior mean one but variance 10^3, and $c = 10^3$, giving μ prior variance 10^6.

Example: Cardiac surgery data

- The joint density is

$$\left[\prod_j \binom{m_j}{r_j} \frac{e^{r_j\beta_j}}{\left(1+e^{\beta_j}\right)^{m_j}} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta_j-\mu)^2\right\}\right] \times \pi(\mu)\pi(\sigma^2),$$

  so the full conditional densities for μ and σ^2 are normal and inverse gamma.

- We use a Metropolis–Hastings step for β, using a random walk proposal with

$$\beta_j' \sim \mathcal{N}\{\beta_j, d^2\sigma^2 v_j/(\sigma^2+v_j)\}, \quad v_j = \frac{m_j+1}{(r_j+1/2)(m_j-r_j+1/2)},$$

  where we choose d to optimise the algorithm.

- This normal approximation comes from Example 75, taking

$$\widehat{\beta}_j \mid \beta \sim \mathcal{N}(\beta, v_j), \quad \beta_j \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

  and then computing $\beta_j \mid \widehat{\beta}_j$.

Example: Cardiac surgery data

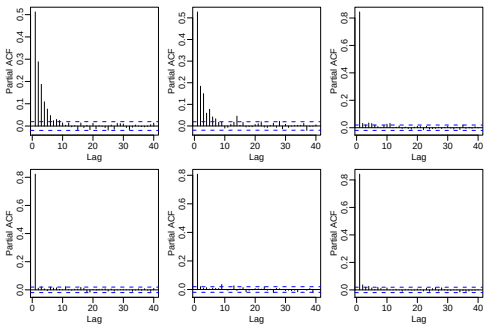
```
cardiac.gibbs <- function(data, mu0=0, a=10^(-3), b=10^(-3), c=10^3, R=10^5, d=1)
{ # parameter is mu, sig2, beta
  card.update <- function(data, mu0, a, b, c, para)
  {
    sig2 <- para[2]
    beta <- para[-c(1,2)]
    n <- length(beta)
    mu <- rnorm( 1, (mu0/c^2 + sum(beta)/sig2)/(1/c^2+n/sig2),sqrt(1/(1/c^2+n/sig2)) )
    sig2 <- rigamma( a+n/2, b+0.5*sum((beta-mu)^2) )
    v <- (data$m+1)/((data$r+0.5)*(data$m-data$r+0.5))
    var.beta <- sig2*v/(v+sig2)
    beta.prop <- rnorm(n, beta, sd=d*sqrt(var.beta))
    acc.prob <- exp( data$r*beta.prop - data$m*log(1+exp(beta.prop)) -
                                  0.5*(beta.prop-mu)^2/sig2 - data$r*beta +
                                  data$m*log(1+exp(beta)) + 0.5*(beta-mu)^2/sig2 )
    acc.prob <- pmin(1,acc.prob)   # use pmin and ifelse to do all
    beta <- ifelse(runif(n)<=acc.prob,beta.prop, beta) # acceptances/rejections at once
    c( mu, sig2, beta)
  }
  rigamma <- function(a, b) 1/rgamma(1, shape=a, rate=b)
  logit <- function(p) log(p/(1-p))
  out <- matrix(NA, 2+nrow(data), R)
  out[, 1] <- c(0, 1, rep(0,nrow(data)))
  for(r in 2:R)
   out[, r] <- card.update(data, mu0, a, b, c, out[,r-1])
  out
}

system.time( cardiac.sim <- cardiac.gibbs(cardiac, R=10^4, d=4) )  # around 3.5 seconds
acc.rate <- function(x) mean((diff(x)!=0))
apply(cardiac.sim,1,acc.rate)   # compute acceptance rates for the proposals
```

Effect of d

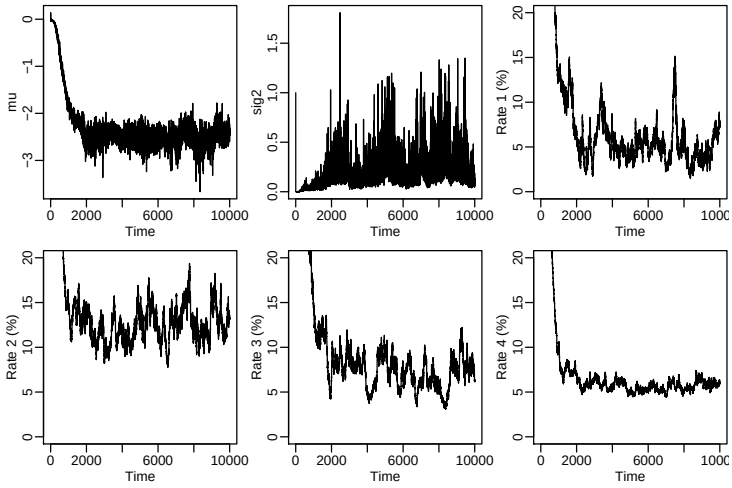
Acceptance probabilities for different values of d:

d	0.1	0.5	1	2	3	5	10	20	30
μ	1	1	1	1	1	1	1	1	1
σ^2	1	1	1	1	1	1	1	1	1
β	0.95	0.82	0.7	0.5	0.37	0.25	0.12	0.06	0.05

PACF for $d = 1$:

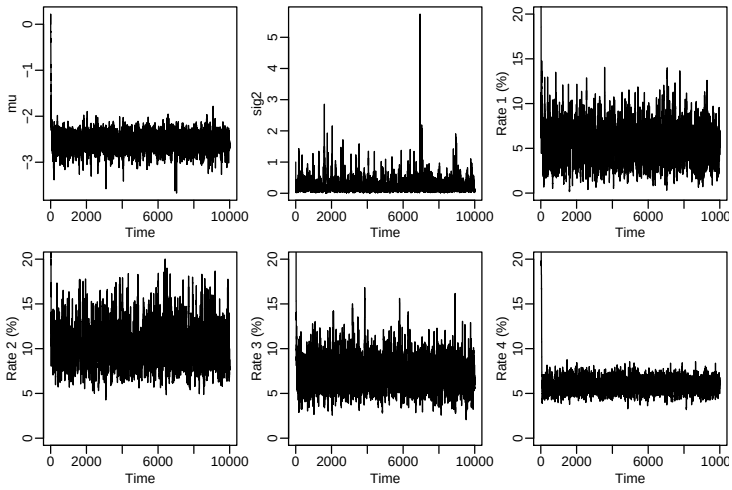


Effect of d: $d = 0.1$



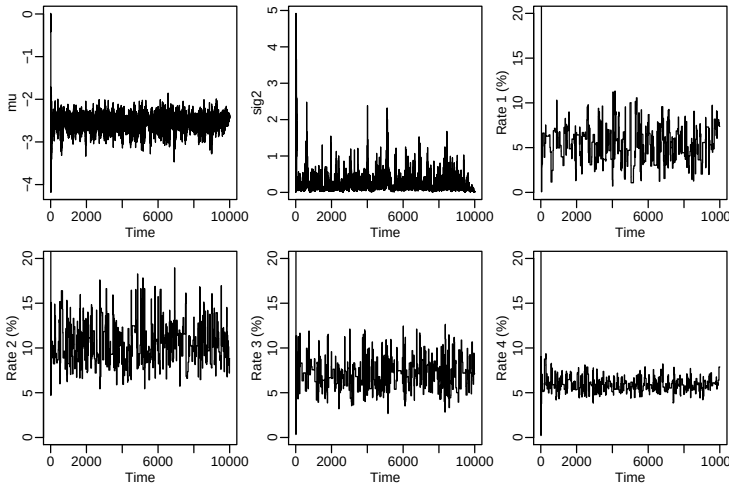
Taking $d = 0.1$ makes the acceptance probability too high, so the chain mixes too slowly.

Effect of d: $d = 1$



Taking $d = 1$ is OK, but theory suggest that the acceptance rate should be around 0.2–0.4, so taking $d \approx 4$ seems somewhat better.

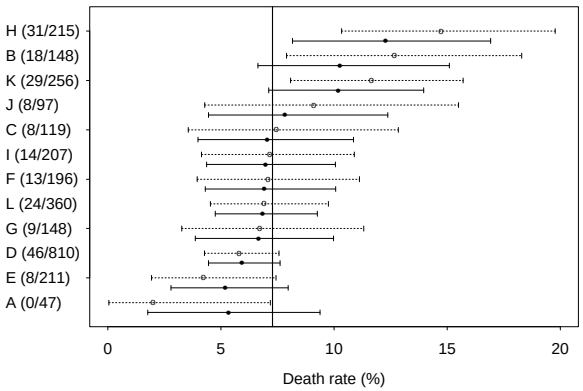
Effect of d: $d = 30$



Taking $d = 30$ makes the acceptance probability too low, so the chain sticks.

Example: Cardiac surgery data, effect of shrinkage

Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical model. Note the shrinkage ('borrowing of strength') of the estimates for the hierarchical model towards the overall posterior mean rate, shown as the solid vertical line; the hierarchical intervals are slightly shorter than those for the simpler model.



Summary

- ☐ Graphical representation of dependence relations very useful.
- ☐ Hierarchical modelling allows us to fit complex models to data
- ☐ Key idea is to treat parameters as coming from a distribution, and to use the data to estimate the distribution
  - Appropriate when exchangeable elements are present
  - Not appropriate when we are interested in certain pre-specified parameters or where prior knowledge distinguishes them
  - Example of inappropriate use: economic modelling with countries of Europe treated as exchangeable
- ☐ Can be hard to count the number of parameters: the prior 'ties together' some parameters, so there are 'really' fewer—but how many?