Statistical Inference

Anthony Davison

http://stat.epfl.ch

1 Introduction slide 2

1.1 Background slide 3

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 reach a **decision** to guide future actions.
- ☐ 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**.

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?
- ☐ For now we suppose these questions have satisfactory answers . . .

Some statistical activities

- Conventionally divided into
  - **design of investigations** — how do we get reliable data to answer a question efficiently and securely?
  - **descriptive statistics/exploratory data analysis** — how can we get insight into a specific dataset?
  - **inference** — what can we learn about the properties of a 'population' underlying the data?
  - **decision analysis** — what is the optimal decision in a given situation?

  to which we nowadays add

  - **machine learning** — algorithms, generally complex and computationally demanding, often used for prediction/decision-making.

Descriptive statistics

- In principle concerns **only the data available**, mainly involving
  - **graphical summaries** — histograms, boxplots, scatterplots, ...
  - **numerical summaries** — averages, variances, medians, ...
- Some summaries presuppose the existence of 'population' quantities (e.g., a density).
- We use probability models to analyse the properties of these summaries (e.g., formulation of a boxplot, 'is that difference significant?', ...).
- Even when we have 'all the data' (e.g., loyalty card transactions) we may want to ask 'what if?' questions, and these require further assumptions (e.g., temporal stability, future and current customers are similar, ...).

Statistical inference

- Use observed data to draw conclusions about a 'population' from which the data are assumed to be drawn, or about future data.
- The 'population' and observed data are linked by concepts of probability.
- Two distinct roles of probability in statistical analysis:
  - as a description of **variation** in data ('aleatory probability', 'chance'), treating the observed data y as an outcome of a random process/probability model, perhaps
    - ▷ suggested by the context, or
    - ▷ imposed by the investigator (via some sampling procedure);
  - to formulate **uncertainty** ('epistemic probability') about the reality modelled in terms of the random experiment, based on y.
- Most of the course concerns the formulation and expression of uncertainty.
- We first revise some concepts from probability and basic statistics.

Probability spaces

- ☐ Ordered triples $(\Omega, \mathcal{F}, \mathrm{P})$ consisting of
  - a set Ω of **elementary outcomes** ω corresponding to distinct potential outcomes of a random experiment;
  - an **event space** $\mathcal{F}$ of subsets of Ω that satisfy (a) $\Omega \in \mathcal{F}$, (b) if $\mathcal{A} \in \mathcal{F}$, then $\mathcal{A}^c \in \mathcal{F}$, and (c) if $\mathcal{A}_1, \mathcal{A}_2, \ldots \in \mathcal{F}$, then $\bigcup A_j \in \mathcal{F}$;
  - a **probability measure** $\mathrm{P} : \mathcal{F} \to [0,1]$ that satisfies (i) if $\mathcal{A} \in \mathcal{F}$, then $0 \leq \mathrm{P}(\mathcal{A}) \leq 1$, (ii) $\mathrm{P}(\Omega) = 1$, (iii) if $\mathcal{A}_1, \mathcal{A}_2, \ldots \in \mathcal{F}$ satisfy $\mathcal{A}_j \cap \mathcal{A}_k = \emptyset$ for $j \neq k$, then $\mathrm{P}(\bigcup \mathcal{A}_j) = \sum \mathrm{P}(\mathcal{A}_j)$.
- ☐ We call $(\Omega, \mathcal{F})$ a **measure space** and any $\mathcal{A} \in \mathcal{F}$ an **event (measurable set)**.
- ☐ From these we deduce
  - the **inclusion-exclusion formulae**, and
  - computation of probabilities in simple problems using **combinatorial formulae**.
- ☐ If $\mathrm{P}(\mathcal{B}) > 0$ we define **conditional probabilities** $\mathrm{P}(\mathcal{A} \mid \mathcal{B}) = \mathrm{P}(\mathcal{A} \cap \mathcal{B})/\mathrm{P}(\mathcal{B})$, and derive
  - a new **conditional probability distribution** $\mathrm{P}_{\mathcal{B}}(\mathcal{A}) = \mathrm{P}(\mathcal{A} \mid \mathcal{B})$ for $\mathcal{A} \in \mathcal{F}$,
  - the **law of total probability**,
  - **Bayes' theorem**, and
  - the notion of **independent events**, for which $\mathrm{P}(\mathcal{A} \cap \mathcal{B}) = \mathrm{P}(\mathcal{A})\mathrm{P}(\mathcal{B})$.

Random variables

- ☐ Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $(\mathcal{X}, \mathcal{G})$ a measurable space. A **random function** X from Ω into $\mathcal{X}$ has the property that $X^{-1}(\mathcal{C}) = \{\omega : X(\omega) \in \mathcal{C}\} \in \mathcal{F}$ for any $\mathcal{C} \in \mathcal{G}$, so $\mathrm{P}(X \in \mathcal{C}) = \mathrm{P}\{X^{-1}(\mathcal{C})\}$ is well-defined. Such a function is called **measurable**.
- ☐ If $\mathcal{X} = \mathbb{R}$ or $\mathbb{R}^n$ we call X a **random variable** and there exists a **cumulative distribution function (CDF)** F such that $\mathrm{P}\{X \in (-\infty, x_1] \times \cdots \times (-\infty, x_n]\} = F(x_1, \ldots, x_n)$.
- ☐ A CDF increases from 0 when any of its arguments increases from $-\infty$ to $+\infty$.
- ☐ F can be written as a sum of (sub-)distributions $F_{\mathrm{ac}} + F_{\mathrm{dis}} + F_{\mathrm{sing}}$, where
  - F_{ac} is absolutely continuous, i.e., there exists a non-negative **probability density function (PDF)** $f_{\mathrm{ac}}(x) = \mathrm{d}F_{\mathrm{ac}}(x)/\mathrm{d}x$,
  - F_{dis} is discrete, i.e., its **probability mass function (PMF)** $f_{\mathrm{dis}}(x)$ is positive only on a finite or countable set $\mathcal{S}$, and
  - F_{sing} is singular, and can be ignored (look up 'Cantor distribution' if interested).
- ☐ We call X **continuous** or **discrete** respectively if F_{dis} or F_{ac} is absent.
- ☐ If necessary we use **Lebesgue–Stieltjes integration**, whereby

$$\mathrm{P}(X \in \mathcal{C}) = \int_{\mathcal{C}} \mathrm{d}F(x) = \int_{\mathcal{C}} f_{\mathrm{ac}}(x)\,\mathrm{d}x + \sum_{x \in \mathcal{C} \cap \mathcal{S}} f_{\mathrm{dis}}(x), \quad \mathcal{C} \subset \mathcal{X};$$

the notation $\int_a^b$ is unwise because it doesn't distinguish $\mathcal{C} = [a, b]$ from $\mathcal{C} = (a, b)$.

New distributions and random variables

- We define the **conditional distribution** of X given an event $\mathcal{B} \in \mathcal{F}$ by

$$\mathrm{P}(X \in \mathcal{A} \mid \mathcal{B}) = \mathrm{P}(\{X \in \mathcal{A}\} \cap \mathcal{B})/\mathrm{P}(\mathcal{B}).$$

- If $Y = g(X) \in \mathcal{Y}$ and we write $g^{-1}(\mathcal{B}) = \{x : g(x) \in \mathcal{B}\}$ for $\mathcal{B} \subset \mathcal{Y}$, then

$$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\}.$$

- If X is continuous and $Y = g(X)$ with g a smooth bijection, then (in obvious notation)

$$f_Y(y) = f_X\{g^{-1}(y)\} \left| \frac{\partial g^{-1}(y)}{\partial y} \right|,$$

  where the last term is the Jacobian of the transformation.
- If $X = (X_1, X_2)$ is continuous, we obtain **marginal** and **conditional** densities

$$f_{X_2}(x_2) = \int f_{X_1,X_2}(x_1, x_2)\,\mathrm{d}x_1, \quad f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)},$$

  with corresponding formulae in the discrete and mixed cases.
- X_1 and X_2 are **independent ($X_1 \perp\!\!\!\perp X_2$)** iff $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$, $\forall x_1, x_2$.

Moments

- The **expectation** $\mathrm{E}\{g(X)\}$ of $g(X)$ is defined if $\mathrm{E}\{|g(X)|\} < \infty$ as

$$\mathrm{E}\{g(X)\} = \int_{\mathcal{X}} g(x)\,\mathrm{d}F(x).$$

- For scalar X we define **moments** $\mathrm{E}(X^r)$, **mean** $\mu = \mathrm{E}(X)$ and **variance**

$$\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2.$$

- $\mathrm{var}(X) = 0$ iff X is constant with probability one.
- For vector X we define the **mean vector** and **(co)variance matrix**

$$\mu = \mathrm{E}(X), \quad \mathrm{cov}(X_1, X_2) = \mathrm{E}(X_1 X_2^{\mathrm{T}}) - \mathrm{E}(X_1)\mathrm{E}(X_2)^{\mathrm{T}},$$

  and write $\mathrm{var}(X) = \mathrm{cov}(X, X) = E\{(X - \mu)(X - \mu)^{\mathrm{T}}\}$.
- The **correlation**, $\mathrm{corr}(X_1, X_2) = \mathrm{cov}(X_1, X_2)/\{\mathrm{var}(X_1)\mathrm{var}(X_2)\}^{1/2}$, is a measure of dependence between variables that does not depend on their units of measurement.
- Expectation $\mathrm{E}(\cdot)$ is a linear operator, so it is easy to check that

$$\mathrm{E}(a + BX) = a + B\mathrm{E}(X), \quad \mathrm{cov}(a + BX, c + DX) = B\mathrm{var}(X)D^{\mathrm{T}}.$$

Conditional moments

- The **conditional expectation** of $g(X, Y)$ given $X = x$ is

$$\mathrm{E}\{g(X,Y) \mid X = x\} = \int_{\mathcal{Y}} g(x,y)\,\mathrm{d}F(y \mid x),$$

  which in the continuous and discrete cases equals

$$\int_{\mathcal{Y}} g(x,y) f_{Y|X}(y \mid x)\,\mathrm{d}y, \quad \sum_{y \in \mathcal{Y}} g(x,y) f_{Y|X}(y \mid x),$$

  and other conditional moments are defined likewise.
- This is a function of x, so it defines a random variable $\tilde{g}(X) = \mathrm{E}\{g(X,Y) \mid X\}$.
- The **law of total expectation (tower property)** gives

$$\begin{aligned} \mathrm{E}\{g(X,Y)\} &= \mathrm{E}_X\left[\mathrm{E}\{g(X,Y) \mid X = x\}\right], \\ \mathrm{var}\{g(X,Y)\} &= \mathrm{E}_X\left[\mathrm{var}\{g(X,Y) \mid X = x\}\right] + \mathrm{var}_X\left[E\{g(X,Y) \mid X = x\}\right], \end{aligned}$$

  where E_X denotes expectation with respect to the marginal distribution of X, etc., with a similar expression (which you should give) for $\mathrm{cov}\{g(X,Y), h(X,Y)\}$.
- We ignore mathematical issues arising from conditioning on events of probability zero — look up 'Borel–Kolmogorov paradox' if interested.

Terminology and notation

- PDFs and PMFs are not the same but we henceforth use the term **density** for both.
- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ means that the X_j are independent and all have density f, and we then 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 $X_j \sim f_j$.
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} (\mu, \sigma^2)$ means that the X_j are independent with mean μ and variance σ^2 (with $0 < \sigma^2 < \infty$). The X_j need not be normal or have the same distribution.
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} (\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2)$ means that the X_j are independent with means μ_j and variances σ_j^2 (where $0 < \sigma_j^2 < \infty$).
- The **p quantile** of the distribution F of a scalar random variable X is

$$x_p = \inf\{x : F(x) \geq p\}, \quad 0 < p < 1.$$

  Usually $x_p = F^{-1}(p)$ for continuous X, but not for discrete (or mixed) X.
- A **standard normal** variable $Z \sim \mathcal{N}(0,1)$ has PDF and CDF

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

  and p quantile $z_p = \Phi^{-1}(p)$, so $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$ has p quantile $\mu + \sigma z_p$.

Inequalities

- A real-valued **convex function** g defined on a vector space $\mathcal{V}$ has the property that for any $x, y \in \mathcal{V}$,

$$g\{tx + (1-t)y\} \leq tg(x) + (1-t)g(y), \quad 0 \leq t \leq 1.$$

  Equivalently, for all $y \in \mathcal{V}$, there exists a vector $b(y)$ such that

$$g(x) \geq g(y) + b(y)^{\mathrm{T}}(x - y)$$

  for all x. If $g(x)$ is differentiable, then we can take $b(y) = g'(y)$.

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

$$\begin{aligned} \mathrm{P}\{h(X) \geq a\} &\leq \mathrm{E}\{h(X)\}/a, \quad \text{(basic inequality)} \\ \mathrm{P}(|X| \geq a) &\leq \mathrm{E}(|X|)/a, \quad \text{(Markov's inequality)} \\ \mathrm{P}(|X| \geq a) &\leq \mathrm{E}(X^2)/a^2, \quad \text{(Chebyshov's inequality)} \\ \mathrm{E}\{g(X)\} &\geq g\{\mathrm{E}(X)\}. \quad \text{(Jensen's inequality)} \end{aligned}$$

- On replacing X by $X - \mathrm{E}(X)$, Chebyshov's inequality gives

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\} \leq \mathrm{var}(X)/a^2.$$

Note: Inequalities

(a) Let $Y = h(X)$. If $y \geq 0$, then for any $a > 0$, $y \geq yI(y \geq a) \geq aI(y \geq a)$. Therefore

$$\mathrm{E}\{h(X)\} = \mathrm{E}(Y) \geq \mathrm{E}\{YI(Y \geq a)\} \geq \mathrm{E}\{aI(Y \geq a)\} = a\mathrm{P}(Y \geq a) = a\mathrm{P}\{h(X) \geq a\},$$

and division by $a > 0$ gives the result.
(b) Note that $h(x) = |x|$ is a non-negative function on $\mathbb{R}$, and apply (a).
(c) Note that $h(x) = x^2$ is a non-negative function on $\mathbb{R}$, and that $\mathrm{P}(X^2 \geq a^2) = \mathrm{P}(|X| \geq a)$.
(d) A convex function has the property that, for all y, there exists a value $b(y)$ such that $g(x) \geq g(y) + b(y)(x - y)$ for all x. If $g(x)$ is differentiable, then we can take $b(y) = g'(y)$. (Draw a graph if need be.) To prove this result, we take $y = \mathrm{E}(X)$, and then have

$$g(X) \geq g\{\mathrm{E}(X)\} + b\{\mathrm{E}(X)\}\{X - \mathrm{E}(X)\},$$

and taking expectations of this gives $\mathrm{E}\{g(X)\} \geq g\{\mathrm{E}(X)\}$.

MGFs and KGFs

- The **moment-generating function (MGF)** and **cumulant-generating function (KGF)** of a scalar random variable X are

$$M_X(t) = \mathrm{E}\left(e^{tX}\right), \quad K_X(t) = \log M_X(t), \quad t \in \mathcal{N} = \{t : M_X(t) < \infty\}.$$

- $\mathcal{N}$ is non-empty, because $M_X(0) = 1$, but the MGF and KGF are non-trivial only if $\mathcal{N}$ contains an open neighbourhood of the origin, since then

$$M_X(t) = \mathrm{E}\left(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mathrm{E}(X^r), \quad K_X(t) = \sum_{r=1}^{\infty} \frac{t^r}{r!}\kappa_r,$$

  and one can obtain the **moments** $\mathrm{E}(X^r)$ and **cumulants** κ_r by differentiation.
- In the vector case we define

$$M_X(t) = \mathrm{E}\left(e^{t^{\mathrm{T}}X}\right), \quad K_X(t) = \log M_X(t),$$

  and differentiation with respect to the elements of $t = (t_1, \ldots, t_n)^{\mathrm{T}}$ gives the mean vector and covariance matrix of X.
- There is a 1–1 mapping between MGFs/KGFs and distributions.
- KGFs for linear combinations are computed as $K_{a+BX}(t) = a^{\mathrm{T}}t + K_X(B^{\mathrm{T}}t)$.

Note: Moments and cumulants

- We consider scalar X, as the calculations for vector X are analogous.
- First note that $M_X(t) = 1$ when $t = 0$, since $\mathrm{E}(e^{tX}) = \mathrm{E}(1) = 1$; thus $0 \in \mathcal{N}$ for any X.
- If $\mathcal{N}$ contains an open set $(-a, a)$ for some $a > 0$, and $\mu_r = \mathrm{E}(X^r)$ denotes the rth moment of X, then if $|t| < a$,

$$K_X(t) = \sum_{r=1}^{\infty} \frac{t^r \kappa_r}{r!} = \log M_X(t) = \log\left(\sum_{r=0}^{\infty} \frac{t^r \mu_r}{r!}\right) = \log(1+b) = b - b^2/2 + b^3/3 + \cdots,$$

  where $b = t\mu_1 + t^2\mu_2/2! + t^3\mu_3/3! + \cdots$. If we expand and compare coefficients of $t, t^2, t^3, \ldots$ in the two expansions we get

$$\kappa_1 = \mu_1, \quad \kappa_2 = \mu_2 - \mu_1^2, \quad \kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3, \quad \kappa_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4, \quad \ldots,$$

  so $\kappa_1 = \mathrm{E}(X)$, $\kappa_2 = \mathrm{var}(X)$, $\kappa_3 = \mathrm{E}\{(X - \mu_1)^3\}$, $\ldots$

Exponential tilting

- A baseline density f_0 with a non-trivial MGF can be used to construct a family of densities by **exponential tilting**, i.e.,

$$f(x;t) = f_0(x)\exp\left\{t^{\mathrm{T}}x - K_X(t)\right\}, \quad t \in \mathcal{N},$$

  where

$$\mathcal{N} = \{t : K_X(t) < \infty\}$$

  and individual members of the family are determined by the value of t.
- Hölder's inequality gives

$$M_X\{\alpha t_1 + (1-\alpha)t_2\} \le M_X(t_1)^{\alpha} M_X(t_2)^{1-\alpha} < \infty, \quad 0 \le \alpha \le 1,$$

  for any $t_1, t_2 \in \mathcal{N}$, so the set $\mathcal{N}$ and the function $K_X(t)$ are both convex.
- This implies that $f(x;t)$ is log-concave in t, which is very useful for statistics.
- Slightly generalized, this construction leads to an elegant general theory that puts many well-known distributions (Poisson, binomial, normal, ...) under the same roof.

Exponential family models

- If $\theta \in \Theta \subset \mathbb{R}^d$, where $\dim\Theta = d$, and there exists a $d \times 1$ function $s = s(y)$ of data y and a **parametrisation** (i.e., a 1–1 function) $\varphi \equiv \varphi(\theta)$ such that

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

  then this is an (d,d) **exponential family** of distributions, with
  - **canonical statistic** $S = s(Y)$,
  - **canonical parameter** φ,
  - **cumulant generator** k, which is convex on $\mathcal{N} = \{\varphi : k(\varphi) < \infty\}$, and
  - **mean parameter** $\mu \equiv \mu(\varphi) = \mathrm{E}(S;\varphi) = \nabla k(\varphi)$, where $\nabla\cdot = \partial\cdot/\partial\varphi$.
- We suppose that there is no vector a such that $a^{\mathrm{T}}S$ is constant, and call the model a **minimal representation** if there is no vector a such that $a^{\mathrm{T}}\varphi$ is constant.
- The cumulant-generating function for S is

$$K_S(t) = \log M_S(t) = k(\varphi + t) - k(\varphi), \quad t \in \mathcal{N}' \subset \mathbb{R}^d,$$

  where $0 \in \mathcal{N}'$. On writing $\nabla^2\cdot = \partial^2\cdot/\partial\varphi\partial\varphi^{\mathrm{T}}$, one can check that

$$\mathrm{E}(S) = \nabla k(\varphi), \quad \mathrm{var}(S) = \nabla^2 k(\varphi).$$

Note: Cumulant-generating functions

☐ The MGF 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,

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

which yields $K_S(t) = k(\varphi + t) - k(\varphi)$.

☐ Now $M_S(0) = 1$, $K_S(0) = 0$, $\partial K_S(t)/\partial t = \nabla k(\varphi + t)$ and $\partial^2 K_S(t)/\partial t \partial t^{\mathrm{T}} = \nabla^2 k(\varphi + t)$, so

$$\mathrm{E}(S) = \partial M_S(t)/\partial t|_{t=0} = \partial e^{K_S(t)}/\partial t\Big|_{t=0} = \partial K_S(t)/\partial t\, e^{K_S(t)}\Big|_{t=0} = \nabla k(\varphi).$$

A similar calculation for the variance gives

$$\mathrm{E}(SS^{\mathrm{T}}) = \partial^2 M_S(t)/\partial t \partial t^{\mathrm{T}}\big|_{t=0} = \nabla^2 k(\varphi) + \nabla k(\varphi)\nabla k(\varphi)^{\mathrm{T}},$$

and thus

$$\mathrm{var}(S) = \mathrm{E}(SS^{\mathrm{T}}) - \mathrm{E}(S)\mathrm{E}(S)^{\mathrm{T}} = \nabla^2 k(\varphi) + \nabla k(\varphi)\nabla k(\varphi)^{\mathrm{T}} - \nabla k(\varphi)\nabla k(\varphi)^{\mathrm{T}} = \nabla^2 k(\varphi).$$

Examples

Example 1 (Poisson sample) *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{Poiss}(\theta)$, *find the corresponding exponential family.*

Example 2 (Satellite conjunction) *A simple model for the position* Y *of a satellite in* $\mathbb{R}^2$ *relative to the origin is*

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

where $d_1, d_2 > 0$ *are known and* $\psi > 0$, $0 < \lambda \leq 2\pi$. *Write the corresponding density*

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

as an exponential family.

☐ NB: avoid confusion — exponential family ≠ exponential distribution!

☐ The exponential distribution is just one example of an exponential family.

Note to Example 1

Independent Poisson $Y_1, \ldots, Y_n$ have joint density

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

where $m(y) = (\prod y_j)^{-1}$. This is a $(1,1)$ exponential family with

- ☐ canonical statistic $s = s(y) = \sum y_j$,
- ☐ canonical parameter $\log\theta = \varphi \in \mathcal{N} = \mathbb{R}$,
- ☐ cumulant generator $k(\varphi) = n\theta = ne^{\varphi}$ and
- ☐ mean parameter $\mu = \nabla k(\varphi) = ne^{\varphi} = n\theta = \mathrm{E}(S)$.

Two standard parametrizations use the real parameter φ or the mean $\mu = ne^{\varphi} \in \mathbb{R}_+$.

Note to Example 2

- ☐ 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 if Ω is known then the exponent can be written as

$$-\tfrac{1}{2}\log\{(2\pi)^n|\Omega|\} - \tfrac{1}{2}y^{\mathrm{T}}\Omega^{-1}y + y^{\mathrm{T}}\Omega^{-1}\mu - \tfrac{1}{2}\mu^{\mathrm{T}}\Omega^{-1}\mu = \log m(y) + s(y)^{\mathrm{T}}\varphi - k(\varphi),$$

  where $s(y) = \Omega^{-1}y$, $\varphi = \mu$ and $k(\varphi) = \frac{1}{2}\varphi^{\mathrm{T}}\Omega^{-1}\varphi$. It is easy to check that $\nabla k(\varphi) = \Omega^{-1}\varphi = \mathrm{E}(S)$ and $\nabla^2 k(\varphi) = \Omega^{-1} = \mathrm{var}(S)$.
- ☐ In the satellite example $d = 2$, $\Omega = D^{-1}$ is diagonal, and with $\theta^{\mathrm{T}} = (\psi, \lambda)$ we have

$$\varphi^{\mathrm{T}} = (\varphi_1, \varphi_2) = (\psi\cos\lambda, \psi\sin\lambda), \quad s(Y) = (d_1Y_1, d_2Y_2), \quad k(\varphi) = d_1\varphi_1^2/2 + d_2\varphi_2^2/2.$$

  The θ parametrisation gives the polar coordinates of the mean φ, but these are clearly equivalent because there is a 1–1 mapping between them.

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 every constant vector $a_{n\times 1}$, and then

- $M_Y(t) = \exp(t^{\mathrm{T}}\mu + \frac{1}{2}t^{\mathrm{T}}\Omega t)$ and 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 constants $a_{m\times 1}$ and $B_{m\times n}$,

$$a + BX \sim \mathcal{N}_m\left(a + B\mu, B\Omega B^{\mathrm{T}}\right);$$

- $a + BX$ and $c + DX$ are independent iff $B\Omega D^{\mathrm{T}} = 0$;
- 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:

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

Exponential family models II

- When $\dim s = d' > \dim\theta = d$ the model is called a (d', d) **curved exponential family**, and the $d' \times 1$ vector $\varphi(\theta)$ gives a d-dimensional sub-manifold of $\mathbb{R}^{d'}$.
- Exponential families are **closed under sampling**: the joint density of independent observations $Y_1, \ldots, Y_n$ from an exponential family with the same $s(Y_j)^{\mathrm{T}}\varphi = S_j^{\mathrm{T}}\varphi$ is

$$\prod_{j=1}^n f(y_j;\theta) = \prod_{j=1}^n m(y_j)\exp\left\{s_j^{\mathrm{T}}\varphi - k_j(\varphi)\right\} = \prod_{j=1}^n m(y_j)\exp\left\{\left(\sum_{j=1}^n s_j\right)^{\mathrm{T}}\varphi - \sum_{j=1}^n k_j(\varphi)\right\},$$

  so with $k_S(\varphi) = \sum_j k_j(\varphi)$, the density of $S = \sum_j S_j = \sum_j s(Y_j)$ is

$$f(s;\theta) = m^*(s)e^{s^{\mathrm{T}}\varphi - k_S(\varphi)}, \quad \text{with} \quad m^*(s) = \int_{\{y:\sum_j s(y_j)=s\}} \prod_{j=1}^n m(y_j)\,\mathrm{d}y.$$

  This is an exponential family, with canonical statistic S, canonical parameter φ and cumulant generator $k_S(\varphi)$.

Example 3 (Satellite conjunction) *Show that taking ψ known in Example 2 gives a $(2, 1)$ exponential family.*

Note to Example 3

We previously had

$$\varphi^{\mathrm{T}} = (\varphi_1, \varphi_2) = (\psi\cos\lambda, \psi\sin\lambda), \quad s(Y) = (d_1Y_1, d_2Y_2), \quad k(\varphi) = d_1\varphi_1^2/2 + d_2\varphi_2^2/2,$$

but with ψ known we can write

$$\varphi^{\mathrm{T}} = (\varphi_1, \varphi_2) = (\cos\lambda, \sin\lambda), \quad s(Y) = (\psi d_1Y_1, \psi d_2Y_2), \quad k(\varphi) = \psi^2(d_1\varphi_1^2 + d_2\varphi_2^2)/2,$$

where λ is the only unknown parameter. This is a $(2,1)$ exponential family because it cannot be written in terms of a scalar φ; the mean traces a curve (a circle) as λ varies.

Order statistics

- The **order statistics** of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ are the ordered values

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n-1)} \le X_{(n)}.$$

- In particular, the **minimum** is $X_{(1)}$, the **maximum** is $X_{(n)}$, and the **median** is

$$X_{(m+1)} \quad (n = 2m+1, \text{ odd}), \qquad \tfrac{1}{2}(X_{(m)} + X_{(m+1)}) \quad (n = 2m, \text{ even}).$$

  The median is the central value of $X_1, \ldots, X_n$.
- If f is continuous then the X_j must be distinct, and for $r = 1, \ldots, n$ we have

$$\begin{aligned} \mathrm{P}(X_{(r)} \le x) &= \sum_{j=r}^{n} \binom{n}{j} F(x)^j \{1-F(x)\}^{n-j}, \\ f_{X_{(r)}}(x) &= \frac{n!}{(r-1)!\,1!\,(n-r)!} F(x)^{r-1} f(x) \{1-F(x)\}^{n-r}. \end{aligned}$$

- Joint densities can be obtained using the argument that gives $f_{X_{(r)}}(x)$, and in particular

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! f(x_1) \cdots f(x_n), \quad x_1 < \cdots < x_n.$$

Example 4 *Find the joint density of $X_{(2)}, \ldots, X_{(n-1)}$ given that $X_{(1)} = x_1$ and $X_{(n)} = x_n$.*

Note: densities of order statistics

- The event $X_{(r)} \le x$ occurs iff at least r of the independent variables $X_1, \ldots, X_n$ are less than or equal to x, and each of them does this with probability $F(x)$. Hence the probability of the event is given by a binomial probability, and a little thought shows that this is the stated formula.
- The density can be obtained by differentiation of $\mathrm{P}(X_{(r)} \le x)$, whereupon one finds that almost all the terms cancel, giving the stated density. A nicer argument is as follows: for the event $X_{(r)} \in [x, x+dx)$, we need to split the sample into three groups of respective sizes $r-1$, 1 and $n-r$ (hence the multinomial coefficient) and probabilities $F(x)$, $f(x)\mathrm{d}x$, and $1-F(x)$. An application of the multinomial distribution gives the required formula.
- For the joint density we divide the sample into n parts, each with one observation, and apply a version of the multinomial argument just given.

Note to Example 4

- The joint density of $X_{(1)}$ and $X_{(n)}$ is given by splitting the sample into three parts, with respective probabilities $f(x_1)\mathrm{d}x_1$, $F(x_n) - F(x_1)$ and $f(x_n)\mathrm{d}x_n$, and noting that we want to have 1, $n-2$ and 1 of the total n observations in the three parts, giving

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = \frac{n!}{1!(n-2)!1!} f(x_1)\{F(x_n) - F(x_1)\}^{n-2} f(x_n), \quad x_1 < x_n,$$

  where we have dropped the $\mathrm{d}x_1\mathrm{d}x_n$.
- Hence the conditional density of $X_{(2)}, \ldots, X_{(n-1)}$ given that $X_{(1)} = x_1$ and $X_{(n)} = x_n$ is

$$\frac{n! f(x_1)\cdots f(x_n)}{n!/(n-2)! \times f(x_1)\{F(x_n) - F(x_1)\}^{n-2} f(x_n)} = (n-2)! \prod_{j=2}^{n-1} \frac{f(x_j)}{F(x_n) - F(x_1)},$$

  where $x_1 < x_2 < \cdots < x_{n-1} < x_n$. This is the joint density of the order statistics of a random sample of size $n-2$ from the truncated distribution $f(x)/\{F(x_n) - F(x_1)\}$, where $x_1 < x < x_n$.

Modes of convergence

- Let $X, X_1, X_2, \ldots$ be random variables with cumulative distribution functions $F, F_1, F_2, \ldots$. Then
  - 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 $\varepsilon > 0$;
  - 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.
  - A sequence $X_1, X_2, \ldots$ of estimators of a parameter θ is **(weakly) consistent** if $X_n \xrightarrow{P} \theta$.
- 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 called the **continuous mapping theorem** and **Slutsky's theorem**.

Limit theorems

Theorem 5 (Weak law of large numbers, WLLN) *If* $X, X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$ *and* $\mathrm{E}(X)$ *is finite, then* $\overline{X} = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{P} \mathrm{E}(X)$.

Theorem 6 (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 7 ('Delta method') *If* $a_n(X_n - \mu) \xrightarrow{D} Y$, $a_n, \mu \in \mathbb{R}$, $a_n \to \infty$ *as* $n \to \infty$, *and* g *is continuously differentiable at* μ *with* $g'(\mu) \neq 0$, *then* $a_n\{g(X_n) - g(\mu)\} \xrightarrow{D} g'(\mu)Y$.

- ☐ The CLT provides the finite-sample approximation $Z_n \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$, where $\dot{\sim}$ means 'is approximately distributed as'.
- ☐ 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 $J_g(\mu) = \partial g(\mu)/\partial \mu^{\mathrm{T}}$.

1.3 Statistics Revision

Statistical activities

- ☐ Planning of investigations
- ☐ Obtaining reliable data
- ☐ Exploratory data analysis/visualisation
- ☐ **Model formulation**
- ☐ **Point estimation** of a population parameter
- ☐ **Interval estimation** for a population parameter
- ☐ **Hypothesis testing** to assess whether observed data support a particular model
- ☐ **Prediction** of a future or unobserved random variable
- ☐ **Decision analysis** to choose an action based on data and the costs of potential actions

This course covers some aspects of those activities in red above.
Many inferential tasks can be formulated in decision-theoretic terms, but we shall mostly avoid this.

Statistical models

- ☐ Use observed data to draw conclusions about a 'population', i.e., a model from which the data are assumed to be drawn, or about future data.
- ☐ A **statistical model** is a family of probability distributions for data y in a sample space $\mathcal{Y}$.
- ☐ A **parametric model (family of models)** $f \equiv f(y;\theta)$ or equivalently $F \equiv f(y;\theta)$ is determined by **parameters** $\theta \in \Theta \subset \mathbb{R}^d$, for fixed finite d.
- ☐ If no such θ exists, F is **nonparametric**, and then the parameter is often determined by F through a **statistical functional** $\theta = t(F)$, e.g.,

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

- ☐ Parameters have different roles (which can change during an investigation):
  - **interest parameters** represent targets of inference (e.g., the mean of a population, the slope of a line, a baseline blood pressure) with direct substantive interpretations;
  - **nuisance parameters** are needed to complete a model specification, but are not themselves of main concern.
- ☐ A parametric model should have a 1–1 map from θ to $f(\cdot;\theta)$, so parameters identify models.

Model formulation

- ☐ Two broad types of statistical model:
  - **substantive** — based on fundamental subject-matter theory (e.g., quantum theory, Mendelian genetics, Navier–Stokes equations);
  - **empirical** — a convenient, adequately realistic, representation of data variation;
  - and of course a broad spectrum between them.
- ☐ We aim that
  - primary questions/issues are encapsulated in the interest parameter;
  - secondary aspects can be accounted for, often via nuisance parameters;
  - variation in the data is realistically modelled, leading to reasonable statements of uncertainty;
  - any special feature of the data or 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.

Some notation

- 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.
- We distinguish the data actually observed, y^{o}, from other possible values y, and likewise for estimators $\widehat{\theta}^{\mathrm{o}}$, probabilities $p^{\mathrm{o}} = \mathrm{P}(Y \geq y^{\mathrm{o}})$, $\ldots$, based on y^{o}.
- We write $\nabla \cdot = \partial \cdot / \partial \varphi$ and $\nabla^2 \cdot = \partial^2 \cdot / \partial \varphi \partial \varphi^{\mathrm{T}}$ for differentiation with respect to a parameter, and ∇_y etc., for other derivatives.
- In general discussion we often suppose that data Y come from some unknown 'true' density g, but we fit a candidate density $f(y; \theta)$ that may be different from g.

Point estimation

- An **estimator** of a parameter $\theta \in \Theta$ based on data Y is a random variable $\tilde{\theta} = \tilde{\theta}(Y)$ taking values in Θ. A specific value is an **estimate** $\tilde{\theta}(y)$.
- An **M**(aximisation)**-estimator** is computed using a function $\rho(y; \theta')$ as

$$\tilde{\theta} = \mathrm{argmax}_{\theta'} \frac{1}{n} \sum_{j=1}^{n} \rho(Y_j; \theta').$$

  Under certain conditions $\tilde{\theta}$ also solves

$$\frac{1}{n} \sum_{j=1}^{n} \nabla \rho(Y_j; \theta') = 0$$

  and is then called a **Z**(ero)**-estimator**.
- If the true underlying model is g, then $\tilde{\theta}$ is replaced by θ_g, where

$$\theta_g = \mathrm{argmax}_{\theta'} \int \rho(y; \theta') g(y) \, \mathrm{d}y, \qquad \int \nabla \rho(y; \theta_g) g(y) \, \mathrm{d}y = 0.$$

  Clearly if $g(y) = f(y; \theta)$, then we want $\theta_g = \theta$, uniquely.

Examples

- Equivalently we could minimise the **loss function** $-\rho$ with respect to θ.
- Some examples (for a d-dimensional parameter θ):
  - **maximum likelihood estimation** has $\rho(y;\theta') = \log f(y;\theta')$;
  - **method of moments estimation** has $h(y) = (y, y^2, \ldots, y^d)^{\mathrm{T}}$, $\mu(\theta') = \mathrm{E}\{h(Y)\}$, and
$$-\rho(y;\theta') = \{h(y) - \mu(\theta')\}^{\mathrm{T}}\{h(y) - \mu(\theta')\};$$
  - **generalized method of moments estimation** (widely used in econometrics) also has a symmetric positive definite $d \times d$ matrix $w(\theta')$ and
$$-\rho(y;\theta') = \{h(y) - \mu(\theta')\}^{\mathrm{T}} w(\theta')\{h(y) - \mu(\theta')\};$$
  - **least squares estimation** is method of moments estimation with $h(y_j) = y_j$ and $\mu_j(\theta') = \mathrm{E}(Y_j) = x_j^{\mathrm{T}}\theta'$;
  - **score-matching estimation** (unfortunate misnomer) with $Y \sim g$ has
$$-\rho(y;\theta') = \{\nabla_y \log f(y;\theta) - \nabla_y \log g(y)\}^2 .$$
- There are many (many!) other approaches to estimation.

Examples

Example 8 *Discuss maximum likelihood estimation of the parameters of the normal distribution.*

Example 9 *Discuss moment estimation of the parameters of the Weibull distribution.*

Example 10 *Show that under mild (but not entirely trivial) conditions on the density g, the population version of the score-matching estimator is*

$$\mathrm{argmin}_\theta \mathrm{E}\left[\{\nabla_y \log f(Y;\theta)\}^2 + 2\nabla_y^2 \log f(Y;\theta)\right],$$

and give the sample version.

Note to Example 8

☐ The density function of a normal random variable with mean μ and variance σ^2 is $(2\pi\sigma^2)^{-1/2}\exp\{-(y-\mu)^2/(2\sigma^2)\}$, so here $\theta_{2\times 1} = (\mu, \sigma^2)^{\mathrm{T}} \in \mathbb{R}\times\mathbb{R}_+$, and the likelihood for a random sample $y_1, \ldots, y_n$ equals

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_j;\theta) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_j-\mu)^2}{2\sigma^2}\right\}.$$

Therefore the log likelihood is

$$\ell(\mu,\sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2, \quad \mu\in\mathbb{R}, \sigma^2>0.$$

Its first derivatives are

$$\frac{\partial\ell}{\partial\mu} = \sigma^{-2}\sum_{j=1}^{n}(y_j-\mu), \quad \frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{n}(y_j-\mu)^2,$$

and its (negative) second derivatives are

$$\frac{\partial^2\ell}{\partial\mu^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2\ell}{\partial\mu\partial\sigma^2} = -\frac{n}{\sigma^4}(\overline{y}-\mu), \quad \frac{\partial^2\ell}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{j=1}^{n}(y_j-\mu)^2.$$

☐ To obtain the MLEs, we solve simultaneously the equations

$$\begin{pmatrix} \frac{\partial\ell(\mu,\sigma^2)}{\partial\mu} \\ \frac{\partial\ell(\mu,\sigma^2)}{\partial\sigma^2} \end{pmatrix} = \begin{pmatrix} \sigma^{-2}\sum_{j=1}^{n}(y_j-\mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{n}(y_j-\mu)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Now

$$\frac{\partial\ell(\widehat{\mu},\widehat{\sigma}^2)}{\partial\mu} = 0 \Rightarrow \frac{1}{\widehat{\sigma}^2}\sum_{j=1}^{n}(y_j-\widehat{\mu}) = 0 \Rightarrow n\widehat{\mu} = \sum_{j=1}^{n} y_j \Rightarrow \widehat{\mu} = n^{-1}\sum_{j=1}^{n} y_j = \overline{y}$$

and

$$\frac{\partial\ell(\widehat{\mu},\widehat{\sigma}^2)}{\partial\sigma^2} = 0 \Rightarrow \frac{n}{2\widehat{\sigma}^2} = \frac{1}{2\widehat{\sigma}^4}\sum_{j=1}^{n}(y_j-\widehat{\mu})^2 \Rightarrow \widehat{\sigma}^2 = n^{-1}\sum_{j=1}^{n}(y_j-\widehat{\mu})^2 = n^{-1}\sum_{j=1}^{n}(y_j-\overline{y})^2.$$

The first of these has the sole solution $\widehat{\mu} = \overline{y}$ for all values of σ^2, and therefore $\ell(\widehat{\mu},\sigma^2)$ is unimodal with maximum at $\widehat{\sigma}^2 = n^{-1}\sum(y_j-\overline{y})^2$. At the point $(\widehat{\mu},\widehat{\sigma}^2)$, the hessian matrix is diagonal with elements $\mathrm{diag}\{n/\widehat{\sigma}^2, n/(2\widehat{\sigma}^4)\}$, and so is positive definite. Hence $\widehat{\mu} = \overline{y}$ and $\widehat{\sigma}^2 = n^{-1}\sum(y_j-\overline{y})^2$ are the sole solutions to the likelihood equation, and therefore are the maximum likelihood estimates.

Note to Example 9

☐ A Weibull variable X has CDF $F(x) = 1 - e^{-(\lambda x)^\alpha}$, for $x > 0$ and $\lambda, \alpha > 0$, and is exponential when $\alpha = 1$. Note that $W = (\lambda X)^\alpha \sim \exp(1)$, so

$$\mathrm{E}(X^r) = \mathrm{E}\{(W^{1/\alpha}/\lambda)^r\} = \lambda^{-r}\mathrm{E}(W^{r/\alpha}) = \lambda^{-r}\int_0^\infty w^{r/\alpha}e^{-w}\,\mathrm{d}w = \lambda^{-r}\Gamma(1 + r/\alpha),$$

where $\Gamma(\cdot)$ is the gamma function. Hence with $\theta = (\lambda, \alpha)$ the moment estimators solve

$$\overline{Y} = \mu_1(\theta) = \lambda^{-1}\Gamma(1 + 1/\alpha), \quad \overline{Y^2} = \mu_2(\theta) = \lambda^{-2}\Gamma(1 + 2/\alpha), \quad \lambda, \alpha > 0,$$

i.e.,

$$\overline{Y^2}/(\overline{Y})^2 = \Gamma(1 + 2/\tilde{\alpha})/\Gamma(1 + 1/\tilde{\alpha})^2, \quad \tilde{\lambda} = \Gamma(1 + 1/\tilde{\alpha})/\overline{Y}.$$

Note to Example 10

☐ This approach to estimation can be useful when $\log f(y;\theta) = h(y;\theta) - k(\theta)$ with $k(\theta)$ intractable. It is a misnomer because the standard use of the term 'score' is for the derivative of the log likelihood with respect to θ (not y).

☐ On writing

$$\{\nabla_y \log f(y;\theta) - \nabla_y \log g(y)\}^2 = \{\nabla_y \log f(y;\theta)\}^2 - 2\nabla_y \log f(y;\theta)\nabla_y \log g(y) + \{\nabla_y \log g(y)\}^2,$$

we see that the population version of the estimator is

$$\theta_g = \mathrm{argmin}_\theta \int \{\nabla_y \log f(y;\theta)\}^2 g(y)\,\mathrm{d}y - 2\int \{\nabla_y \log f(y;\theta)\nabla_y \log g(y)\}\, g(y)\,\mathrm{d}y,$$

because θ does not appear in the third term of the square. Now $g(y)\nabla_y \log g(y) = \nabla_y g(y)$, so

$$\int \nabla_y \log f(y;\theta)\nabla_y \log g(y) g(y)\,\mathrm{d}y = \int \nabla_y \log f(y;\theta)\nabla_y g(y)\,\mathrm{d}y$$

and integration by parts implies that if the first term here is identically zero, this equals

$$[\nabla_y \log f(y;\theta) g(y)] - \int \nabla_y^2 \log f(y;\theta) g(y)\,\mathrm{d}y = -\mathrm{E}\left\{\nabla_y^2 \log f(Y;\theta)\right\}$$

Hence

$$\theta_g = \mathrm{argmin}_\theta \mathrm{E}\left[\{\nabla_y \log f(Y;\theta)\}^2 + 2\nabla_y^2 \log f(Y;\theta)\right],$$

whose sample version is

$$\tilde{\theta} = \mathrm{argmin}_\theta \sum_{j=1}^n \left[\{\nabla_y \log f(Y_j;\theta)\}^2 + 2\nabla_y^2 \log f(Y_j;\theta)\right],$$

which can be computed from the sample.

☐ Weighted versions can be used to kill the first term of the integration, if necessary.

Comparison of point estimators

- There are two generic bases for comparing point estimators:
  - **asymptotic** — what happens when $n \to \infty$?
  - **finite-sample** — what happens for sample sizes met in practice?
- **Consistency** is a key asymptotic criterion: does $\tilde{\theta}$ approach θ when $n \to \infty$?

Definition 11 *An estimator $\tilde{\theta}$ of θ is* **consistent** *if $\tilde{\theta} \xrightarrow{P} \theta$ as $n \to \infty$.*

- Consistency is necessary but not sufficient for an estimator to be good, because

$$\tilde{\theta} \xrightarrow{P} \theta \quad \Rightarrow \quad \tilde{\theta} + 10^6/\sqrt{\log\log n} \xrightarrow{P} \theta, \quad n \to \infty,$$

  but the second estimator here is useless: consistency can be considered a 'safety net'.
- Obviously we would like $\tilde{\theta}$ to be 'suitably close' to θ, by minimising

$$\mathrm{MSE}(\tilde{\theta};\theta) = \mathrm{E}\left\{(\tilde{\theta}-\theta)^2\right\}, \quad \mathrm{MAD}(\tilde{\theta};\theta) = \mathrm{E}\left(|\tilde{\theta}-\theta|\right),$$

  or other similar measures of distance (loss functions), asymptotically or in finite samples.

Bias-variance and other tradeoffs

- Using the **bias** $b(\tilde{\theta};\theta) = \mathrm{E}(\tilde{\theta}) - \theta$, the **mean square error** can be expressed as

$$\mathrm{MSE}(\tilde{\theta};\theta) = b(\tilde{\theta};\theta)^2 + \mathrm{var}(\tilde{\theta}),$$

  so we must balance ('trade off') the bias and the variance when choosing $\tilde{\theta}$.
- In simple problems we could insist that the estimator is **unbiased**, i.e., $b(\tilde{\theta};\theta) \equiv 0$, but this is usually artificial because
  - many good estimators are biased, and some unbiased estimators are useless;
  - it may be impossible to find an unbiased estimator; and
  - other properties may be more desirable (e.g., robustness).

  An exception is **meta-analysis**, which involves combining different estimators with possibly very varied sample sizes.

Example 12 *The method of moments estimator of a scalar θ based on a random sample $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ with sample average $\overline{Y}$ solves the equation $\mu(\theta) = \overline{Y}$. Show that if $\mu(\cdot)$ has two smooth derivatives and is 1–1, then the estimator is consistent and asymptotically normal, with bias and variance both of order n^{-1}.*

Note to Example 12

- ☐ As the function $\mu(\cdot)$ is smooth and 1–1, it has a differentiable inverse, and thus by the continuous mapping theorem, $\tilde{\theta} = \mu^{-1}(\overline{Y}) \xrightarrow{P} \mu^{-1}\{\mu(\theta)\} = \theta$, i.e., $\tilde{\theta}$ is consistent. For simplicity of notation write $g(x) = \mu^{-1}(x)$ below.
- ☐ Now $\overline{Y} = \mu + \sigma n^{-1/2} Z_n$, where $Z_n = (\overline{Y} - \mu)/(\sigma^2/n)^{1/2} \xrightarrow{D} Z \sim \mathcal{N}(0,1)$, and Taylor expansion gives

$$g(\overline{Y}) = g(\mu) + g'(\mu)\sigma n^{-1/2} Z_n + \frac{\sigma^2}{2} n^{-1} g''(Z_n') Z_n^2,$$

where $Z_n' \in (0, Z_n)$, i.e.,

$$\tilde{\theta} = \theta + n^{-1/2}\sigma g'(\mu) Z_n + n^{-1} A_n,$$

say, where A_n is a random variable of order 1. Taking expectations gives

$$b(\tilde{\theta}; \theta) = \mathrm{E}(\tilde{\theta}) - \theta = n^{-1}\mathrm{E}(A_n) = O(n^{-1}),$$

under mild further conditions on g''.

- ☐ Now

$$n^{1/2}(\tilde{\theta} - \theta)/\{\sigma g'(\mu)\} = Z_n + n^{-1/2} A_n' \xrightarrow{D} Z,$$

using this (or the delta method), so in large samples we have

$$\tilde{\theta} \mathrel{\dot\sim} \mathcal{N}\{\theta, \sigma^2 g'(\mu)^2/n\}.$$

Efficiency and the Cramèr–Rao lower bound

Definition 13 *If $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are estimators of scalar θ, then the* **relative efficiency** *of $\tilde{\theta}_1$ compared to $\tilde{\theta}_2$ can be defined as*

$$\frac{\mathrm{MSE}(\tilde{\theta}_2; \theta)}{\mathrm{MSE}(\tilde{\theta}_1; \theta)}.$$

In large samples the squared bias is often negligible compared to the variance, and we define the **asymptotic relative efficiency** *as $\mathrm{var}(\tilde{\theta}_2)/\mathrm{var}(\tilde{\theta}_1)$. Similar expressions apply if the parameter has dimension d.*

- ☐ Under mild conditions on the underlying model, a scalar estimator $\tilde{\theta}$ based on $Y \sim f(y; \theta)$ satisfies the **Cramèr–Rao lower bound**,

$$\mathrm{var}(\tilde{\theta}) \geq \frac{\{1 + \nabla b(\tilde{\theta}; \theta)\}^2}{\imath_n(\theta)},$$

where $\imath_n(\theta)$ is the **Fisher information**. This applies for any sample size n, but

  - as $n \to \infty$ the lower bound $\to 1/\imath_n(\theta)$, the asymptotic variance of the maximum likelihood estimator, which hence is most efficient in large samples; and
  - a similar result applies for vector θ.

Bartlett identities

- For data $Y \sim f(y;\theta)$ we define the **log likelihood function** $\ell(\theta) = \log f(Y;\theta)$ and $d \times 1$ **score vector** $U(\theta) = \nabla\ell(\theta)$.
- If we can differentiate with respect to θ under the integral sign, we 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 $U(\theta)$, viz

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

  where $\mathrm{var}\{U(\theta)\} = \imath_n(\theta)$ is called the **Fisher (or expected) information**.
- Later we shall see that in large samples, the maximum likelihood estimator $\widehat{\theta}$ satisfies

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

Note: 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 $nI_1(\theta)$.

Note: CRLB

- We have
$$\mathrm{E}(\tilde{\theta}) = \int \tilde{\theta}(y) f(y;\theta)\,\mathrm{d}y = \theta + b(\tilde{\theta};\theta),$$
and differentiation with respect to θ gives (setting $b'(\theta) = \mathrm{d}b(\tilde{\theta};\theta)/\mathrm{d}\theta$)
$$1 + b'(\theta) = \int \tilde{\theta}(y)\mathrm{d}f(y;\theta)/\mathrm{d}\theta\,\mathrm{d}y = \int \tilde{\theta}(y)\nabla\ell(\theta) f(y;\theta)\,\mathrm{d}y = \mathrm{E}\{\tilde{\theta}U(\theta)\} = \mathrm{cov}\{\tilde{\theta}, U(\theta)\},$$
because $U(\theta)$ has mean zero. Hence the definition of correlation gives
$$\mathrm{cov}\{\tilde{\theta}, U(\theta)\}^2 = \{1 + b'(\theta)\}^2 \leq \mathrm{var}(\tilde{\theta})\mathrm{var}\{U(\theta)\} = \mathrm{var}(\tilde{\theta}) I(\theta),$$
which gives the result.
- If the bias is of order n^{-1}, so too is its derivative, so in large samples we obtain
$$\mathrm{var}(\tilde{\theta}) \geq I(\theta)^{-1} = \mathrm{var}(\widehat{\theta}).$$

Interval estimation

- Point estimation does not express uncertainty — we would like to say whether the observed data y^{o} are consistent with different possible values of a parameter.
- We assess the plausibility of different values of θ by asking how well they explain y^{o}, often using a pivot.

Definition 14 *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 15 *If $M = \max(Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n \overset{\mathrm{iid}}{\sim} U(0,\theta)$, show that $Q_1 = M/\theta$ is a pivot and find a pivot based on $\overline{Y}$.*

Note to Example 15

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

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

so

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

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

- If $Y \sim U(0, \theta)$, then $\mathrm{E}(Y) = \theta/2$ and $\mathrm{var}(Y) = \theta^2/12$. Hence $\overline{Y}$ has mean $\theta/2$ and variance $\theta^2/(12n)$, and for large n, $\overline{Y} \dot{\sim} \mathcal{N}\{\theta/2, \theta^2/(12n)\}$ using the central limit theorem. Therefore

$$Q_2 = \frac{\overline{Y} - \theta/2}{\sqrt{\theta^2/(12n)}} = (3n)^{1/2}(2\overline{Y}/\theta - 1) \dot{\sim} \mathcal{N}(0, 1).$$

Thus Q_2 depends on both data and θ, and has an (approximately) known distribution: hence Q_2 is an (approximate) pivot.

- As $Y/\theta \sim U(0, 1)$, we see that we could use simulation to compute the exact distribution of Q_2, and thus obtain an exact pivot (apart from simulation error). This is called a bootstrap calculation, about which more later.

Confidence intervals

Definition 16 *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 $\leq$), 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

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

- 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

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

    into the form

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

    where the bounds $L = l(Y; \alpha_L, \alpha_U)$, $U = u(Y; \alpha_L, \alpha_U)$ do not depend on θ;
  - then we replace Y by its observed value y^{o} to get a realisation of the CI.
- Going from quantiles of Q to L, U is known as **inverting the pivot**.
- Often we use an approximate pivot of form $(\widehat{\theta} - \theta)/V^{1/2} \dot{\sim} \mathcal{N}(0,1)$, where V estimates $\mathrm{var}(\widehat{\theta})$ and $V^{1/2}$ is called a **standard error**. The resulting (approximate) 95% interval is $\widehat{\theta} \pm 1.96 V^{1/2}$.

Example 17 *In Example 15, find CIs based on Q_1 and on Q_2.*

Note to Example 17

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

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

  and a little algebra gives that

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

  so

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

- For $Q_2 = (3n)^{1/2}(2\overline{Y}/\theta - 1) \dot{\sim} \mathcal{N}(0,1)$, the quantiles are $z_{1-\alpha_L}$ and z_{α_U}, so

$$\mathrm{P}\{z_{\alpha_U} \le (3n)^{1/2}(2\overline{Y}/\theta - 1) \le z_{1-\alpha_L}\} = 1 - \alpha_L - \alpha_U,$$

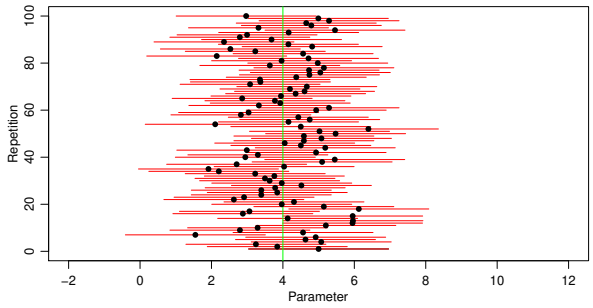
  and hence we obtain

$$L = \frac{2\overline{Y}}{1 + z_{1-\alpha_L}/(3n)^{1/2}}, \quad U = \frac{2\overline{Y}}{1 + z_{\alpha_U}/(3n)^{1/2}};$$

  note that for large n these are $L \approx 2\overline{Y}\{1 - z_{1-\alpha_L}/(3n)^{1/2}\}$ and $U \approx 2\overline{Y}\{1 - z_{\alpha_U}/(3n)^{1/2}\}$.

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 CI based on y^{o} is regarded as randomly chosen from the resulting infinity of CIs.
- Although we do not know if $\theta \in (l(y^{\text{o}}; \alpha_L, \alpha_U), u(y^{\text{o}}; \alpha_L, \alpha_U))$, 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.



More about CIs

- Almost invariably CIs are **two-sided** and **equi-tailed**, i.e., $\alpha_L = \alpha_U = \alpha$, but **one-sided** CIs of form $(-\infty, U)$ or (L, ∞) are sometimes required:
  - compute a two-sided interval with $\alpha_L = \alpha_U = \alpha$, then replace the unwanted limit by $\pm\infty$ (or another value if required in the context).
- For a two-sided CI we define the **lower- and upper-tail errors**

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

  and if these equal the required value for each possible α_L, α_U, then the **empirical coverage** of the CI exactly equals the desired value:
  - this occurs when the distribution of the corresponding pivot is known, but in practice this distribution is usually approximate, and then we use simulation to assess if and when CIs are adequate;
  - it's better to consider the two errors separately, as their sum may be OK even when they are individually incorrect;
  - lower- and upper-tail errors are properties of the CI procedure, not of individual intervals!

Prediction

- ☐ Prediction refers to 'estimation' of unobserved (future, latent, ...) random variables Y_+, say.
- ☐ Often require **prediction (or tolerance) intervals** based on existing data Y, by finding a pivot that depends on both Y_+ and Y, and predict Y_+ using this pivot, e.g., using its mean or median.
- ☐ Here's a very simple example ...

Example 18 *If* $Y_1, \ldots, Y_n, Y_+ \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$*, give prediction limits and a predictor for* Y_+ *based on the other variables.*

Note to Example 18

- ☐ Standard results give $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ independent of $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, both independent of $Y_+ \sim \mathcal{N}(\mu, \sigma^2)$, so $Y_+ - \overline{Y} \sim \mathcal{N}(0, \sigma^2 + \sigma^2/n)$, independent of S^2, leading to

$$Q = \frac{Y_+ - \overline{Y}}{\{(1+1/n)S^2\}^{1/2}} \sim t_{n-1},$$

  leading to two-sided equi-tailed $(1-2\alpha)$ prediction interval

$$\overline{Y} \pm (1+1/n)^{1/2} S t_{n-1}(1-\alpha).$$

  Note that even as $n \to \infty$ this interval does not vanish, rather it approaches $\mu \pm \sigma z_{1-\alpha}$.
- ☐ The Y_j are replaced by y_j^{o} to give the realisation of the interval.
- ☐ One obvious scalar predictor $\widehat{Y}_+$ is given by taking the median for Q, i.e., solving

$$q_{0.5} = \frac{\widehat{Y}_+ - \overline{Y}}{\{(1+1/n)S^2\}^{1/2}},$$

  where in this case $q_{0.5} = 0$, giving $\widehat{Y}_+ = \overline{Y}$ and realised value $\overline{y}^{\text{o}}$.

Hypothesis testing

- A **statistical hypothesis** is an assertion about the population underlying some data, or equivalently a restriction on possible models for the data, such as:
  - the population has mean μ_0;
  - the population is $\mathcal{N}(\mu_0, \sigma_0^2)$, with both parameters specified;
  - the population is $\mathcal{N}(\mu, \sigma^2)$, with the parameters unspecified;
  - the data are sampled from the discrete uniform distribution on $\{1, \ldots, 9\}$;
  - the population density is symmetric about some μ;
  - the population mean $\mu(x)$ increases when a covariate x increases.
- These are assertions about populations, not about a dataset, but they have implications for datasets.
- In some cases the distribution is fully specified, but not always.
- Some, but not all, hypotheses concern parameters.
- A **hypothesis test** uses a stochastic 'argument by contradiction' to make an inference about a statistical hypothesis: we assume that the hypothesis is true, and attempt to disprove it using data.

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.
- Tests on parameters are often based on pivots: if $\theta = \theta_0$, then $T = |q(Y; \theta_0)|$ has a known distribution G_0, say, and observing a value $t_{\text{obs}} = |q(y^{\text{o}}; \theta_0)|$ that is unusual relative to G_0 'contradicts' H_0.
- In other cases we choose a test statistic that seems plausible, such as Pearson's statistic,

$$T = \sum_{k=1}^{K} (O_k - E_k)^2 / E_k,$$

  used for testing how the agreement of observed counts O_k in K categories with their theoretical expectations E_k.
- In any case we need to know (or be able to approximate) the distribution of T under H_0.

1.4 Bases for Uncertainty Statements

Uncertainty

- Essentially three bases for statements of uncertainty:
  - 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;
  - a **Bayesian (inverse probability) inference** expresses it via a prior probability density and uses Bayes' theorem to update this in light of the data;
  - 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.
- A frequentist should choose the **reference set** $\mathcal{S}$ thoughtfully.

Example 19 (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 $\mathcal{S} = \{(y, m) : y \in \mathbb{R}, m \in \{1, 2\}\}$.

Comments on sampling theory inference

- 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 20 *What would the confidence intervals look like in Example 19? How would the image on slide 41 change? What hypothetical repetitions form the reference set?*

Bayesian inference

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

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

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

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

  here θ is regarded as random and y^{o} as fixed.
- A point estimate $\tilde{\theta}(y^{\mathrm{o}})$ of θ is obtained by minimising a **posterior expected loss**, i.e.,

$$\tilde{\theta}(y^{\mathrm{o}}) = \mathrm{argmin}_{\tilde{\theta}} \mathrm{E}\left\{ L(\theta, \tilde{\theta}) \mid y^{\mathrm{o}} \right\} = \mathrm{argmin}_{\tilde{\theta}} \int L(\theta, \tilde{\theta}) f(\theta \mid y^{\mathrm{o}}) \, \mathrm{d}\theta,$$

  where the **loss function** $L(\theta, \tilde{\theta}) \geq 0$ measures the loss when θ is estimated by $\tilde{\theta}$.

Comments on Bayesian inference

- 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.

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

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

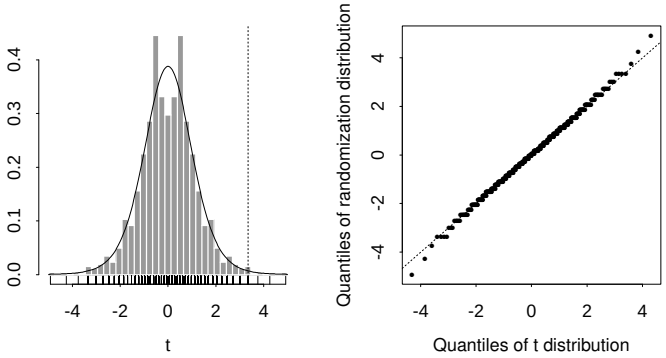
- Hence $\mathrm{E}(\overline{D}) = \psi$ and $\mathrm{var}(\overline{D}) = m^{-2}\sum_{j=1}^{m} c_j^2$, which is unknown because the c_j are unknown, is estimated by (exercise)

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

- $\overline{D}$ 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 $Z \dot{\sim} t_9$, 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 **repeated sampling (frequentist) approach**, which invokes hypothetical repetitions of the data-generating mechanism, or
  - a **Bayesian approach**, which requires that 'prior information' on unknown quantities be expressed as a probability distribution, 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.

2 Some Basic Concepts slide 58

2.1 Likelihood slide 59

Likelihood

- We now suppose that the data are provisionally believed to come from a parametric model $f_Y(y;\theta)$ for which $\theta \in \Theta$.
- Given observed data y, 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 $n^{-1}\ell(\theta)$.

- Comments:
  - the posterior density based on data y and prior $f(\theta)$ is proportional to $L(\theta) \times f(\theta)$;
  - the above formula is readily extended — for example, if $y_1, \ldots, y_n$ are in time order, then

$$\ell(\theta) = \sum_{j=2}^{n} \log f(y_j \mid y_1, \ldots, y_{j-1};\theta) + \log f(y_1;\theta).$$

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**

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

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

$$\jmath(\theta) = -\nabla^2 \ell(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad \imath(\theta) = \mathrm{E}\{\jmath(\theta)\};$$

these are $d \times d$ matrices if θ has dimension d and otherwise are scalars.

- To evaluate $\imath(\theta)$ we replace y by the random variable Y and take expectations.

Example 21 (Exponential family) *Find the likelihood quantities when $Y_1, \ldots, Y_n$ is a random sample from a (d, d) exponential family.*

Note to Example 21

☐ The density for a single observation is

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

where $s = s(y)$, so the corresponding log likelihood based on $y_1, \ldots, y_n$ is

$$\ell(\theta) = \sum_{j=1}^{n} \log f(y_j;\theta) \equiv \sum_{j=1}^{n} s_j^{\mathrm{T}}\varphi(\theta) - nk\{\varphi(\theta)\} = s^{\mathrm{T}}\varphi(\theta) - nk\{\varphi(\theta)\}, \quad \theta \in \Theta,$$

where $s = \sum_j y_j$ and $\equiv$ means that we have dropped additive constants from the log likelihood.

☐ If ∇ denotes gradient with respect to θ and k_φ and $k_{\varphi\varphi}$ denote the gradient and Hessian matrix of k with respect to φ, then the score equation is

$$\nabla\varphi(\theta)^{\mathrm{T}} s - n\nabla\varphi(\theta)^{\mathrm{T}} k_\varphi\{\varphi(\theta)\} = 0,$$

so if the $d \times d$ matrix $\varphi(\theta)^{\mathrm{T}}$ is invertible (which is the case for a smooth $1-1$ transformation), then the MLE $\widehat{\varphi}$ satisfies $k_\varphi(\widehat{\varphi}) = \overline{s} = s/n$ (note that $\mathrm{E}(S/n) = k_\varphi(\varphi)$, so $\widehat{\varphi}$ is also a moments estimate), and therefore $\widehat{\theta} = \varphi^{-1}(\widehat{\varphi})$.

☐ To compute the observed information we write the likelihood derivatives as

$$\frac{\partial\varphi_t}{\partial\theta_r} s_t - n\frac{\partial\varphi_t}{\partial\theta_r}\frac{\partial k(\varphi)}{\partial\varphi_t}, \quad r = 1, \ldots, d,$$

using the Einstein summation convention that implies summation over repeated indices (here t), and then differentiate with respect to θ_u to obtain

$$\jmath(\theta)_{r,u} = -\frac{\partial^2\varphi_t}{\partial\theta_r\partial\theta_u} s_t + n\frac{\partial^2\varphi_t}{\partial\theta_r\partial\theta_u}\frac{\partial k(\varphi)}{\partial\varphi_t} + n\frac{\partial\varphi_t}{\partial\theta_r}\frac{\partial\varphi_v}{\partial\theta_u}\frac{\partial^2 k(\varphi)}{\partial\varphi_t\partial\varphi_v}, \quad r, u = 1, \ldots, d.$$

Note that

- if $\varphi(\theta) \equiv \theta$, i.e., the exponential family is in canonical form, then $\nabla\varphi(\theta) = I_d$ and the second derivatives are zero, so this entire expression reduces to $n\nabla^2 k(\varphi)$, which is non-random;
- $\mathrm{E}(S_t) = n\partial k(\varphi)/\partial\varphi_t$, so in any case

$$\imath(\theta) = n\nabla\varphi(\theta)^{\mathrm{T}} k_{\varphi\varphi}\{\varphi(\theta)\}\left\{\nabla\varphi(\theta)^{\mathrm{T}}\right\}^{\mathrm{T}};$$

- the MLE satisfies the score equation, so the observed information at the MLE is

$$\jmath(\widehat{\theta}) = n\nabla\varphi(\widehat{\theta})^{\mathrm{T}} k_{\varphi\varphi}\{\varphi(\widehat{\theta})\}\left\{\nabla\varphi(\widehat{\theta})^{\mathrm{T}}\right\}^{\mathrm{T}}.$$

Invariance

- ☐ We 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

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

  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

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

  so the respective maximum likelihood estimates satisfy $\widehat{\phi} = \phi(\widehat{\theta})$.

Interest and nuisance parameters

- ☐ In most cases $\theta = (\psi, \lambda)$, where the
  - (low-dimensional, often scalar) **interest parameters** ψ represent targets of inference with direct substantive interpretations;
  - (maybe high-dimensional) **nuisance parameters** λ are needed to complete a model specification, but are not themselves of main concern.
- ☐ Ideally inference on ψ should be invariant to **interest-respecting (or interest-preserving) transformations**

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

- ☐ For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then the log-normal variable $Y = \exp(X)$ has mean $\psi = \exp(\mu + \sigma^2/2)$, and
  - confidence intervals for ψ should be the same whether the nuisance parameter λ is chosen as μ or σ^2 or $\mu - \sigma^2/2$ or ...;
  - if (L, U) is a confidence interval for ψ, then a confidence interval for $\log \psi$ should be $(\log L, \log U)$.
- ☐ Later we will try to construct likelihoods that depend only on the interest parameters.

2.2 Complications

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 — though the likelihood is readily constructed when the data are independent.
- 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;
  - 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

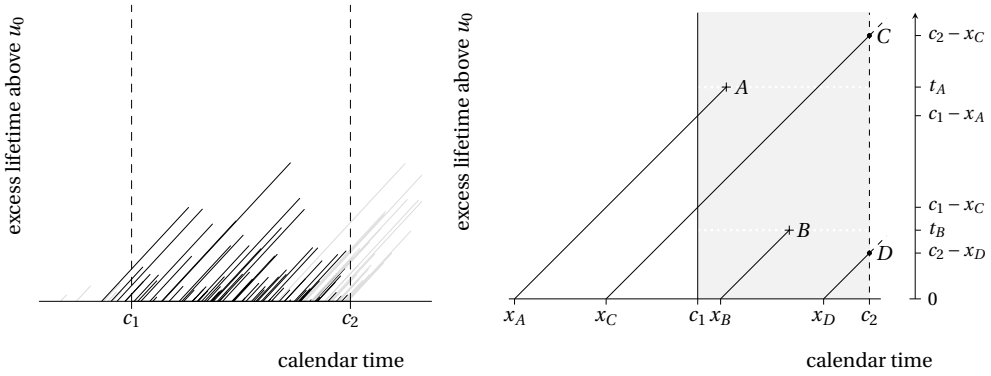
$$\prod_{j=1}^{n} f(y_j \mid y_j \in \mathcal{I}_j; \theta).$$

Example 22 *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.

Biased sampling

- Arises when the probability of selecting (sampling) an observation depends on its value.
- If $p(y) = \mathrm{P}(\mathcal{S} \mid Y = y)$ denotes the probability that an observation of size y is selected, then the density of a selected observation is

$$f_{\mathcal{S}}(y) = f(y \mid \mathcal{S}) = \frac{\mathrm{P}(\mathcal{S} \mid Y = y) f(y)}{\mathrm{P}(\mathcal{S})} = \frac{p(y) f(y)}{\int p(y) f(y)\,\mathrm{d}y}.$$

- A common example, **length-biased sampling**, occurs when $p(y) \propto y$, giving

$$f_{\mathcal{S}}(y) = \frac{y f(y)}{\int x f(x)\,\mathrm{d}x} = \frac{y f(y)}{\mu}, \quad y > 0,$$

  say, and the mean length for the selected observations is not the population mean $\mathrm{E}(Y) = \mu$ but

$$\mathrm{E}(Y \mid \mathcal{S}) = \int y f_{\mathcal{S}}(y)\,\mathrm{d}y = \int y^2 f(y)/\mu\,\mathrm{d}y = \mu + \sigma^2/\mu,$$

  where $\sigma^2 = \mathrm{var}(Y)$ is the population variance.
- Many other types of biased sampling arise in medical and epidemiological studies.

Censoring

- Selection determines which observations appear in a sample, whereas censoring reduces the information available in the sample.
- **Censoring** is very common in 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 together; see the Lexis diagram.

Dependent data

- If the joint density of $Y = (Y_1, \ldots, Y_n)$ is known, then the **prediction decomposition**

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

  gives the density (and hence the likelihood).
- 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 23 (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 23

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 common 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 the 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

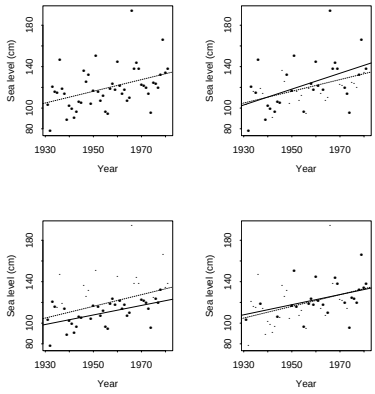
$$\mathrm{P}(I = 1 \mid x, y) f(y \mid x; \theta) f(x; \theta), \qquad \int \mathrm{P}(I = 0 \mid x, y) f(y \mid x; \theta) f(x; \theta) \, \mathrm{d}y,$$

  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. 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, 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.

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.
- ☐ 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.

2.3 Data Reduction

Sufficiency

- ☐ When can a lot of data from a particular model be reduced to a few relevant quantities without any loss of information?
- ☐ 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.

Example 24 (Uniform model) *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(\theta)$, find a sufficient statistic for θ and say how to use $f(y \mid s)$ to assess model fit.*

Note to Example 24

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

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

  where $m = \max(y_1, \ldots, y_n)$; note that $\prod_j I(0 < y_j < \theta) = I(0 < m < \theta)$. Clearly the likelihood depends on the data only through n and m, and as n is taken to be fixed, a sufficient statistic is $M = \max Y_j$.
- We have $\mathrm{P}(M \leq m) = (m/\theta)^n$ for $0 < m < \theta$, so M has density nm^{n-1}/θ^n for $0 < m < \theta$, but noting that $(Y_1/M, \ldots, Y_n/M)$ has a 1 somewhere unknown, we see that computing the conditional density of the observations given M it is easiest to first compute that of the order statistics, i.e.,

$$f(y_1, \ldots, y_{n-1}, m) = n!\theta^{-n}, \quad 0 < y_1 < \cdots < y_{n-1} < m < \theta,$$

  so the joint density of $Y_{(1)}, \ldots, Y_{(n-1)}$ given $M = m$ is

$$\frac{n!\theta^{-n}}{nm^{n-1}/\theta^n} = \frac{(n-1)!}{m^{n-1}}, \quad 0 < y_1 < \cdots < y_{n-1} < m,$$

  which is the density of the order statistics of a random sample of size $n-1$ from the $U(0,m)$ density. Tests of fit will be based on this density, which does not depend on θ.
- Equivalently we compute $(Y_1/M, \ldots, Y_n/M)$, throw away the 1, and treat the remaining values as $n-1$ independent $U(0,1)$ variables if we want to compute tests of fit of the model.

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 statistic', we define a **minimal sufficient statistic** to be a function of any other sufficient statistic. This is unique up to 1–1 maps.
- To formalise this idea, we note that
  - any statistic $T = t(Y)$ taking values $t \in \mathcal{T}$ partitions the sample space $\mathcal{Y}$ into equivalence classes $\mathcal{C}_t = \{y' \in \mathcal{Y} : t(y') = t\}$;
  - the partition $\mathcal{C}_t$ corresponding to T is sufficient if and only if the distribution of Y within each $\mathcal{C}_t$ does not depend on θ; and
  - a minimal sufficient statistic gives the coarsest possible sufficient partition.
- We use the following results to identify (minimal) sufficient statistics.

Theorem 25 (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 26 *If $Y \sim f(y;\theta)$ and $S = s(Y)$ is such that $\log f(z;\theta) - \log f(y;\theta)$ is free of θ if and only if $s(z) = s(y)$, then S is minimal sufficient for θ.*

Note to Theorem 25

- ☐ 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 26

- ☐ We must show that that S is sufficient and that it is minimal. For simplicity let Y be discrete.
- ☐ 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

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

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.

Examples

Example 27 (Uniform model) *Discuss minimal sufficiency when* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(\theta)$.

Example 28 (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.*

Note to Example 27

- The density of $Y_1, \ldots, Y_n$ is

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

  where $m = \max(y_1, \ldots, y_n)$, so the factorisation theorem implies that $M = \max(Y_1, \ldots, Y_n)$ is sufficient, as we already deduced in Example 24 .

- In the calculation below we set $0/0 = 1$. To show that M is minimal sufficient, note that if we have two samples $y_1, \ldots, y_n$ and $z_1, \ldots, z_{n'}$, then (in an obvious notation)

$$\frac{f(z;\theta)}{f(y;\theta)} = \frac{\theta^{-n} I(0 < m_z < \theta)}{\theta^{-n'} I(0 < m_y < \theta)},$$

  which is independent of θ iff $n = n'$ and $m_y = m_z$, i.e., the samples have the same size and the same maxima. Since we usually take the size as non-random (for reasons seen later), the sample maximum is minimal sufficient for θ.

- To illustrate the idea of sufficient partitions, let $U = \min(Y_1, \ldots, Y_n)$, so $S = (U, M)$ is also sufficient. The partitions of the sample space $\mathcal{Y} = (0, \theta)^n$ corresponding to the statistics U, M and (U, M) have elements $\mathcal{C}_u = \{y \in \mathcal{Y} : u(y) = u\}$, $\mathcal{C}_m = \{y \in \mathcal{Y} : m(y) = m\}$ and

$$\mathcal{C}_{u,m} = \{y \in \mathcal{Y} : u(y) = u, m(y) = m\}, \quad 0 < u < m < \theta,$$

  where for brevity we write $y = (y_1, \ldots, y_n)$; C_u contains all the samples that have minimum u, for example. Notice that the same partition $\mathcal{C}_u$ would arise if we replaced u by 1–1 function $g(u)$.

- Sketch the partitions on the board!

- We already saw that the density of $(Y_1, \ldots, Y_n)$ given that $M = m$, i.e., the conditional density of $Y = y$ inside $\mathcal{C}_m$, is the density of $n-1$ independent $U(0, m)$ variables, which does not depend on θ, so the partition $\{\mathcal{C}_m : 0 < m < \theta\}$ is sufficient. Obviously the same is also true of $\{\mathcal{C}_{um} : 0 < u < m < \theta\}$.

- The density of U is given by differentiation of $\mathrm{P}(U \leq u) = 1 - (1 - u/\theta)^n$, for $0 < u < \theta$, i.e., $n\theta^{-1}(1 - u/\theta)^{n-1}$ for $0 < u < \theta$, so the conditional density of $Y_1, \ldots, Y_n$ given U is

$$\frac{\theta^{-n} I(0 < m < \theta)}{n\theta^{-1}(1 - u/\theta)^{n-1} I(0 < u < \theta)} = \frac{1}{n(\theta - u)^{n-1}} I(0 < u < m < \theta),$$

  which depends on θ. Hence the partition $\{\mathcal{C}_u : 0 < u < \theta\}$ is not sufficient.

Note to Example 28

- The density g is continuous, so all the y_j are distinct with probability one. The joint density of $Y_1, \ldots, Y_n$ is therefore

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

and that of the order statistics $S = (Y_{(1)}, \ldots, Y_{(n)})$ is

$$f(s;\theta) = n! \prod_{j=1}^{n} g(y_j - \theta), \quad y_1 < \cdots < y_n,$$

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 the order of $(y_1, \ldots, y_n)$, all of which have order statistics s; clearly $|\mathcal{Y}_s| = n!$, because there are no ties.
As $f(y \mid s)$ does not depend on g or θ, the set of order statistics S is sufficient for g and θ.

- To show minimality, take another sample $z_1, \ldots, z_n$ and note that

$$\frac{f(z;\theta)}{f(y;\theta)} = \frac{\prod_{j=1}^{n} g(z_j - \theta)}{\prod_{j=1}^{n} g(y_j - \theta)},$$

which (for general g) is free of θ only if the y_j are a permutation of the z_j, and this occurs only if the order statistics of the samples are the same.

- Here $|s| = n$ in general. In special cases (e.g., the normal density) there is a minimal sufficient statistic of lower dimension.

Using sufficiency: Rao–Blackwell theorem

Theorem 29 (Rao–Blackwell) *If $\tilde{\theta}$ is an unbiased estimator of a parameter θ of a statistical model $f(y;\theta)$ and if $S = s(Y)$ is sufficient for θ, then $T = \mathrm{E}(\tilde{\theta} \mid S)$ is also unbiased, and $\mathrm{var}(T) \leq \mathrm{var}(\tilde{\theta})$.*

Example 30 (Exponential family) *Find a minimal sufficient statistic for θ based on a random sample $Y_1, \ldots, Y_n$ from a (d, d) exponential family. If $d = 1$ and $s(Y) = Y$, find a better unbiased estimator of $\mu = \mathrm{E}(Y_1)$ than Y_1.*

- The Rao–Blackwell theorem is non-asymptotic: it holds for any n.
- The process of getting a better estimator, **Rao–Blackwellization**, is useful in many contexts (e.g., as a variance reduction technique in MCMC estimation).

Note to Theorem 29

☐ We must show that that T is a statistic, that it is unbiased, and that it has smaller variance than $\tilde{\theta}$.

☐ We have

$$T = \mathrm{E}(\tilde{\theta} \mid S) = \int \tilde{\theta}(y) f(y \mid s) \, \mathrm{d}y,$$

which does not depend on θ by sufficiency of S, so T is indeed a statistic.

☐ Moreover

$$\mathrm{E}(T) = \int \left\{ \int \tilde{\theta}(y) f(y \mid s) \, \mathrm{d}y \right\} f(s; \theta) \, \mathrm{d}s = \int \tilde{\theta}(y) f(y; \theta) \, \mathrm{d}y = \theta,$$

by unbiasedness of $\tilde{\theta}$.

☐ Finally we write $\tilde{\theta} - \theta = \tilde{\theta} - T + T - \theta = A + B$, say, and note that $\mathrm{E}(A \mid S) = \mathrm{E}(B) = 0$, so

$$\mathrm{cov}(A, B) = \mathrm{E}_S \mathrm{E}_{Y|S}(AB) = \mathrm{E}_S \left\{ B \mathrm{E}_{Y|S}(A \mid S) \right\} = \mathrm{E}_S(B0) = 0,$$

and thus

$$\mathrm{var}(\tilde{\theta}) = \mathrm{var}(A + B) = \mathrm{var}(A) + \mathrm{var}(B) = \mathrm{var}(\tilde{\theta} - T) + \mathrm{var}(T) \geq \mathrm{var}(T),$$

with equality iff $\mathrm{E}\{(T - \tilde{\theta})^2\} = 0$, i.e., T and $\tilde{\theta}$ are equal almost everywhere.

Note to Example 30

☐ The joint density is

$$\prod_{j=1}^{n} m(y_j!) \exp\left[s(y_j)^{\mathrm{T}} \varphi(\theta) - k\{\varphi(\theta)\} \right] = \prod_{j=1}^{n} m(y_j!) \times \exp\left[s^{\mathrm{T}} \varphi(\theta) - nk\{\varphi(\theta)\} \right], \quad \theta \in \Theta,$$

so $(s = \sum s(y_j), n)$ is sufficient by the factorisation theorem. It is also minimal, because the log density for samples $z_1, \ldots, z_n$ and $y_1, \ldots, y_m$,

$$\sum_{j=1}^{n} \log f(z_j; \theta) - \sum_{j=1}^{m} \log f(y_j; \theta),$$

does not depend on θ iff $\sum s(y_j) = \sum s(z_j)$ (and $n = m$). As usual we drop n from the minimal sufficient statistic.

☐ To find the unbiased estimator we argue by symmetry: clearly $\mathrm{E}(Y_1 \mid S) = \cdots = \mathrm{E}(Y_n \mid S)$ because S is symmetric in the Y_j and the latter were IID. Hence

$$\mathrm{E}(Y_1 \mid S) = n^{-1} \sum_{j=1}^{n} \mathrm{E}(Y_j \mid S) = \mathrm{E}\left(n^{-1} \sum_{j=1}^{n} Y_j \mid S \right) = \mathrm{E}(S \mid S) = S,$$

and clearly $\mathrm{var}(S) = \mathrm{var}(Y_1)/n$.

Complete statistics

- If we have numerous unbiased estimators, all of which could be improved, then we would like to find the best.
- To force uniqueness we introduce **completeness**: a statistic S (or its density) is **complete** if for any function h,
$$\mathrm{E}\{h(S)\} = 0 \text{ for all } \theta \implies h(s) \equiv 0,$$
and S is **boundedly complete** if this is true provided h is bounded.
- If S is complete, then two unbiased estimators based on S satisfy
$$\mathrm{E}\{\tilde{\theta}_1(S) - \tilde{\theta}_2(S)\} = 0 \quad \text{for all} \quad \theta,$$
so by completeness $\tilde{\theta}_1(S) = \tilde{\theta}_2(S)$ is unique.

Example 31 *Show that the maximum of a uniform sample is complete, and hence find the unique minimum variance unbiased estimator of θ.*

Theorem 32 (No proof) *The minimal sufficient statistic in a (d, d) exponential family (i.e., one for which the parameter space contains an open d-dimensional set) is complete.*

Note to Example 31

- The density of M is of the form
$$f(m;\theta) = a(m)b(\theta)I(0 < m < \theta), \quad 0 < m < \theta, \quad \theta > 0,$$
where $a(m) = nm^{n-1}$ and $b(\theta) = \theta^{-m}$, so suppose for a contradiction that there exists a function h for which $h(m) \neq 0$ but
$$0 = \mathrm{E}\{h(M)\} = \int_0^\theta a(m)b(\theta)h(m)\,\mathrm{d}m \propto \int_0^\theta a(m)h(m)\,\mathrm{d}m, \quad \theta > 0.$$
- The integral here equals zero for all θ so its derivative $a(\theta)h(\theta)$ with respect to θ must be zero. However, $a(m) \neq 0$, so $h(\theta) = 0$ for all $\theta > 0$, which is a contradiction. Hence M is complete.
- For the unbiased estimator, we note that $\mathrm{E}(M) = n\theta/(n+1)$, so $\tilde{\theta} = (n+1)M/n$ is unbiased and must therefore be the unique minimum variance unbiased estimator of θ.

Using sufficiency: Eliminating nuisance parameters

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

Lemma 33 *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 34 *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 35 (2×2 table) *Apply Lemma 34 to the 2×2 table.*

Note to Lemma 33

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 34

In the discrete case, let $\sum_{t,w}$ and $\sum_w$ denote sums over the sets $\{y : t(y)=t^{\mathrm{o}}, w(y)=w^{\mathrm{o}}\}$ and $\{y : w(y)=w^{\mathrm{o}}\}$, and note that

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

so

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

say, where the cumulant generator for the conditional density depends on w^{o}. This is the announced exponential family.

Note to Example 35

☐ 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

$$\pi_0 = \frac{e^{\lambda}}{1+e^{\lambda}}, \quad \pi_1 = \frac{e^{\lambda+\psi}}{1+e^{\lambda+\psi}}, \quad \psi, \lambda \in \mathbb{R}$$

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**

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

Ancillary statistics

- Sometimes we can write a minimal sufficient statistic as $S = (T, A)$ where $A = a(Y)$ is an **ancillary statistic**, defined as 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 A considered as fixing the reference set $\mathcal{S}$ used in repeated sampling inference.
- A **distribution-constant statistic** is one whose distribution does not depend on the parameter.

Example 36 (Sample size) *If $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} f(y;\theta)$, with the sample size N stemming from a random mechanism, then clearly the most general sufficient statistic is $(Y_1, \ldots, Y_N, N)$. If the distribution of N that does not depend on θ, however,*

$$f(y, n;\theta) = f(y \mid n;\theta) f(n) = \prod_{j=1}^{n} f(y_j;\theta) \times f(n),$$

so N is ancillary for θ, and we should use the reference set consisting of vectors $y_1, \ldots, y_n$ of length n.

Ancillary statistics II

Example 37 (Regression) *In a regression setting a response vector $Y_{n\times 1}$ depends on a matrix $X_{n\times p}$ of covariates. If their joint density factorises as $f(y \mid x;\psi) f(x)$, so that the interest parameters ψ only appear in the first term, then we should treat the X matrix as fixed, even if (Y, X) are actually sampled from some distribution.*

Example 38 (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

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

Theorem 39 (Basu) *A complete minimal sufficient statistic is independent of any distribution-constant statistic.*

Note to Example 38

☐ 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.

Note to Theorem 39

☐ In the discrete case, note that for any c and θ, the marginal density of C may be written using the sufficient statistic S as

$$f_C(c) = \sum_s f_{C|S}(c \mid s) f_S(s; \theta),$$

so for all θ we have

$$\sum_s \{f_C(c) - f_{C|S}(c \mid s)\} f_S(s; \theta) = 0,$$

and completeness of S implies that $f_C(c) = f_{C|S}(c \mid s)$ for every c and s, i.e., $C \perp\!\!\!\perp S$.

☐ The argument in the continuous case is analogous.

'Ideal' frequentist inference

- Frequentist recipe for inference on an interest parameter ψ:
  - find the likelihood function for the data Y;
  - find a sufficient statistic $S = s(Y)$ of the same dimension as θ;
  - eliminate any nuisance parameters λ;
  - find a function T of S whose distribution depends only on ψ;
  - use the distribution of T (conditioned on any ancillary statistics) for inference (confidence limits/tests) for ψ;
  - (use the conditional distribution of Y given S to assess model adequacy).
- For inference note that if T is continuous with distribution F, observed value t^o and the true value of ψ is ψ_0, then

$$F(T;\psi_0) \sim U(0,1) \quad \text{is a pivot,}$$

so confidence limits for ψ_0 are given by inverting it, i.e., solving $F(t^o;\psi_\alpha) = \alpha$ for appropriate values of α.

Note: Why is a P-value uniform?

- For simplicity write $F_0(t) = \mathrm{P}(T \le t;\psi_0)$, and note if $T \sim F_0$, then

$$\mathrm{P}\{F_0(T) \le u\} = \mathrm{P}\{T \le F_0^{-1}(u)\} = F_0\{F_0^{-1}(u)\} = u, \quad 0 < u < 1,$$

i.e., $F_0(T) \sim U(0,1)$ is a pivot, because it depends on the data (through T), the parameter ψ_0, and has a known distribution.

- The above proof works for any continuous T, but is only approximate if T is discrete (e.g., has a Poisson distribution). In such cases $F_0(T)$ can only take a finite or countable number of values that give the **achievable confidence levels**.

Significance functions

- It is useful to plot the **P-value (or significance) function**

$$p(\psi) = \mathrm{P}(T \ge t^o;\psi) = 1 - F(t^o;\psi) \quad \text{against} \quad \psi.$$

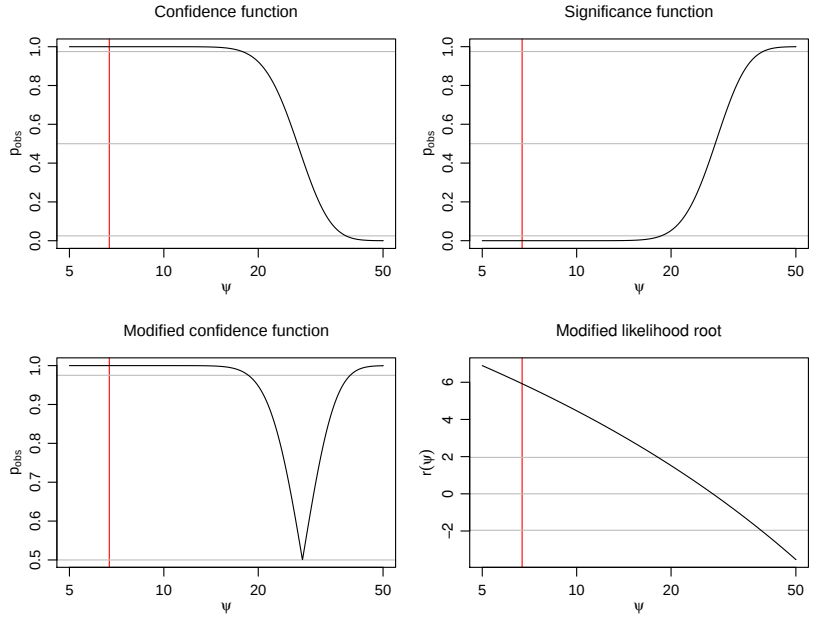
- As $F_0(T) \sim U(0,1)$ when $\psi = \psi_0$, we regard values of ψ for which $p(\psi)$ is too extreme as incompatible with t^o, leading to the (two-sided) $(1-\alpha)$ confidence set

$$\{\psi : \alpha/2 \le p(\psi) \le 1 - \alpha/2\},$$

or to using $p(\psi_0)$ as the P-value for a test of $H_0 : \psi = \psi_0$ against $H_1 : \psi > \psi_0$.

- Equivalent functions include
  - the **confidence function** $1 - p(\psi)$;
  - the **modified confidence function** $\max\{p(\psi), 1 - p(\psi)\}$; and
  - a **pivot function** showing how a (standard normal) pivot varies with ψ.

Significance and related functions



Examples

Example 40 (Normal sample) *Apply the recipe above to inference for the mean of a normal random sample with known variance.*

Example 41 (Uniform sample) *Apply the recipe above to inference for the upper limit of a uniform sample.*

Example 42 (2×2 table) *Apply the recipe above to the 2×2 table.*

Note to Example 40

- ☐ Suppose that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\psi, 1)$. This is a (1,1) exponential family, so the minimal sufficient statistic is $S = \overline{Y} \sim \mathcal{N}(\psi, 1/n)$, and clearly we should take $T = \overline{Y}$, so $\sqrt{n}(\overline{Y} - \psi) \sim \mathcal{N}(0, 1)$.
- ☐ Here the significance function is

$$p(\psi) = \mathrm{P}(T \geq t^{\mathrm{o}}; \psi) = 1 - \Phi\{n^{1/2}(\overline{y}^{\mathrm{o}} - \psi)\} = \Phi\{n^{1/2}(\psi - \overline{y}^{\mathrm{o}})\},$$

  and solving this for $p(\psi_\alpha) = \alpha$ gives $n^{1/2}(\psi_\alpha - \overline{y}^{\mathrm{o}}) = z_\alpha$, i.e., $\psi_\alpha = \overline{y}^{\mathrm{o}} + n^{-1/2} z_\alpha$, leading to the familiar $(1 - \alpha)$ confidence interval (L, U) with observed value

$$(\overline{y}^{\mathrm{o}} + n^{-1/2} z_{\alpha/2}, \quad \overline{y}^{\mathrm{o}} + n^{-1/2} z_{1-\alpha/2}).$$

- ☐ For the model assessment step we could note that as $S = \overline{Y}$ is a complete minimal sufficient statistic, the distribution-constant statistic $C = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$ is independent of $\overline{Y}$ (by Basu's theorem), and therefore plots and tests of the suitability of the model would be based on C.

Note to Example 41

We have already seen that M is minimal sufficient and that its distribution $\mathrm{P}(M \le x) = (x/\theta)^n$, for $0 < x < \theta$, depends only on θ. Hence the corresponding significance function based on an observed m^{o} would be

$$p(\theta) = 1 - (m^{\mathrm{o}}/\theta)^n \quad \theta > m^{\mathrm{o}},$$

from which we read off the limits using the equation $\alpha = 1 - (m^{\mathrm{o}}/\theta_\alpha)^n$, i.e., $\theta_\alpha = m^{\mathrm{o}}(1-\alpha)^{-1/n}$.

Note to Example 42

- In this case

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

  and we can vary ψ to (numerically) solve

$$\mathrm{P}(T \le t \mid W = w; \psi_\alpha) = \alpha,$$

  thus giving limits for confidence intervals (approximate because the model is discrete).

Comments

- The essence of the recipe on slide 85 is to base an exact pivot $Q = q(Y; \psi)$ on a minimal sufficient statistic and use the **significance (or p-value) function**

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

  to invert Q and thus make inference on ψ using the quantiles q_p 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.

3 Likelihood Theory

slide 90

3.1 Basic Results

slide 91

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:
  - give (fairly) general setup;
  - prove main results for scalar parameter;
  - discussion of inference;
  - vector parameter, nuisance parameters, ...

Basic setup

- ☐ Let $Y, Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, and define the **Kullback–Leibler divergence** from the **data-generating model** g to a **candidate 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 holds because $-\log x$ is convex and is strict unless $f \equiv g$.
- ☐ 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

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

- ☐ $\theta_g = \mathrm{argmax}_\theta \mathrm{E}_g \log f(Y; \theta)$ gives the optimal theoretical 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).
- ☐ The natural estimator of θ_g is

$$\widehat{\theta} = \mathrm{argmax}_\theta\, \overline{\ell}(\theta),$$

  but we need conditions on $\overline{\ell}$ to ensure that $\widehat{\theta} \xrightarrow{P} \theta_g$ as $n \to \infty$.

Regular models

- **Regularity conditions** are needed ensure asymptotic consistency and normality of the MLE, such as
  - **(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

$$\imath_1(\theta) = \mathrm{E}_g\left\{-\nabla^2 \log f(Y;\theta)\right\}, \quad \hbar_1(\theta) = \mathrm{E}_g\left\{\nabla \log f(Y;\theta)\nabla^{\mathrm{T}} \log f(Y;\theta)\right\},$$

  are finite and positive definite. When $g = f_{\theta_g}$ we shall see that $\hbar_1(\theta_g) = \imath_1(\theta_g)$.
- Above $\nabla g(\theta) = \partial g(\theta)/\partial\theta$ and $\nabla^2 g(\theta) = \nabla\nabla^{\mathrm{T}} g(\theta) = \partial^2 g(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$.

Regularity conditions

- (C1) ensures that $\widehat{\theta}$ can be 'on all sides' of θ_g in the limit — if it fails, then a limiting normal distribution cannot arise;
- (C2) is essential for consistency, otherwise $\widehat{\theta}$ might not converge — it often fails in mixture models, for which care is needed;
- (C3) is a technical condition needed to bound terms of a Taylor series — can be replaced by a variety of other conditions, see for example van der Vaart (1998, *Asymptotic Statistics*, Chapter 5); and
- (C4) ensures that the asymptotic variance of $\widehat{\theta}$ is positive definite.

Consistency of the MLE

Lemma 43 *If $Y_1, \ldots, Y_n \sim g$ and $n \to \infty$, 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 concave (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.

Note to Lemma 43

- ☐ We prove this for θ scalar.
- ☐ As the θs correspond to different densities, precisely one θ_g 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 44 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, the regularity conditions hold and $n \to \infty$, then the sequence of consistent maximum likelihood estimators $\widehat{\theta}$ satisfies*

$$n^{1/2}(\widehat{\theta} - \theta_g) \xrightarrow{D} \mathcal{N}_d\{0, \imath_1^{-1}(\theta_g)\hbar_1(\theta_g)\imath_1^{-1}(\theta_g)\},$$

where for a single observation Y we define

$$\imath_1(\theta) = \mathrm{E}_g\left\{-\nabla^2 \log f(Y;\theta)\right\}, \quad \hbar_1(\theta) = \mathrm{E}_g\left\{\nabla \log f(Y;\theta) \nabla^{\mathrm{T}} \log f(Y;\theta)\right\}.$$

- ☐ This implies that for large n,

  $$\widehat{\theta} \dot{\sim} \mathcal{N}_d\{\theta_g, \imath^{-1}(\theta_g)\hbar(\theta_g)\imath^{-1}(\theta_g)\},$$

  where $\imath(\theta) = n\imath_1(\theta)$, $\hbar(\theta) = n\hbar_1(\theta)$ correspond to a sample of size n.
- ☐ This provides tests and confidence intervals based on the approximate pivots

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

  where v_{rr} are the diagonal elements of an estimate of $\imath^{-1}(\theta_g)\hbar(\theta_g)\imath^{-1}(\theta_g)$.
- ☐ When $g = f_{\theta_g}$, $\imath_1(\theta_g) = \hbar_1(\theta_g)$ and the variance (matrix) becomes $\imath(\theta_g)^{-1}$.

Note to Theorem 44

☐ 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

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

☐ 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)

$$\mathrm{var}\left\{n^{-1/2}\sum_{j=1}^n \nabla \log f(Y_j;\theta_g)\right\} = n^{-1}\sum_{j=1}^n \mathrm{E}_g\{\nabla \log f(Y_j;\theta_g)\nabla^{\mathrm{T}} \log f(Y_j;\theta_g)\} = \hbar_1(\theta_g).$$

so the numerator of (2) converges in distribution to $\mathcal{N}\{0, \hbar_1(\theta_g)\}$, using the CLT.

☐ Moreover the weak law of large numbers gives

$$-\nabla^2\overline{\ell}(\theta_g) = -\frac{1}{n}\sum_{j=1}^n \nabla^2 \log f(Y_j;\theta_g) \xrightarrow{P} \imath_1(\theta_g).$$

☐ Lemma **??** shows that $R_n \xrightarrow{P} 0$, so the denominator of (2) tends in probability to $\imath_1(\theta_g)$.

☐ Putting the pieces together, we find that

$$n^{1/2}(\widehat{\theta}-\theta_g) \xrightarrow{D} \mathcal{N}_d\{0, \imath_1(\theta_g)^{-1}\hbar_1(\theta_g)\imath_1(\theta_g)^{-1}\}, \quad n\to\infty,$$

where the variance formula is also valid when $\imath_1$ and $\hbar_1$ are $d\times d$ matrices.

☐ The information quantities based on a random sample of size n are $\imath(\theta_g) = n\imath_1(\theta_g)$ and $\hbar(\theta_g) = n\hbar_1(\theta_g)$, giving

$$\widehat{\theta} \dot{\sim} \mathcal{N}_d(\theta_g, \imath(\theta_g)^{-1}\hbar(\theta_g)\imath(\theta_g)^{-1}\},$$

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

Note: Secret Lemma

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

Note: A simpler cleaner argument

☐ Write

$$0 = \nabla\overline{\ell}(\widehat{\theta}) = \nabla\overline{\ell}(\theta_g) + \int_0^1 \nabla^2\overline{\ell}\{\theta_g + t(\widehat{\theta} - \theta_g)\}(\widehat{\theta} - \theta_g)\,\mathrm{d}t,$$

and note that $U_n = n^{1/2}\nabla\overline{\ell}(\theta_g) \xrightarrow{D} U \sim \mathcal{N}_d\{0, \hbar(\theta_g)\}$, so writing $Z_n = n^{1/2}(\widehat{\theta} - \theta_g)$ we have

$$U_n = -\int_0^1 \nabla^2\overline{\ell}(\theta_g + tn^{-1/2}Z_n)\,\mathrm{d}t\, Z_n,$$

and as $n \to \infty$ the integral here is approximately $\int_0^1 \nabla^2(\theta_g)\,\mathrm{d}t$, which converges in probability to $-\imath(\theta_g)$. Hence $Z_n \xrightarrow{D} Z = \imath(\theta_g)^{-1}U$, i.e.,

$$Z_n \xrightarrow{D} \mathcal{N}_d\{0, \imath(\theta_g)^{-1}\hbar(\theta_g)\imath(\theta_g)^{-1}\}, \quad n \to \infty.$$

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 saw earlier that under mild conditions the Bartlett identities give 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 $\imath(\theta) = \hbar(\theta)$, and $\imath(\theta) = n\imath_1(\theta) = n\hbar_1(\theta)$ when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$.
- ☐ Mathematically speaking the assumption that $g \in \mathcal{F}$ is always false, but
  - the asymptotic results are supposed to provide guidelines on what to expect when fitting models — checking the regularity conditions in practice would require knowledge of g, in which case there's no need for inference!
  - 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' — if the model is not quite right, then the interpretation of the crucial parameters will be unchanged.

In practice ...

- ☐ We usually assume classical asymptotics and replace the sandwich matrix $\imath(\theta_g)^{-1}\hbar(\theta_g)\imath(\theta_g)^{-1}$ by the inverse of the **observed information matrix**

$$\widehat{\jmath} = -\nabla^2 \ell(\widehat{\theta}),$$

  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 $\widehat{\jmath}^{-1}$.
- ☐ If we must make the sandwich we can replace $\imath(\theta_g)$ by $\widehat{\jmath}$ and $\hbar(\theta_g)$ by (e.g.)

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

  though $\widehat{\jmath}^{-1}\widehat{\hbar}\widehat{\jmath}^{-1}$ can be unstable because $\widehat{\hbar}$ misbehaves.

Related statistics

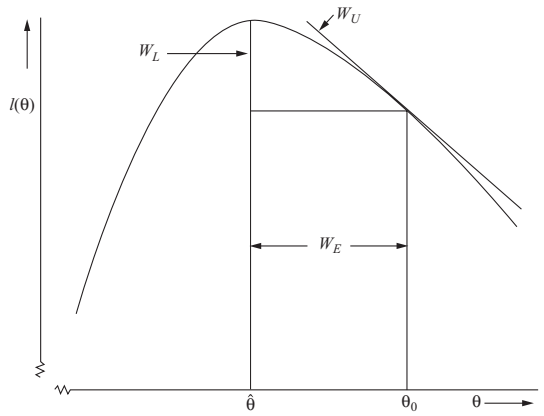


Figure 6.2. Three asymptotically equivalent ways, all based on the log likelihood function of testing null hypothesis $\theta = \theta_0$: W_E, horizontal distance; W_L vertical distance; W_U slope at null point.

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

Related statistics

- Classical asymptotics support inference for scalar θ based on any of the (approximate) pivots

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

  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 $(1-\alpha)$ **Wald 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.

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 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 $\jmath(\theta_g)$;
- ☐ 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* $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \exp(\theta)$ *and compare the resulting inferences with those from an exact pivot.*

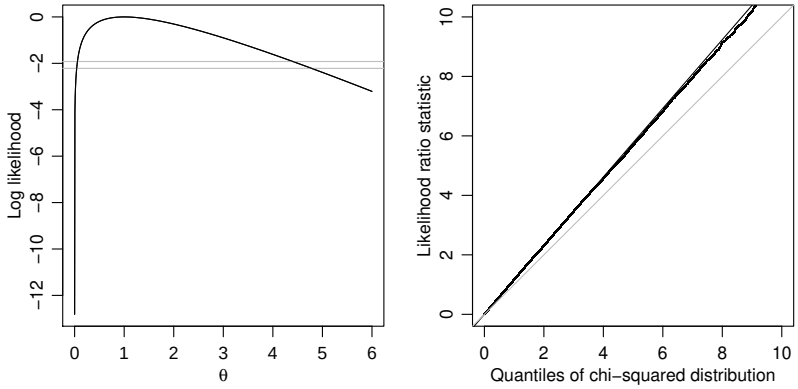
Note to Example 45

- ☐ 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

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

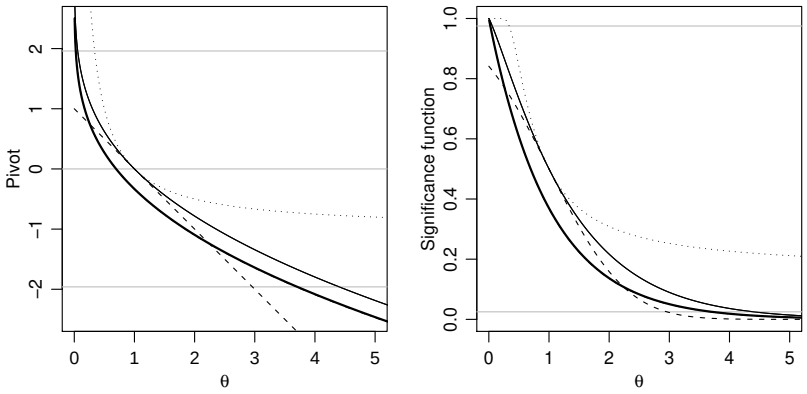
- ☐ The exact pivot is $\theta \sum Y_j$ whose distribution is gamma with unit scale and shape parameter n.
- ☐ Consider an exponential sample with $n = 1$ and $\overline{y} = 1$; then $\widehat{\jmath} = 1$. The log likelihood $\ell(\theta)$, shown in the left-hand panel of the figure, is unimodal but 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 confidence functions, with horizontal lines at 0.025 and 0.975.

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

Note to Example 46

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

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

Now

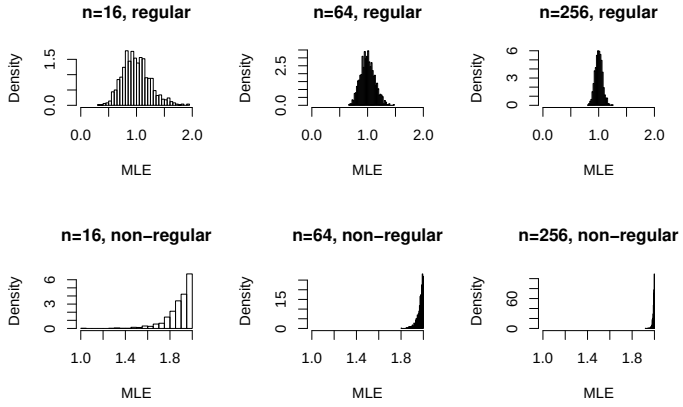
$$\mathrm{P}\left\{n(\theta - \widehat{\theta})/\theta \le x\right\} = \mathrm{P}(\widehat{\theta} \ge \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n \to 1 - \exp(-x),$$

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 46, 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.



3.2 Vector Parameter

Vector case

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

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

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 $s(\theta_g)^{\mathrm{T}} s(\theta_g) \dot{\sim} \chi^2_d$.

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

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

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

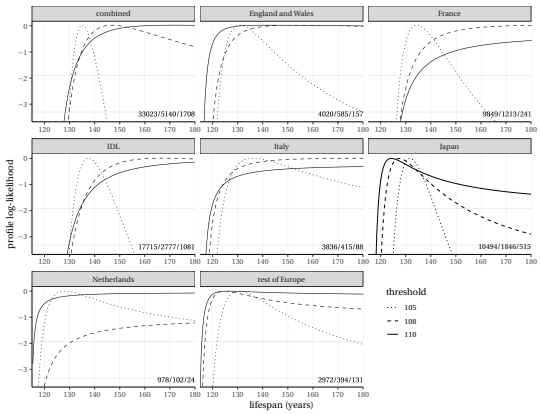
$$\begin{aligned}w_{\mathrm{p}}(\psi_g) &\doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g) - (\widehat{\lambda}_\psi - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda}_\psi - \lambda_g) \\ &\doteq (\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\psi}(\widehat{\psi}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

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

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

Example: Human lifespan

Example 47 *Profile log likelihoods for the endpoint ψ of a generalized Pareto model fitted to data on lifetimes of persons aged over 105 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 Review of Statistics and its Application*).

Model selection

- ☐ The fact that

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

  is minimised when $f = g$ suggested comparing competing models $\mathcal{F}_1, \ldots, \mathcal{F}_M$ by 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

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

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

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

$$2\left\{\ell(\widehat{\theta}) - \ell^+(\widehat{\theta})\right\} = 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} + 2\{\ell(\theta_g) - \ell^+(\theta_g)\} + 2\{\ \ell^+(\theta_g) - \ell^+(\widehat{\theta})\}, \tag{3}$$

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 (3) 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

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

☐ The second term on the right-hand side of (3) has expectation zero.

☐ The third term on the right-hand side of (3) 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

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

☐ Hence

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

If $\hbar(\theta_g) \doteq \imath_{\theta\theta}(\theta_g)$, 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* $\mathrm{NIC} = 2\{\mathrm{tr}(\widehat{\hbar}\widehat{\jmath}^{-1}) - \ell(\widehat{\theta})\}$, 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).

Effect of nuisance parameters

Example 48 (Neyman–Scott) *Find the profile log likelihood for σ^2 when $(y_{j1}, y_{j2}) \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, for $j = 1, \ldots, n$. Comment.*

- ☐ Profiling over many nuisance parameters can lead to completely wrong inferences, as the previous example shows.
- ☐ Even when the number of nuisance parameters is $o(n)$ we may run into trouble: in general

$$\text{Bias } (\widehat{\psi}; \psi) = O(d^3/n),$$

  so for the bias to tend to zero in large samples we require $d = o(n^{1/3})$ for consistency of $\widehat{\psi}$. Hence bias increases with $\dim(\lambda)$, at least in general.
- ☐ How can we rescue 'ordinary' likelihood inference when there are many nuisance parameters?

Note to Example 48

- ☐ The overall log likelihood is

$$\ell(\sigma^2, \mu_1, \ldots, \mu_n) \equiv -\tfrac{1}{2}\left[(2n)\log\sigma^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}\left\{(y_{j1}-\mu_j)^2 + (y_{j2}-\mu_j)^2\right\}\right],$$

  and differentiation with respect to μ_j gives that $\widehat{\mu}_j = (y_{j1} + y_{j2})/2$, so as

$$\{a - (a+b)/2\}^2 + \{b - (a+b)/2\}^2 = (a-b)^2/2,$$

  we obtain

$$\ell_{\text{p}}(\sigma^2) = -n\log\sigma^2 - \frac{1}{4\sigma^2}\sum_{j=1}^{n}(y_{j1} - y_{j2})^2, \quad \sigma^2 > 0.$$

- ☐ This is maximised at $\widehat{\sigma}^2_p = (4n)^{-1}\sum_{j=1}^{n}(y_{j1} - y_{j2})^2$, but as $Y_{j1} - Y_{j2} \overset{\text{iid}}{\sim} \mathcal{N}(0, 2\sigma^2)$, we see that $\sigma^2_p \xrightarrow{P} \sigma^2/2$ as $n \to \infty$; this is a completely inconsistent estimator. Hence the profile log likelihood has its asymptotic maximum in completely the wrong place.
- ☐ In this example there are $d = n + 1$ parameters of which n are nuisance parameters.

Dealing with nuisance parameters

- ☐ Approaches to dealing with high-dimensional 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 33 and 34) — 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**, via, e.g., a **modified profile likelihood** or a **modified likelihood root**, which can approximate both conditional and marginal likelihoods;
  - using **orthogonal parameters**, i.e., mapping $\lambda \mapsto \zeta(\lambda,\psi)$ which is orthogonal to ψ;
  - using a **composite likelihood** in which λ does not appear; or
  - 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.
- ☐ We have already seen examples of marginal and conditional likelihoods.
- ☐ Below we sketch some of the other approaches.

Modified profile likelihood

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

$$\ell_{\mathrm{mp}}(\psi) = \ell_{\mathrm{p}}(\psi) + m(\psi),$$

  with $m(\psi)$ chosen to make ℓ_{p} closer to a marginal or conditional log likelihood.
- ☐ Taking

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

  does this in some generality.
- ☐ The
  - first term of $m(\psi)$ can be obtained numerically if need be, but
  - the second term, a Jacobian needed to make ℓ_{mp} invariant to interest-preserving reparametrisation, 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, i.e., with ψ scalar

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

so tests and one-sided confidence sets

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

based on the observed data y^{o} have error $n^{-1/2}$.

- If we replace $r(\psi)$ by 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 for continuous responses the error drops to $O(n^{-3/2})$, so

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

so a one-sided confidence set

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

has error of order $n^{-3/2}$; often this almost exact even for tiny n (recall Example 45).

. . . with nuisance parameters

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

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

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,

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

- In general for independent continuous observations we write

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

where the $1 \times d$ vectors $V_j = \partial y_j/\partial\theta^{\mathrm{T}}$ 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 49 (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 49

- ☐ 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 parameters η and τ 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

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

  where we have set $\varepsilon_j^{\mathrm{o}} = (y_j^{\mathrm{o}}-\widehat{\eta}^{\mathrm{o}})/\widehat{\tau}^{\mathrm{o}}$.
- ☐ 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**.
- To see the effect of this, we expand the equation defining $\widehat{\lambda}_\psi$ around $\widehat{\theta}$, giving

$$\begin{aligned} 0 &= \frac{\partial \ell(\widehat{\theta}_\psi)}{\partial \lambda} = \frac{\partial \ell(\widehat{\theta})}{\partial \lambda} + \frac{\partial^2 \ell(\widehat{\theta})}{\partial \lambda \partial \theta^{\mathrm{T}}}(\widehat{\theta}_\psi - \widehat{\theta}) + \cdots \\ &= \frac{\partial^2 \ell(\widehat{\theta})}{\partial \lambda \partial \lambda^{\mathrm{T}}}(\widehat{\lambda}_\psi - \widehat{\lambda}) + \frac{\partial^2 \ell(\widehat{\theta})}{\partial \lambda \partial \psi^{\mathrm{T}}}(\psi - \widehat{\psi}) + \cdots \\ &= \widehat{\jmath}_{\lambda\lambda}(\widehat{\lambda}_\psi - \widehat{\lambda}) + \widehat{\jmath}_{\lambda\psi}(\psi - \widehat{\psi}) + \cdots \end{aligned}$$

  which implies that

$$\widehat{\lambda}_\psi = \widehat{\lambda} + \widehat{\jmath}_{\lambda\lambda}^{-1}\widehat{\jmath}_{\lambda\psi}(\widehat{\psi} - \psi) + \cdots$$

- Hence if we can arrange the model so that $\widehat{\jmath}_{\lambda\psi} \approx 0$, for example by parametrising it so that $\imath_{\lambda\psi}(\theta) \equiv 0$, then $\widehat{\lambda}_\psi$ will depend only weakly on ψ, and we can ignore the Jacobian term in the modified profile likelihood.
- This suggests mapping an original parametrisation (ψ, γ) to (ψ, λ), where $\lambda = \lambda(\psi, \gamma)$ is orthogonal to ψ.

Orthogonalisation

- 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

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

- In fact an explicit expression for λ in terms of ψ and γ is not needed to compute ℓ_{mp} in the new parametrisation.

Orthogonal parametrisation

- A solution (possibly numerical) 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 50 (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 50

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

$$\frac{\partial \gamma}{\partial \psi^{\mathrm{T}}} = -k_{\gamma\gamma}^{-1}(\psi, \gamma) k_{\gamma\psi}(\psi, \gamma). \tag{4}$$

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 (4). 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**

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

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), and
  - **empirical likelihood**, an extension of nonparametric modelling (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;
  - general (and 'optimal') approximate theory for inference in regular models;
  - close links to Bayesian inference (later).
- Weaknesses of likelihood approach:
  - requires 'parametric' model for data;
  - can fail in high-dimensional settings;
  - not all models are regular.

4 Hypothesis Testing

slide 124

4.1 Pure Significance Tests

slide 125

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+ ≥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:

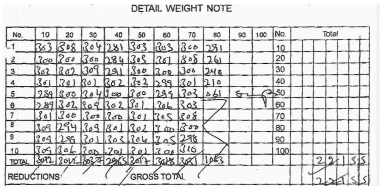
  H_0: the top quark does not exist.

  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 $\text{Poiss}(\lambda_0 = 6.7)$ and represents the baseline noise under H_0.
- We compute the **P-value**

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

  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?


DETAIL WEIGHT NOTE

No.	10	20	30	40	50	60	70	80	90	100	No.	Total
1	263	288	264	281	303	303	300	291			10	
2	300	287	269	284	305	291	288	261			20	
3	262	287	309	281	300	298	304	248			30	
4	261	291	301	302	292	299	301	210			40	
5	284	280	284	300	300	285	305	261			50	
6	248	282	284	302	291	284	303				60	
7	281	290	299	298	291	295	298				70	
8	305	294	289	281	302	290	279				80	
9	293	296	291	302	304	295	298				90	
10	305	286	290	294	291	298	310				100	
TOTAL	2891	2904	2892	2925	2973	2948	2981	1185				
REDUCTIONS					GROSS TOTAL							

- $n = 92$ weighings of sacks on the 'delivery' (or not?) of a commodity:

```
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?

Pearson's statistic

Definition 51 *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 $\mathrm{E}(O_k) = E_k > 0$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} E_k = n$, then* **Pearson's statistic (aka the 'χ^2 statistic')** *is*

$$T = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}.$$

- If

$$(O_1, \ldots, O_K) \sim \mathrm{Mult}\{n, (p_1 = E_1/n, \ldots, p_K = E_K/n)\},$$

  then $T \dot{\sim} \chi^2_{K-1}$ (approximation OK if average $E_k \geq 5$), giving a test of whether data $O_1, \ldots, O_K$ agree with specified probabilities $p_1, \ldots, p_K$.
- Here Benford's law suggests all $p_k \doteq 1/10$, so take $E_k = 92/10 = 9.2$.
- For the original dataset we found $t_{\mathrm{obs}} = 158.2$ and hence

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

  which is essentially impossible for uniformly distributed digits.
- Very strong evidence for industrial 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 T is continuous and H_0 is true, then we can treat p_{obs} as 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**.

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 or marginalisation.

Example 52 *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 53 *Find an exact test on a canonical parameter in a logistic regression model.*

Note to Example 54

- ☐ here $\overline{Y}$ and S^2 are minimal sufficient and independent, with $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and we can write $\overline{Y} \overset{\text{D}}{=} \mu + \sigma n^{-1/2} Z$ and $S^2 = \overset{\text{D}}{=} \sigma^2 V/(n-1)$, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ are independent. Hence

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

  is pivotal and thus allows tests on μ without reference to σ^2.
- ☐ For a test on σ^2 without regard to μ, we use the marginal distribution of $\S^2$, as $V = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ is a pivot.

Note to Example 53

- 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 34 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.

Comments

- 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

$$p_+ = \mathrm{P}_0(T \geq t_{\mathrm{obs}}), \quad p_- = \mathrm{P}_0(T \leq t_{\mathrm{obs}}), \quad p_{\mathrm{obs}} = 2\min(p_-, p_+),$$

  so $p_- + p_+ = 1 + \mathrm{P}_0(T = t_{\mathrm{obs}})$, which equals 1 unless T is discrete;
- We sometimes avoid minor problems due to discreteness by computing '**continuity-corrected**' P-values

$$p_+ = \sum_{t > t_{\mathrm{obs}}} \mathrm{P}_0(T = t) + \tfrac{1}{2}\mathrm{P}_0(T = t_{\mathrm{obs}}), \quad p_- = \sum_{t < t_{\mathrm{obs}}} \mathrm{P}_0(T = t) + \tfrac{1}{2}\mathrm{P}_0(T = t_{\mathrm{obs}}).$$

- So far we have described **pure significance tests**, where the situation if H_0 is false is not explicitly considered. We look at the effect of alternatives now.

4.2 Neyman–Pearson Approach

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 $\mathcal{Y}$ into

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

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

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

Link to confidence sets

- ☐ In a test on a parameter θ, with hypothesis $H_0 : \theta = \theta_0$ and corresponding size α critical region $\mathcal{Y}_1(\theta_0)$, we reject H_0 at level α if

$$p_{\text{obs}}(y; \theta_0) < \alpha \quad \Longleftrightarrow \quad y \in \mathcal{Y}_1(\theta_0).$$

- ☐ A $(1 - \alpha)$ confidence set $\mathcal{C}_{1-\alpha}$ for the 'true value' of θ, i.e., the value that generated the data, is the set of all values of θ_0 for which H_0 is not rejected at significance level α, i.e.,

$$\mathcal{C}_{1-\alpha} = \{\theta : p_{\text{obs}}(y; \theta) \geq \alpha\} = \{\theta : y \notin \mathcal{Y}_1(\theta)\}.$$

- ☐ This links hypothesis testing and confidence intervals, and enables construction of the latter in general settings, by this process of **test inversion**.

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 decision:

  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.
- 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** $\beta = \mathrm{P}_1(Y \in \mathcal{Y}_1)$.
- Note that the consequences of (e.g., losses due to) bad decisions are not taken into account.

Example 54 *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$*, with* σ^2 *known,* $H_0 : \mu = \mu_0$ *and* $H_1 : \mu = \mu_1$*, find the Type II error as a function of the Type I error.*

Note to Example 54

- 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.
- Although the form of $\mathcal{Y}_0$ is 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.

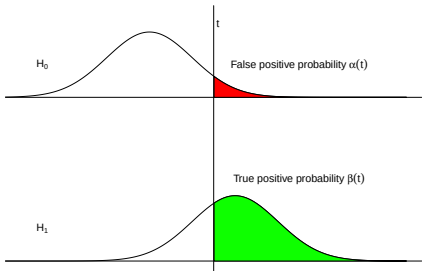
True and false positives: Example

- 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 54 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 55 *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**,

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

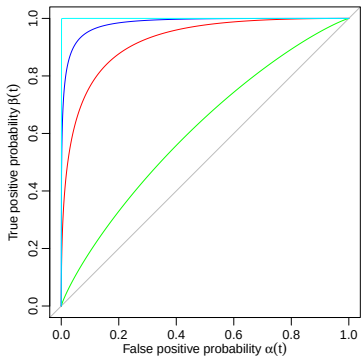
  which ranges between 0.5 for a useless test and 1.0 for a perfect test.

Example

☐ In Example 54 $\alpha(t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma\}$ 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 56 *A* **simple hypothesis** *entirely fixes the distribution of the data Y, whereas a* **composite hypothesis** *does not fix the distribution of Y.*

Definition 57 *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 58 (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 59 *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 58

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. If we write $F(\mathcal{C}) = \int_{\mathcal{C}} f(y)\,\mathrm{d}y$ for any density f with corresponding distribution F, then

$$\int_{\mathcal{Y}_1} f(y)\,\mathrm{d}y - \int_{\mathcal{Y}_1'} f(y)\,\mathrm{d}y = F(\mathcal{Y}_1) - F(\mathcal{Y}_1') \tag{5}$$

equals

$$F(\mathcal{Y}_1 \cap \mathcal{Y}_1') + F(\mathcal{Y}_1 \cap \mathcal{Y}_0') - F(\mathcal{Y}_1' \cap \mathcal{Y}_1) - F(\mathcal{Y}_1' \cap \mathcal{Y}_0) = F(\mathcal{Y}_1 \cap \mathcal{Y}_0') - F(\mathcal{Y}_1' \cap \mathcal{Y}_0) \tag{6}$$

where $\mathcal{Y}_0 \cup \mathcal{Y}_1 = \mathcal{Y}_0' \cup \mathcal{Y}_1' = \mathcal{Y}$.
If $F = F_0$, then (5) is non-negative, because $\alpha = F_0(\mathcal{Y}_1) \geq F_0(\mathcal{Y}')$, so (6) is also non-negative, giving

$$tF_0(\mathcal{Y}_1 \cap \mathcal{Y}_0') \geq tF_0(\mathcal{Y}_1' \cap \mathcal{Y}_0), \quad t \geq 0.$$

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

$$F_1(\mathcal{Y}_1 \cap \mathcal{Y}_0') \geq tF_0(\mathcal{Y}_1 \cap \mathcal{Y}_0') \geq tF_0(\mathcal{Y}_1' \cap \mathcal{Y}_0) \geq F_1(\mathcal{Y}_1' \cap \mathcal{Y}_0).$$

On adding $F_1(\mathcal{Y}_1 \cap \mathcal{Y}_1')$ to both sides we see that $F_1(\mathcal{Y}_1) \geq F(\mathcal{Y}_1')$, as required.

Note to Example 59

☐ 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 in Example 54 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 $\varphi_1 < \varphi_0$.

Discussion: 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;
  - even a tiny p_{obs} can support H_0 better than an alternative H_1 (e.g., $t_{\text{obs}} = 3$ 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, ...

Discussion: 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-specified alternative (e.g., covid testing);
  - model simplification if null hypothesis true;
  - 'dividing hypothesis' used to partition the parameter space into 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 60 *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.

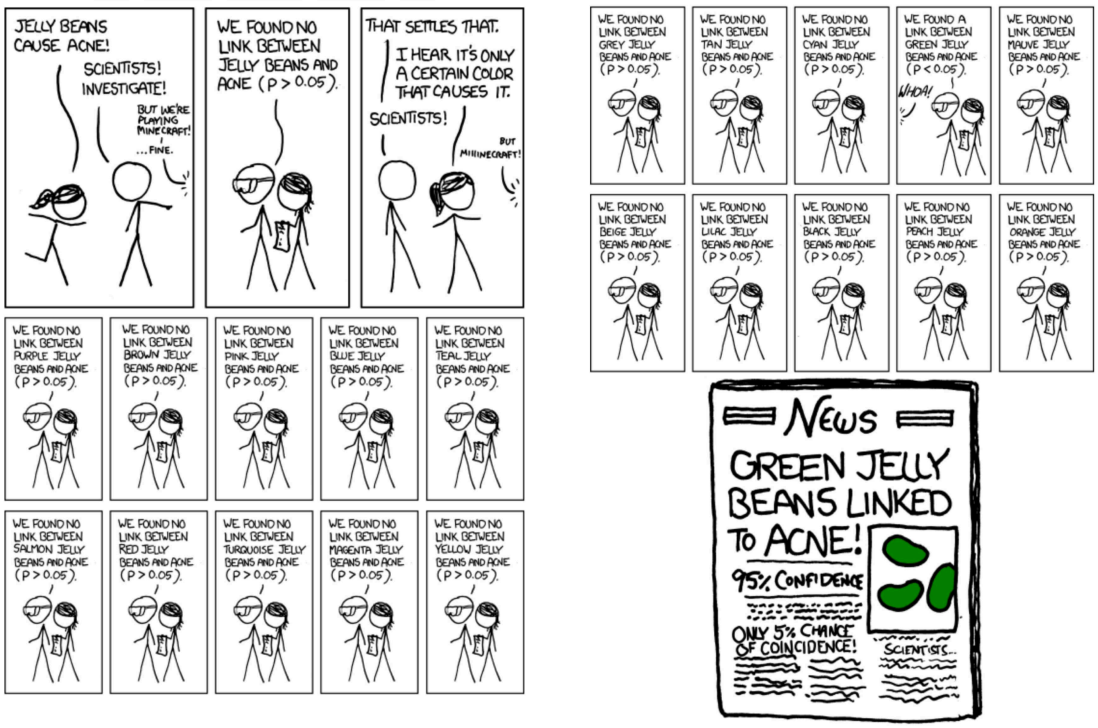
Motivation

- ☐ 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 can 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), cast 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 tail approximations; 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

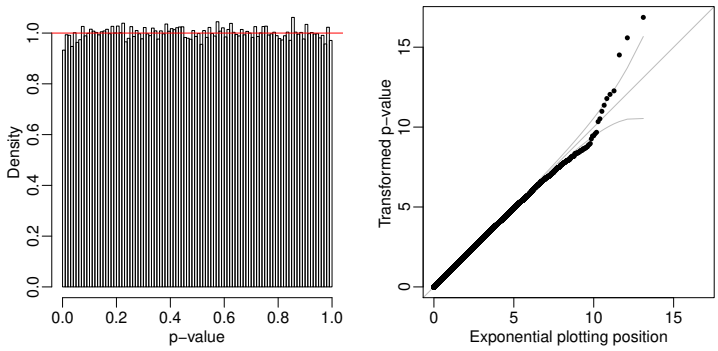
$$p_{\text{obs},j} = \text{P}_0(T_j > t_{\text{obs},j}) = 1 - \text{P}_0(-t_{\text{obs},j} \leq Y_j \leq t_{\text{obs},j}) \doteq 2\Phi(-t_{\text{obs},j}),$$

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 $\mathrm{FWER} \leq \alpha$.
- Control of the error rate:
  - **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 $\alpha \doteq 0.0026$ — 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 **Bonferroni inequality**) 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, because it uses higher rejection thresholds.

Note: Holm–Bonferroni procedure (HB)

- Recall that there are m hypotheses, of which m_0 are true nulls (for which $j \in \mathcal{I}$) and $m_1 = m - m_0$ are false nulls.
- If we apply HB and $\text{FP} \geq 1$, we must have wrongly rejected some H_j with $j \in \mathcal{I}$. If $H_{(s)}$ is the first such hypothesis to be rejected in the sequential procedure, then the $s-1$ hypotheses rejected before it must have been false null hypotheses, so $s - 1 \leq m_1 = m - m_0$, i.e., $m_0 \leq m + 1 - s$.
- As $H_{(s)}$ was rejected, the corresponding P-value satisfies

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

  Thus 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 Boole's inequality gives

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

- The only assumption needed above was that the null P-values are $U(0,1)$ (used in Boole's inequality), so HB strongly controls the FWER.

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

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

  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 the number of rejected hypotheses R satisfies

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

where $\{R_{-1} = r-1\}$ is the event that there are exactly $r-1$ rejections among $H_2, \ldots, H_m$. 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 $R_{-1} \in \{0, 1, \ldots, m-1\}$.

☐ Hence (under the conditions above) the Benjamini–Hochberg procedure strongly controls the FDR.

☐ Note that

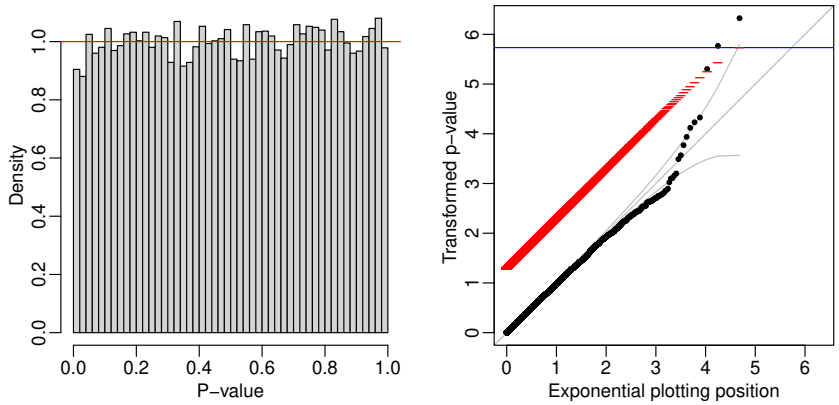
- if $m_0 \ll m$, then the last inequality may be very unequal, so possibly FDR $\ll \alpha$.
- if the P-values are dependent in such a way that

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

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. The red line shows the $U(0, 1)$ density.
- Right: exponential Q-Q plot of $Z_j = -\log Q_j$, with Bonferroni cutoff (blue) and Benjamini–Hochberg cutoffs (red), both with $\alpha = 0.05$. The grey lines are the target and pointwise 95% confidence sets for the order statistics.



Comments

- The Holm–Bonferroni procedure (HB) compares $P_{(1)}, P_{(2)}, \ldots$ to $\alpha/m, \alpha/(m-1), \ldots$, whereas the ordinary Bonferroni procedure (B) 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 Holm–Bonferroni procedure.
- The Benjamini–Hochberg procedure (BH) strongly controls the false discovery rate, comparing the ordered P-values to $\alpha/m, 2\alpha/m, \ldots, \alpha$.
- HB and B also give strong control when the P-values are dependent. So does BH, taking

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

  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.

Selection effects

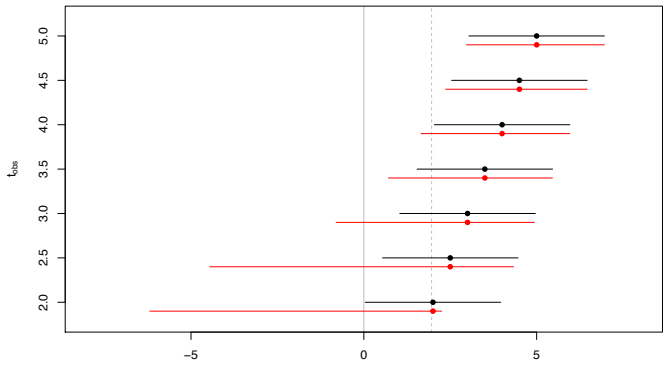
- ☐ Contrast
  - **exploratory analysis**, where we study data with no strong prior hypotheses, aiming to find something 'interesting' for future study, and
  - **confirmatory analysis**, where we specify an analysis protocol (hypotheses/tests/...) in advance and stick to it.
- ☐ Most statistical procedures assume we are doing the second, but there can be a strong temptation to cheat and treat an exploratory analysis as confirmatory.
- ☐ In 'the garden of forking paths' we make a series of choices (which response? transformation? which explanatory variables? ...) but do not then allow for them.
- ☐ This leads to non-reproducible results, 'false discoveries', bad science ...
- ☐ If we compute a confidence interval $\mathcal{I}$ for θ following a sequence of choices summarised in a selection event S that is *based on the data*, and compute

$$\mathrm{P}(\theta \in \mathcal{I}) \quad \text{when we should compute} \quad \mathrm{P}(\theta \in \mathcal{I} \mid S),$$

we are effectively pretending that S did not exist.

Simple example

Example 61 *Suppose $T \sim \mathcal{N}(\theta, 1)$ and we perform a two-sided test of $H_0 : \theta = 0$ at level $\alpha = 5\%$ and then construct a 95% confidence interval around the observed t_{obs} if we reject H_0. Compare the resulting confidence intervals when we do and do not allow for selection.*



95% confidence intervals for θ without (black) and with (red) allowance for selection on event $S = \{T > z_{0.975}\}$.

Note to Example 61

- Recall the basis of confidence intervals for θ based on an estimator T satisfying $T \sim \mathcal{N}(\theta, 1)$. We use the fact that $T - \theta \sim \mathcal{N}(0, 1)$ to argue that

$$\mathrm{P}(T \leq t_{\mathrm{obs}}) = \mathrm{P}(T - \theta \leq t_{\mathrm{obs}} - \theta) = \Phi(t_{\mathrm{obs}} - \theta)$$

  and then set this equal to $\alpha, 1 - \alpha$ to obtain the $(1 - 2\alpha)$ confidence interval $(t_{\mathrm{obs}} - z_{1-\alpha}, t_{\mathrm{obs}} - z_{\alpha})$, which reduces to the 95% confidence interval $t_{\mathrm{obs}} \pm 1.96$ when $\alpha = 0.025$.
- If we condition on the selection event that $T > z_{1-\beta}$ and, if this event occurs, compute the 95% confidence interval for θ, we are effectively using the conditional distribution

$$\begin{aligned} \mathrm{P}(T \leq t_{\mathrm{obs}} \mid T > z_{1-\beta}) &= \mathrm{P}(T - \theta \leq t_{\mathrm{obs}} - \theta \mid T - \theta > z_{1-\beta} - \theta) \\ &= \frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} \end{aligned}$$

  and the $(1 - 2\alpha)$ interval for θ has as endpoints the solutions of the equations

$$\frac{\Phi(t_{\mathrm{obs}} - \theta) - \Phi(z_{1-\beta} - \theta)}{1 - \Phi(z_{1-\beta} - \theta)} = \alpha, 1 - \alpha.$$

- If we set $\beta = 0.025$ and $\alpha = 0.025$, then we get the limits shown in the graph, which shows that even having $t_{\mathrm{obs}} = 3$ still leads to a 95% CI that contains 0 when we allow for selection. Hence making allowance for selection can radically change inferences, especially when H_0 is only just rejected.

Implications

- Need to be aware of possibility of selection effects and to read the literature critically.
- Must be clear if a study is exploratory or confirmatory:
  - if confirmatory, need to clarify protocol for inference **beforehand**;
  - if exploratory, need to avoid (any?) conclusions that might be due to 'forking paths'.
- Active area of research, likely to change in next few years.

5 Bayesian Statistics

slide 158

5.1 Introduction

slide 159

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 all uncertainty in distributional terms.

Example

Example 62 *(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 62

☐ 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,

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1, \quad a, b > 0, \tag{7}$$

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

$$\begin{aligned} \pi(\theta \mid y) &= \frac{\theta^{s+a-1}(1-\theta)^{n-s+b-1}/B(a,b)}{\int_0^1 \theta^{s+a-1}(1-\theta)^{n-s+b-1}\,\mathrm{d}\theta/B(a,b)} \\ &\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1. \end{aligned} \tag{8}$$

As (7) has unit integral for all positive a and b, the constant normalizing (??) must be $B(a+s, b+n-s)$. Therefore

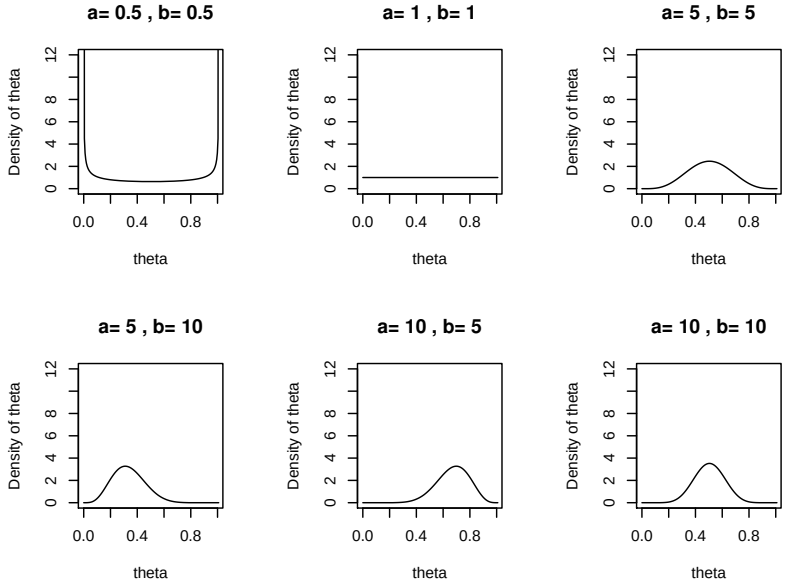
$$\pi(\theta \mid y) = \frac{1}{B(a+s, b+n-s)}\theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1.$$

☐ 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.

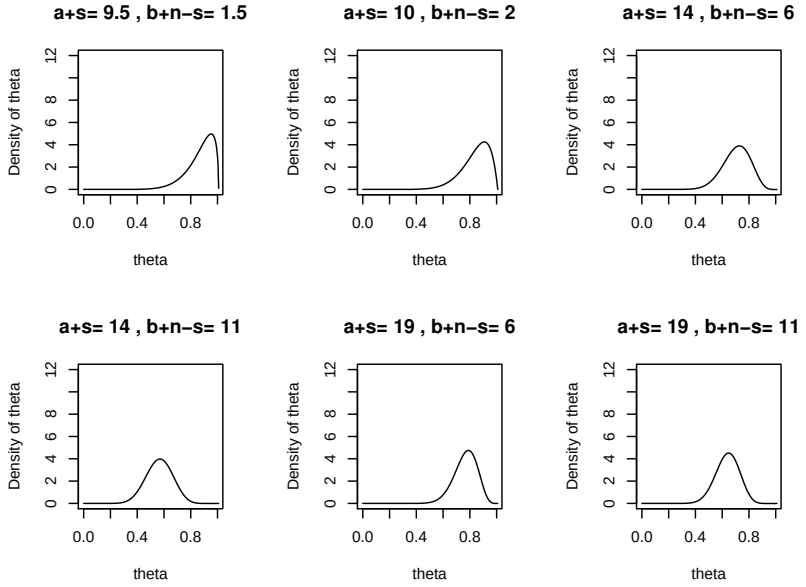
100 spins of a 5Fr coin

```
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
```

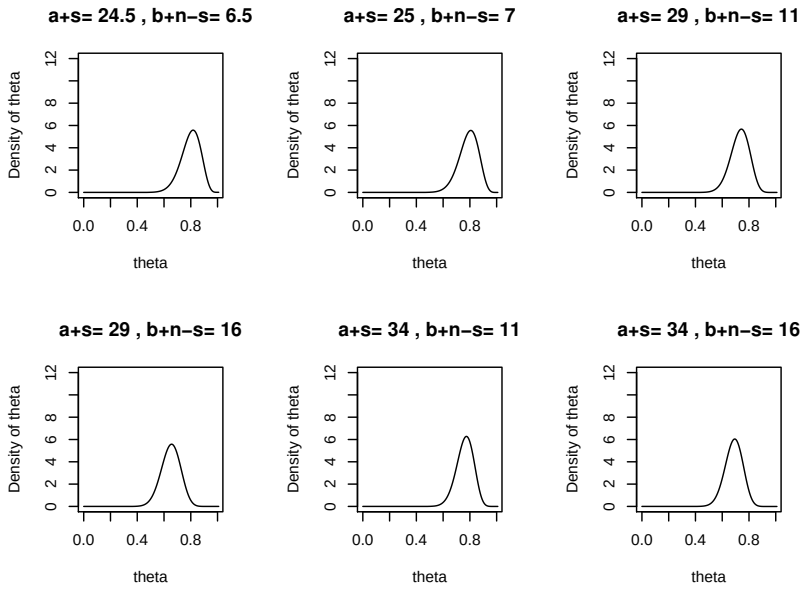

Beta prior densities



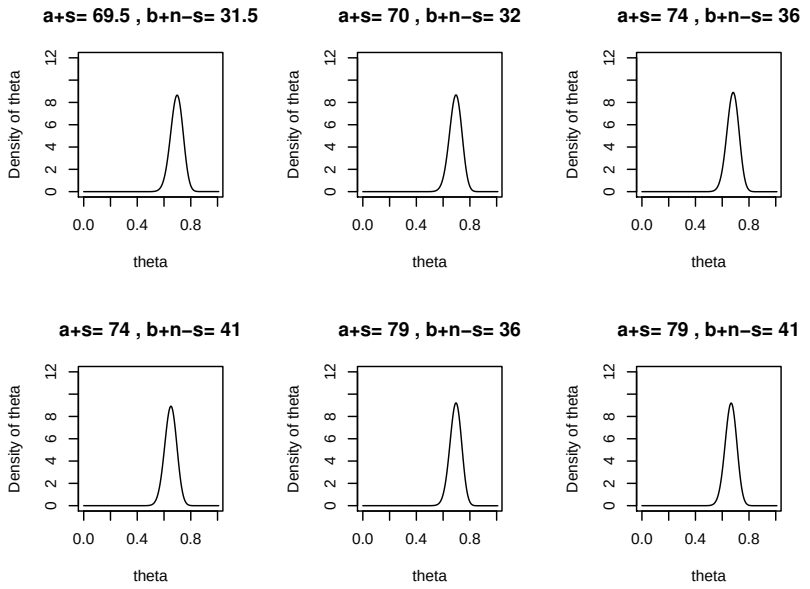
$n = 10,\ s = 9$



$n = 30,\ s = 24$



$n = 100,\ s = 69$



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.

Lemma 63 *An exponential family density*

$$f(y \mid \theta) = m(y) \exp[s(y)\varphi(\theta) - k\{\varphi(\theta)\}], \quad y \in \mathcal{Y}, \theta \in \Theta,$$

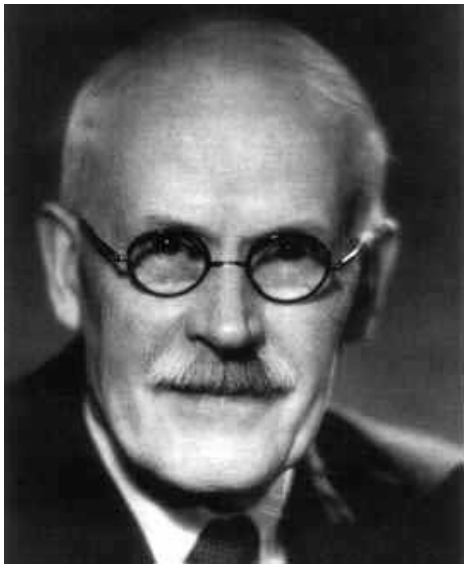
has conjugate prior

$$f(\theta; a, b) = h(a, b) \exp[a\varphi(\theta) - bk\{\varphi(\theta)\}], \quad \theta \in \Theta,$$

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 64

- ☐ *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 65 *What does a uniform prior for* $\theta \in (0, 1)$ *imply for* $\psi = \log\{\theta/(1-\theta)\} \in \mathbb{R}$*?*

Lemma 66 *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 65

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 66

- 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 67 (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 67

- 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 68 *The* **posterior α quantile** $\theta^\alpha(y)$ *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 a 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 $\widehat{\sigma}^2/n = \widehat{j}^{-1}$ 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 had

$$\mathrm{P}_{Y|\theta}\left\{\theta^\alpha(Y) \ge \theta\right\} = \int I\{\theta^\alpha(y) \ge \theta\} f(y \mid \theta)\,\mathrm{d}y = \alpha,$$

  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.

Matching priors II

- For scalar θ it turns out that the Jeffreys prior $\pi(\theta) \propto |\imath(\theta)|^{1/2}$
  - is matching to order n^{-1},
  - but higher-order matching is possible only in special cases.
- In the vector case, inferences for an interest parameter ψ match to order n^{-1}
  - if ψ is orthogonal to the other parameters λ, and

$$\pi(\psi, \lambda) \propto \imath_{\psi\psi}^{1/2}(\psi, \lambda) \times g(\lambda),$$

  - but in general it is impossible to match for all parameters simultaneously—would 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**.

Edgeworth series

We use asymptotic approximations to compare the Bayesian and frequentist solutions.

Definition 69 *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.

5.2 Bayesian Inference

Inference

Once we have a prior, what about

☐ point estimates?

☐ confidence sets?

☐ prediction?

☐ hypothesis tests?

☐ model comparison?

☐ model checking?

Point estimation

- Bayesian analysis yields a joint posterior distribution over the unknowns (parameters, predictands, ...), but this can be unwieldy, and simple summaries are often needed.
- A **Bayes estimator** $\tilde{u}(y^{\mathrm{o}})$ of an unknown u results from minimising a **posterior expected loss**,

$$\tilde{u}(y^{\mathrm{o}}) = \operatorname{argmin}_{\tilde{u}} \mathrm{E}\left\{L(u, \tilde{u}) \mid y^{\mathrm{o}}\right\} = \operatorname{argmin}_{\tilde{u}} \int L(u, \tilde{u}) \pi(u \mid y^{\mathrm{o}}) \, \mathrm{d}u,$$

  where the **loss function** $L(u, \tilde{u}) \geq 0$ measures the loss when u is estimated by $\tilde{u}$.
- The loss functions

$$(\tilde{u} - u)^2, \quad |\tilde{u} - u|,$$

  lead to the posterior mean $\mathrm{E}(u \mid y^{\mathrm{o}})$ and median of u.
- Another common estimator, the **maximum a posteriori (MAP)** estimator

$$\tilde{u} = \operatorname{argmax}_u \pi(u \mid y^{\mathrm{o}}),$$

  is not a Bayes estimator in general. It is superficially similar to the MLE, but is not invariant to parameter transformation because of the appearance of a Jacobian.

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.

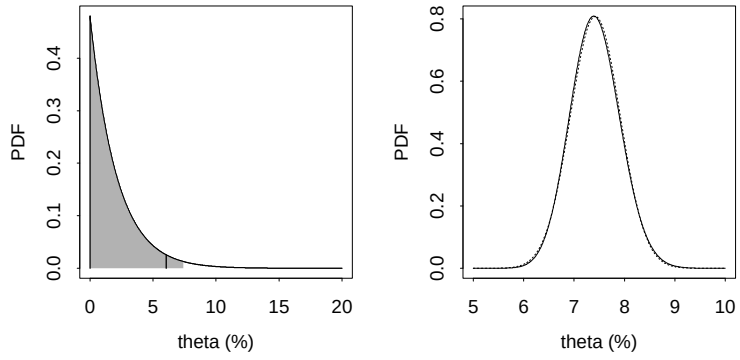
A	0/47	B	18/148	C	8/119	D	46/810	E	8/211	F	13/196
G	9/148	H	31/215	I	14/207	J	8/97	K	29/256	L	24/360

Example 70 (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 MAP 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

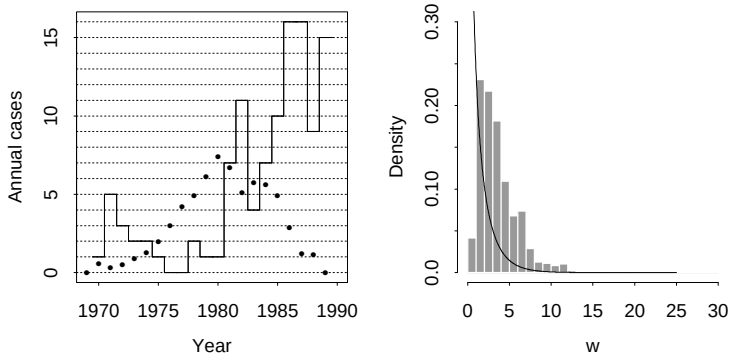
- As $B_{10} = B_{01}^{-1}$, the evidence for H_0 is $2\log B_{01} = -2\log B_{10}$.
- Models $f(y \mid H, \theta)$ for n observations and $d \times 1$ parameter θ often compared using

$$\mathrm{BIC} = -2\ell(\widehat{\theta}) + d\log n,$$

  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 71 (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 (residuals, ...).
- Another measure of plausibility based on possible new dataset $Y_+ \sim f$ is

$$\mathrm{P}\{f(Y_+) \leq f(y^{\mathrm{o}})\},$$

  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 ...

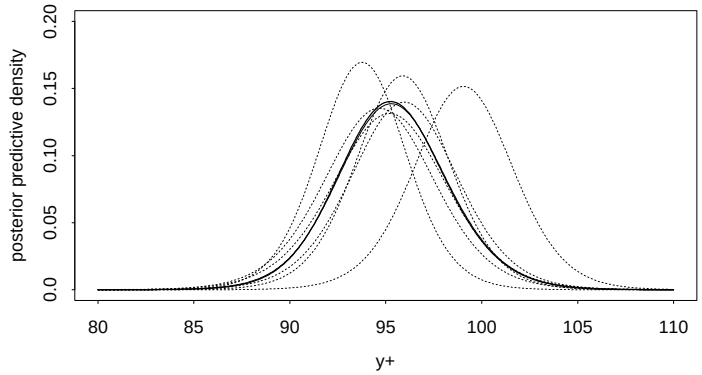
Example

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 . . .

5.3 Bayesian Computation

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 (other courses . . .).

Laplace's method

Lemma 72 *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 72

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 73 *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 74 *Use Laplace approximation to derive the Bayesian information criterion.*

Note to Example 74

- ☐ 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 75 *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

$$r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right), \quad r_0 = \mathrm{sign}(u_0 - \tilde{u})\{2g(u_0)\}^{1/2}, \quad v_0 = \frac{g'(u_0)}{a(u_0)}.$$

Example 76 *Use the methods above to approximate the posterior conditional distribution*

$$\mathrm{P}(\theta \leq \theta_0 \mid y)$$

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 75

☐ 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

$$\begin{aligned} J_n(u_0) &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0^*} e^{-nr^{*2}/2} \left\{1 + O(n^{-1})\right\} dr^* \\ &= \Phi(n^{1/2} r_0^*) + O(n^{-1}), \end{aligned} \tag{9}$$

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 76

☐ 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 75 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 77 *Let $u = (u_1, u_2)$, where u_1 is scalar and u_2 a $p \times 1$ vector, and consider*

$$J_n(u_1^0) = (2\pi)^{-(p+1)/2} c \int_{-\infty}^{u_1^0} \int \exp\{-ng(u_1, u_2)\} \, \mathrm{d}u_2 \mathrm{d}u_1, \tag{10}$$

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 76 can be extended to the multiparameter case using Lemmas 73 and 77, and give

  $$\mathrm{P}(\psi \leq \psi_0 \mid y) = \Phi\{r_B^*(\psi_0)\} \left\{1 + O(n^{-1})\right\},$$

  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 116–118 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

$$\exp\{\log f(y; \theta) + \log \pi(\theta)\},$$

  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 — see Appendix I (and other courses) for a summary of basic MCMC.

5.4 Hierarchical Models

Exchangeability

- 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 is common in modelling complex data, in both classical and Bayesian frameworks.
- Some theoretical justification is provided by the notion of exchangeability: variables are exchangeable if there is no reason to distinguish them.

Definition 78 *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.*

De Finetti's theorem

Theorem 79 (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*

$$f(u_1, \ldots, u_n) = \int_0^1 \prod_{j=1}^{n} \theta^{u_j} (1-\theta)^{1-u_j} \, G(\mathrm{d}\theta) \tag{11}$$

where

$$G(\theta) = \lim_{m\to\infty} \mathrm{P}\left\{m^{-1}(U_1 + \cdots + U_m) \le \theta\right\}, \quad \theta = \lim_{m\to\infty} m^{-1}(U_1 + \cdots + U_m).$$

- 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 continuous and other types of variables.
- 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 80 *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. 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 80

☐ 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

$$\begin{pmatrix}\mu\\ \theta\\ y\end{pmatrix} \sim \mathcal{N}_{2n+1}\left\{\mu_0 1_{2n+1}, \quad \begin{pmatrix}\tau^2 & \tau^2 1_n^{\mathrm{T}} & \tau^2 1_n^{\mathrm{T}}\\ \tau^2 1_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n\\ \tau^2 1_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n & V + \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n\end{pmatrix}\right\}, \tag{12}$$

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

$$\begin{pmatrix}\mu\\ \theta\end{pmatrix} \mid y \sim \mathcal{N}_{n+1}\left\{\mu_0 1_{n+1} + \Omega_{12}\Omega_{22}^{-1}(y - \mu_0 1_n), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right\}. \tag{13}$$

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 (12) 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 (13) 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 80

☐ 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 80, 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

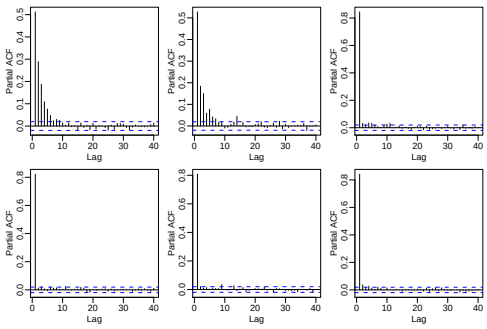
```
cardiac.gibbs <- function(data, mu0=0, a=10^(-3), b=10^(-3), c=10^3, R=10^5, d=1)
{ # parameter is mu, sig2, beta
  card.update <- function(data, mu0, a, b, c, para)
  {
    sig2 <- para[2]
    beta <- para[-c(1,2)]
    n <- length(beta)
    mu <- rnorm( 1, (mu0/c^2 + sum(beta)/sig2)/(1/c^2+n/sig2),sqrt(1/(1/c^2+n/sig2)) )
    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

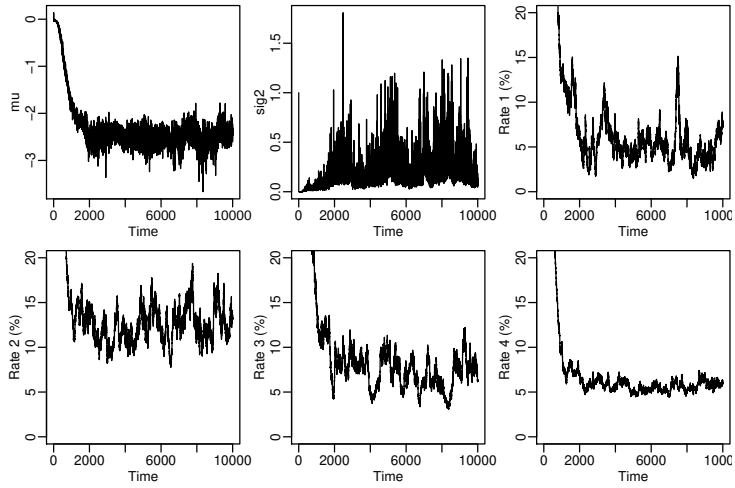
Acceptance probabilities for different values of d:

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 $d = 1$:

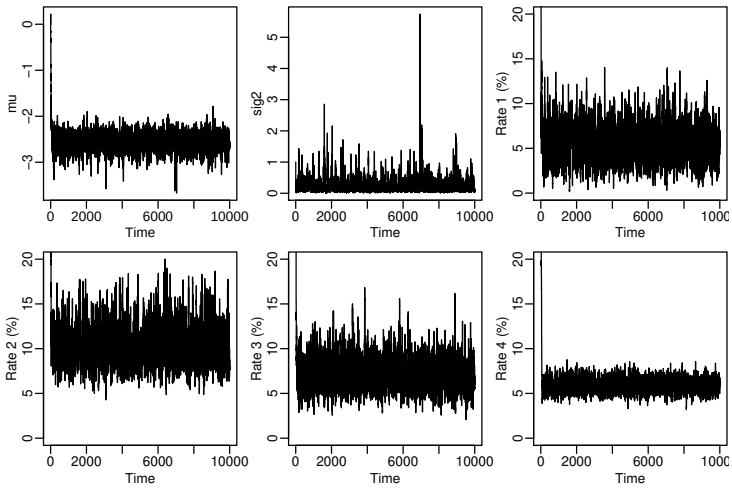


Effect of d: $d = 0.1$



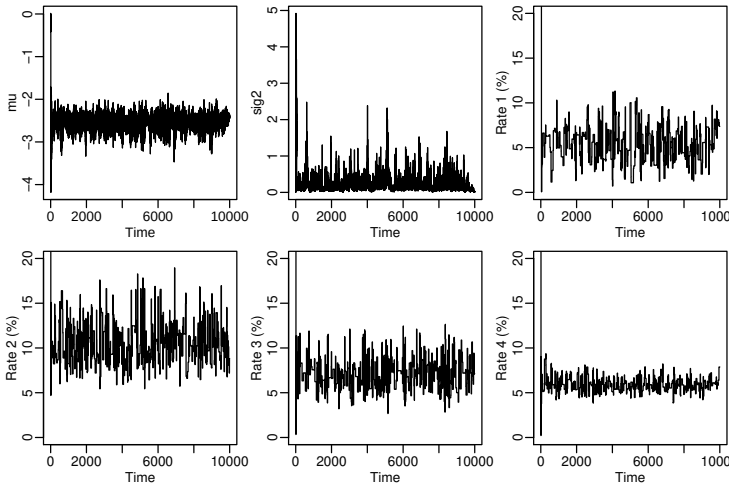
Taking $d = 0.1$ makes the acceptance probability too high, so the chain mixes too slowly.

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.

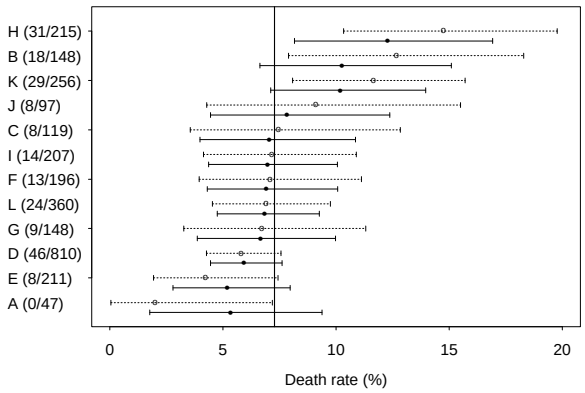
Effect of d: $d = 30$



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

- ☐ 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;
  - inappropriate when we are interested in certain pre-specified parameters or where prior knowledge distinguishes them;
  - an example of inappropriate use: economic modelling with countries of Europe treated as exchangeable.
- ☐ **Empirical Bayes** is also based on hierarchical models, but estimates key parameters — e.g., in Example 80 the parameters μ and σ^2 would be estimated (e.g., by maximising a marginal likelihood) rather than having a prior placed on them. Often this is more attractive than putting a prior at the top level of the hierarchy.
- ☐ Can be hard to count the number of parameters: the prior 'ties together' some parameters, so there are 'really' fewer—but how many?
- ☐ Graphical representation of dependence relations (hierarchical structure, ...) very useful — see Appendix II.

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 81

(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 82 (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 82

☐ 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

$$\begin{aligned}
\mathrm{P}(u \mapsto V \mid u) &= \mathrm{P}\left(W \leq \min[1, \exp\{\tfrac{1}{2}(u^2 - v^2)\}] \mid u\right) \\
&= \mathrm{P}(|V| \leq u \mid u) + \int_{\{v:|v|>|u|\}} e^{(u^2-v^2)/2}\frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(v-u)^2/2\sigma^2}\,du,
\end{aligned}$$

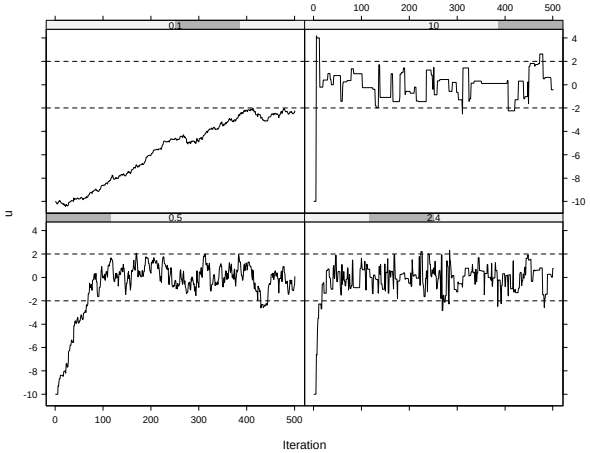
which clearly depends on u and on σ.

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 83 *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 83

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

$$\begin{aligned} \mu \mid \sigma, y &\sim \mathcal{N}\left(\frac{\sum y_j + \sigma^2\kappa\xi}{n + \kappa\sigma^2}, \frac{1}{n\sigma^{-2} + \kappa}\right), \\ \frac{1}{\sigma^2} \mid \mu, y &\sim \Gamma\left(\alpha + n/2, \beta + \sum(y_j - \mu)^2/2\right), \end{aligned}$$

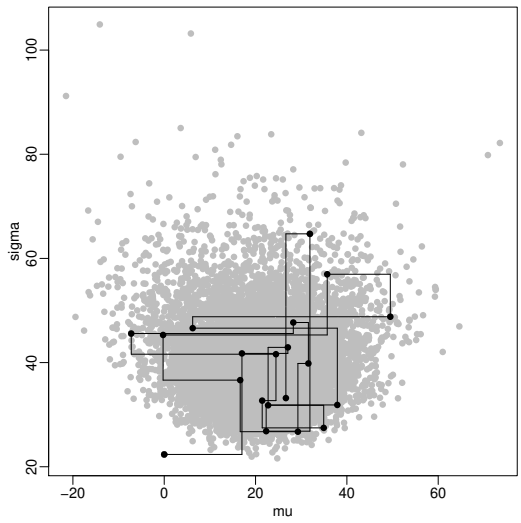
and the Gibbs sampler alternates updates of μ and of σ^{-2} using these two equations.

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])
```

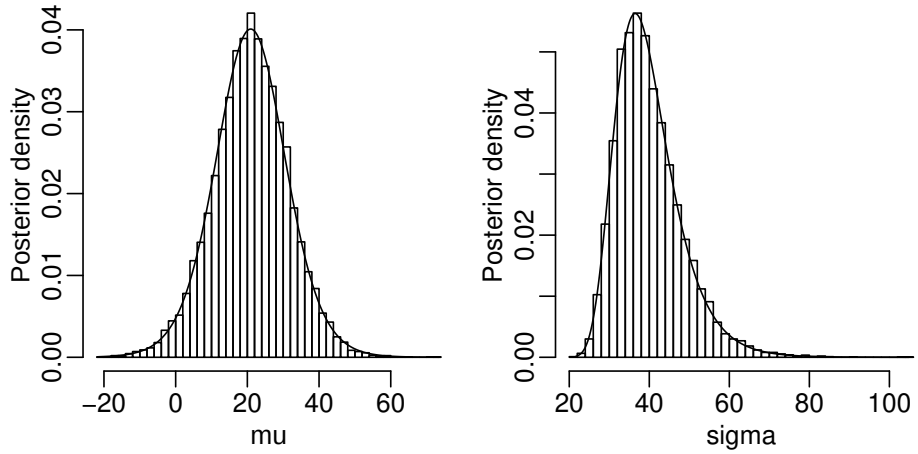

Toy Gibbs example

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:



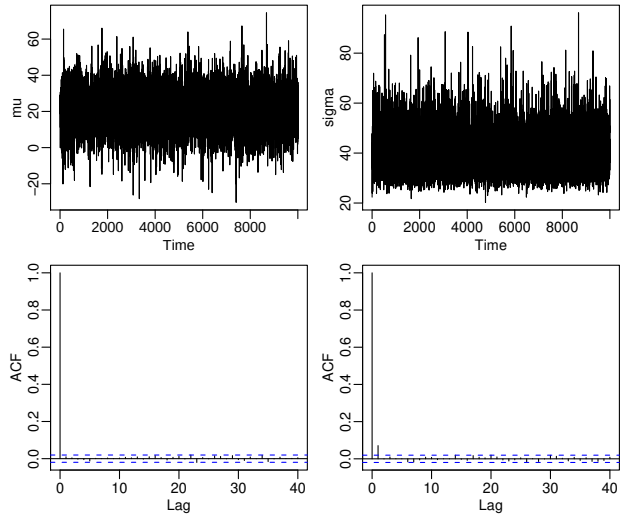
Toy Gibbs example

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.

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

      $$\text{(i)} \quad j \notin \mathcal{N}_j, \qquad \text{(ii)} \quad i \in \mathcal{N}_j \quad \Leftrightarrow \quad j \in \mathcal{N}_i,$$

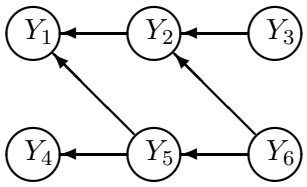
      and let $\tilde{\mathcal{N}}_j = \mathcal{N}_j \cup \{j\}$;
  - the set of nodes and the neighbourhood structure $(\mathcal{J}, \mathcal{N})$ define the graph.

Directed acyclic graphs

Definition 84 *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

$$\begin{aligned} &f(y_1 \mid y_{-1}) = f(y_1 \mid y_2, y_5), \quad f(y_2 \mid y_{-2}) = f(y_2 \mid y_1, y_3, y_5, y_6), \\ &f(y_3 \mid y_{-3}) = f(y_3 \mid y_2, y_6), \quad f(y_4 \mid y_{-4}) = f(y_4 \mid y_5), \quad \ldots \end{aligned}$$