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 new 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 $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$, $\forall x_1, x_2$.

Exchangeability

- Exchangeability is weaker than independence, often used to model variables that are indistinguishable in probabilistic terms, even if not independent.
- de Finetti proved that such variables must be constructed as $U_1, \ldots, U_n \mid \theta \overset{\text{iid}}{\sim} F_\theta$, where $\theta \sim G$ for distributions F_θ and G. The simplest theorem to this effect is the one below.

Definition 1 *Random variables $U_1, \ldots, U_n$ are **finitely exchangeable** if their density satisfies*

$$f(u_1, \ldots, u_n) = f\left(u_{\xi(1)}, \ldots, u_{\xi(n)}\right)$$

*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.*

Theorem 2 (de Finetti) *If $U_1, U_2, \ldots$, is an infinitely exchangeable sequence of binary variables taking values in $\{0, 1\}$, then for any n there is a distribution G such that*

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

where

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

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 $\mu + \sigma z_p$.

Order statistics

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

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

- 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

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

- 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 3 *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 $X_{(r)} \leq x$ 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 $\mathrm{P}(X_{(r)} \leq x)$, whereupon one finds that almost all the terms cancel, giving the stated density. A more easily generalised argument is as follows: for the event $X_{(r)} \in [x, x + \mathrm{d}x)$, we need to split the sample into three groups of respective sizes $r - 1$, 1 and $n - r$ and 'probabilities' $F(x)$, $f(x)\mathrm{d}x$, and $1 - F(x)$. The corresponding multinomial 'probability' is

$$\frac{n!}{(r-1)! \times 1! \times (n-r)!} \{F(x)\}^{r-1} \times f(x)\mathrm{d}x \times \{1 - F(x)\}^{n-r},$$

  and dropping the $\mathrm{d}x$ gives the density function of $X_{(r)}$.
- ☐ 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 3

- ☐ The joint density of $X_{(1)}$ and $X_{(n)}$ is given by splitting the total n observations 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 sizes 1, $n - 2$ and 1, giving

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

  We drop the $\mathrm{d}x_1\mathrm{d}x_n$ to get the joint density.
- ☐ 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 $x_1 < x < x_n$.

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.

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

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

- 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:

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

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 distributions and MGFs/KGFs (if the latter are non-trivial).
- 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 $\kappa_1 = \mathrm{E}(X)$, $\kappa_2 = \mathrm{var}(X)$, $\kappa_3 = \mathrm{E}\{(X - \mu_1)^3\}$, ...

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.,

$$f(y; \varphi) = f_0(y) \exp\{\varphi^{\mathrm{T}} s(y) - k(\varphi)\}, \quad y \in \mathcal{Y}, \varphi \in \mathcal{N},$$

where

$$\mathcal{N} = \{\varphi : k(\varphi) < \infty\}$$

and individual members of the family are determined by the value of φ.

- Hölder's inequality gives

$$M\{\alpha\varphi_1 + (1-\alpha)\varphi_2\} \leq M(\varphi_1)^{\alpha} M(\varphi_2)^{1-\alpha} < \infty, \quad 0 \leq \alpha \leq 1,$$

for any $\varphi_1, \varphi_2 \in \mathcal{N}$, so the set $\mathcal{N}$ and the function k are both convex.

- This implies that $f(y; \varphi)$ is log-concave in φ, which is very useful for statistics.
- This construction leads to an elegant general theory putting many well-known distributions (Poisson, binomial, normal, ...) under the same roof.

Example 4 *Investigate exponential tilting when $f_0(y)$ is uniform on $(0, 2\pi]$ with $s(y) = (\cos y, \sin y)^{\mathrm{T}}$.*

Note to Example 4

Here $\mathcal{Y} = (0, 2\pi]$ is finite, and $s(y)$ has dimension 2 and is bounded, so with $(\varphi_1, \varphi_2) \in \mathbb{R}^2$,

$$\begin{aligned} \int f_0(y) \exp\{\varphi^{\mathrm{T}} s(y)\} \, \mathrm{d}y &= \frac{1}{2\pi} \int_0^{2\pi} \exp(\varphi_1 \cos y + \varphi_2 \sin y) \, \mathrm{d}y \\ &= \frac{1}{2\pi} \int_0^{2\pi} \exp\{\theta_2 \cos(y - \theta_1)\} \, \mathrm{d}y = I_0(\theta_2), \end{aligned}$$

where $\theta_2 = (\varphi_1^2 + \varphi_2^2)^{1/2} \geq 0$, $\theta_1 = \tan^{-1}(\varphi_2/\varphi_1) \in (0, 2\pi]$, and $I_0(\theta_2)$ is a modified Bessel function of the first kind and order 0. Hence $\varphi_1 = \theta_2 \cos\theta_1$ and $\varphi_2 = \theta_2 \sin\theta_1$, and

$$k(\varphi) = \log I_0\{(\varphi_1^2 + \varphi_2^2)^{1/2}\}, \quad \varphi \in \mathcal{N} = \mathbb{R}^2.$$

This is the von Mises–Fisher distribution on the circle, which concentrates around θ_1, with the degree of concentration determined by $\theta_2 \geq 0$; $\theta_2 = 0$ gives the uniform density.

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

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

  then this is 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 $a \neq 0$ such that $a^{\mathrm{T}} S$ is constant, and call the model a **minimal representation** if there is no vector $a \neq 0$ 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 5 (Poisson sample) *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{Poiss}(\theta)$, *find the corresponding exponential family.*

Example 6 (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 5

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 6

- ☐ 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 $n = d = 2$, $\Omega = D^{-1}$ is diagonal, and with $\theta^{\mathrm{T}} = (\psi, \lambda)$ we have

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

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

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 7 (Satellite conjunction) *Show that taking ψ known in Example 6 gives a $(2, 1)$ exponential family.*

Note to Example 7

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.

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 $\mathrm{E}\{g(X)\} \geq g\{\mathrm{E}(X)\}$.

Modes of convergence

- Let $X, X_1, X_2, \ldots$ have CDFs $F, F_1, F_2, \ldots$ and let $\varepsilon > 0$ be arbitrary. Then
  - X_n converges to X **almost surely**, $X_n \xrightarrow{\text{a.s.}} X$, if $\mathrm{P}(\lim_{n\to\infty} X_n = X) = 1$;
  - X_n converges to X **in probability**, $X_n \xrightarrow{P} X$, if $\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) = 0$;
  - 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 **strongly consistent** if $X_n \xrightarrow{\text{a.s.}} \theta$ and **(weakly) consistent** if $X_n \xrightarrow{P} \theta$.
- $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{P}$, but not $\xrightarrow{D}$, require joint distributions of (X_n, X) for every n.
- Let x_0, y_0 be constants, $X, Y, \{X_n\}, \{Y_n\}$ be random variables and $g(\cdot)$ and $h(\cdot,\cdot)$ continuous functions. Then

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

  The last two lines are called the **continuous mapping theorem** (usually used with $\xrightarrow{P}$) and **Slutsky's theorem**.

Limit theorems

Theorem 8 (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 9 (Strong law of large numbers, SLLN) *If* $X, X_1, X_2, \ldots \overset{\text{iid}}{\sim} F$ *and* $\mathrm{E}(X)$ *is finite, then* $\overline{X} = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{\text{a.s.}} \mathrm{E}(X)$.

Theorem 10 (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 11 ('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 $\overline{X} \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$, 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 interest parameters;
  - secondary aspects can be accounted for, often via nuisance parameters;
  - variation in the data is modelled well enough to give realistic assessments 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

- Vectors are always column vectors, with row vectors denoted using the transpose $^{\mathrm{T}}$.
- 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, and
  - α is mostly reserved for significance levels.
- 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. Hence if $g(\varphi)$ is a scalar function of a $d \times 1$ parameter φ, then $\nabla g(\varphi)$ is a $d \times 1$ vector and $\nabla^2 g(\varphi)$ is a $d \times d$ matrix, and if $h(\varphi)$ is a $n \times 1$ vector function of φ, then $\nabla h^{\mathrm{T}}(\varphi)$ is a $d \times n$ matrix and $\nabla^{\mathrm{T}} h(\varphi)$ is an $n \times d$ matrix.
- 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

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

  Often $\tilde{\theta}$ also solves

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

  and is then called a **Z**(ero)**-estimator**.
- Equivalently we could minimise the **loss function** $-\rho$ with respect to θ.
- If the true underlying model is g, then $\tilde{\theta}$ is replaced by θ_g, where

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

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

Examples

- 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

$$-\rho(y; \theta) = \|\nabla_y \log f(y; \theta) - \nabla_y \log g(y)\|_2^2.$$

- There are many (many!) other approaches to estimation.

Examples

Example 12 *Discuss maximum likelihood estimation of the parameters of the normal distribution.*

Example 13 *Discuss moment estimation of the parameters of the Weibull distribution.*

Example 14 *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 12

☐ 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 second derivatives, which give the Hessian matrix, 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 negative 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 13

☐ 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 14

☐ Score-matching 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' in theoretical statistics is for the derivative of the log likelihood with respect to θ (not y).

☐ On writing $\log f(y;\theta) = \ell(\theta)$ for brevity and supposing that y is scalar, we can write

$$\|\nabla_y \log f(y;\theta) - \nabla_y \log g(y)\|_2^2 = \{\nabla_y \ell(\theta)\}^2 - 2\nabla_y \ell(\theta)\nabla_y \log g(y) + \{\nabla_y \log g(y)\}^2,$$

and see that the population version of the estimator is

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

because θ does not appear in the third term of the square. Now g is unknown, so the second integral here appears intractable, but as $g(y)\nabla_y \log g(y) = \nabla_y g(y)$, we have

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

and integration by parts gives

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

when (if!) the first integration term is identically zero. 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,

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

can be computed from the sample.

☐ Weighted versions can be used to kill the first term of the integral, when it is non-zero (exercise).

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_g when $n \to \infty$?

Definition 15 *An estimator $\tilde{\theta}$ of θ_g is* **(weakly) consistent** *if $\tilde{\theta} \xrightarrow{P} \theta_g$ as $n \to \infty$.*

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

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

  but $\tilde{\theta}^*$ is (probably) useless: consistency can be considered a 'safety net'.
- Obviously we would like $\tilde{\theta}$ to be 'suitably close' to θ_g, by minimising

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

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

Bias-variance and other tradeoffs

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

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

  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., $b(\tilde{\theta}; \theta_g) \equiv 0$, 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, in which case we want them to estimate the same thing!

Example 16 *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 16

- ☐ 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 $g(x) = \mu^{-1}(x)$ and $\mu = \mu(\theta)$ 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 we have

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

where $\mu = \mu(\theta)$ and $Z_n' \in (0, Z_n)$, 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 17 *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**,

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

where $\imath(\theta)$ is defined on the next slide. This bound applies for any sample size n. Moreover

  - as $n \to \infty$ the lower bound $\to 1/\imath(\theta)$, 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**:

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

  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 $\mathrm{var}\{U(\theta)\} = \imath(\theta)$ is the $d \times d$ **Fisher (or expected) information** matrix.
- We write $\imath_1(\theta)$ for the Fisher information for a single observation of a random sample $Y_1, \ldots, Y_n$, and then that in the sample is $\imath(\theta) = n\imath_1(\theta)$.
- Later we shall see that in large samples, the maximum likelihood estimator $\widehat{\theta}$ satisfies

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

Note: Bartlett identities

- For any θ we have $1 = \int f(y;\theta)\,\mathrm{d}y$, so 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 $\mathrm{E}\left\{-\nabla^2\ell(\theta)\right\} = \imath(\theta)$, and this in turn equals $n\imath_1(\theta)$.

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

$$\mathrm{cov}\{\tilde{\theta}, U(\theta)\}^2 = \{1 + b'(\theta)\}^2 \leq \mathrm{var}(\tilde{\theta})\mathrm{var}\{U(\theta)\} = \mathrm{var}(\tilde{\theta})\imath(\theta),$$

which gives the result.

☐ If the bias is of order n^{-1}, so too is its derivative, so in large samples we obtain

$$\mathrm{var}(\tilde{\theta}) \geq \imath(\theta)^{-1} = \mathrm{var}(\widehat{\theta}).$$

Pivots

☐ Point estimation does not express uncertainty — we need to assess how well the observed data y^{o} support different possible values of a parameter.

☐ We aim to find subsets of the parameter space that contain the 'true' parameter with a specified probability — when the parameter of interest is scalar, these subsets are usually intervals.

☐ Pivots are useful in finding such subsets.

Definition 18 *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., one that does not depend on θ).*

Example 19 *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 19

- 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 20 *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 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 $1 - \alpha = 0.9, 0.95, 0.99, \ldots$
- A weaker requirement is $\mathrm{P}(L \leq \theta \leq U) \geq 1 - \alpha$, giving confidence level *at least* $1 - \alpha$.

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** — it is convenient if Q is monotone in θ for each Y.
- Often we have an approximate pivot $(\widehat{\theta} - \theta)/V^{1/2} \dot{\sim} \mathcal{N}(0,1)$, 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 21 *In Example 19, find CIs based on Q_1 and on Q_2.*

Note to Example 21

- 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

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

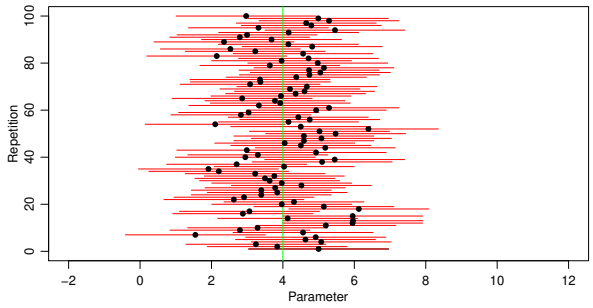
  and hence we obtain

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

  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;
  - these errors are properties of the CI procedure, not of individual intervals!

Prediction

- Prediction refers to 'estimation' of unobserved (future, latent, ...) random variables Y_+.
- In parametric cases we often base **prediction (or tolerance) intervals** on existing data Y by finding a pivot that depends on both Y_+ and Y, and predicting Y_+ using this pivot, e.g., using its mean or median.

Example 22 *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.*

Example 23 (Conformal prediction) *Suppose we seek a prediction interval for the outcome of an ML algorithm. In the simplest case, with $Y_1, \ldots, Y_n, Y_+$ real-valued and exchangeable, $\beta \in (0,1)$, $m = \lceil (n+1)\beta \rceil$ and q_β equal to the mth order statistic of $Y_1, \ldots, Y_n$, show that*

$$\mathrm{P}(Y_+ \leq q_\beta) \geq \beta,$$

and deduce that $\mathrm{P}(q_\alpha < Y_+ \leq q_{1-\alpha}) \geq 1 - 2\alpha$.

Note to Example 22

- 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 $\overline{y}^{\mathrm{o}}$.

Note to Example 23

- Let q_β^+ denote the mth order statistic of $\mathcal{Y}_+ = \{Y_1, \ldots, Y_n, Y_+\}$, and note that under exchangeability Y_+ equals any of the order statistics of $\mathcal{Y}_+$ with probability $1/(n+1)$. Therefore

$$\mathrm{P}(Y_+ \leq q_\beta^+) = m/(n+1) = \lceil (n+1)\beta \rceil / (n+1) \geq (n+1)\beta/(n+1) = \beta.$$

- Now suppose that $m = 2$ and $Y_+ \leq q_\beta^+$, so using an obvious notation $\mathcal{Y}$ can be represented as

$$\bullet \leq + \leq \bullet \leq \cdots \quad \text{or} \quad + \leq \bullet \leq \bullet \leq \cdots .$$

  In both cases $q_\beta \geq q_\beta^+$, so $Y_+ \leq q_\beta^+$ implies that $Y_+ \leq q_\beta$, and conversely. This holds for any m, so

$$\mathrm{P}(Y_+ \leq q_\beta) = \mathrm{P}(Y_+ \leq q_\beta^+) \geq \beta.$$

  Finally

$$\mathrm{P}(q_\alpha < Y_+ \leq q_{1-\alpha}) = \mathrm{P}(Y_+ \leq q_{1-\alpha}) - \mathrm{P}(Y^+ \leq q_\alpha) \geq 1 - \alpha - \alpha = 1 - 2\alpha,$$

  as required.
- For this argument to be practical we must have $1 \leq m \leq n$, so if β is too small or too large, then we must replace the corresponding limit by $\pm\infty$, which does not usually give a useful interval.
- In applications the data are of form (X, Y) and we train a prediction algorithm $\widehat{f}$ using a training subset of $\mathcal{Y} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, giving residuals $Y_j - \widehat{f}(X_j)$ for a test subset of $\mathcal{Y}$ disjoint from the training set, and then apply the argument above to these residuals and $Y_+ - \widehat{f}(X_+)$.

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 data, but they have implications for data.
- Sometimes 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 use our data to disprove it.

Elements of a test

- A **null hypothesis** H_0 to be tested.
- A **test statistic** T, large values of which 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 to check whether observed counts O_k in K categories agree with their expectations $E_k = \mathrm{E}(O_k)$ computed under H_0.
- In any case we need to know (or be able to approximate) the distribution of T under H_0.

1.4 Bases for Uncertainty

Uncertainty

- Essentially three bases for statements of uncertainty:
  - a **frequentist (sampling theory) inference** compares y with a set $\mathcal{S} \subset \mathcal{Y}$ of other data that might have been observed in a hypothetical sampling experiment;
  - a **Bayesian (inverse probability) inference** expresses uncertainty 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** (aka **recognisable subset**) $\mathcal{S}$ of the sample space $\mathcal{Y}$ thoughtfully.

Example 24 (Measuring machines) *A physical quantity θ can be measured with two machines, both giving normal observations $Y \sim \mathcal{N}(\theta, \sigma_m^2)$. A measurement from machine 1 has variance $\sigma_1^2 = 1$, and one from machine 2 has variance $\sigma_2^2 = 100$. A machine is chosen by tossing a fair coin, giving $M = 1, 2$ with equal probabilities. Thus $\mathcal{Y} = \{(y, m) : y \in \mathbb{R}, m \in \{1, 2\}\}$.*
If we observe $(y, m) = (0, 1)$, then clearly we can ignore the fact that we might have observed $m = 2$, i.e., we should take $\mathcal{S}_1 = \{(y, 1) : y \in \mathbb{R}\}$ rather than $\mathcal{S}_2 = \{(y, 2) : y \in \mathbb{R}\}$ or $\mathcal{S} = \mathcal{Y}$.

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 performed with respect to $\mathcal{S}$.
- ☐ We choose $\mathcal{S}$ 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 intervals are used infinitely often, then
  - although any particular interval either does or does not contain θ,
  - it was drawn from a population of intervals with error probability exactly α.
- ☐ Bayesians object that inferences should only be based on the dataset y^{o} actually observed, so the reference set $\mathcal{S}$ is irrelevant.

Example 25 *What would the confidence intervals look like in Example 24? How would the image on slide 42 change? What hypothetical repetitions form the reference sets?*

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 base all our uncertainty statements on the **posterior density** given by Bayes' theorem,

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

- ☐ 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 $\tilde{\theta}$.

Example 26 *Perform Bayesian inference based on* $Y_1, \ldots, Y_n \mid \theta \overset{\text{iid}}{\sim} U(0, \theta)$ *with a Pareto*(a, b) *prior for* θ.

Note to Example 26

- In situations like this, where the support of the density depends on a parameter, it is useful to include an indicator function when writing down the density, viz

$$f(y \mid \theta) = \theta^{-1} I(0 < y < \theta), \quad y \in \mathbb{R}, \theta > 0.$$

  As a function of y for fixed θ, its support is the set $(0, \theta)$, but as a function of θ for fixed y, its support is (y, ∞). Sketch these to appreciate the difference.

- The prior density is $f(\theta) = ab^a/\theta^{a+1} I(\theta > b)$ for $a, b > 0$, and the joint density of the data is

$$f(y \mid \theta) = f(y_1, \ldots, y_n \mid \theta) = \prod_{j=1}^{n} f(y_j \mid \theta) = \prod_{j=1}^{n} I(0 < y_j < \theta)\theta^{-1} = \theta^{-n} I(0 < m < \theta).$$

  where $m = \max(y_1, \ldots, y_n)$, so the posterior density is proportional to

$$f(\theta \mid y) \propto f(y_1, \ldots, y_n \mid \theta) f(\theta) = \theta^{-n} I(0 < m < \theta) \frac{ab^a}{\theta^{a+1}} I(\theta > b) \propto \theta^{-(A+1)} I(\theta > B),$$

  where $A = a + n$ and $B = \max(m, b)$. There are two possibilities here: the prior gives a lower bound b for θ, and if $m < b$ then there is no reason to update this lower bound, but if $m > b$ then clearly $\theta > m > b$, so the bound must be increased at least to m.

- The posterior density has support on (B, ∞) and is proportional to $\theta^{-(A+1)}$, so it is Pareto($A = a + n, B = \max(y_1, \ldots, y_n, b)$). The p quantile of this distribution satisfies $p = 1 - (B/\theta_p)^A$, i.e., $\theta_p = B(1-p)^{-1/A}$, which depends on the data and prior; of course $0 < p < 1$.

- To get a point estimate we might take loss function

$$L(\tilde{\theta}, \theta) = |\tilde{\theta} - \theta| = (\tilde{\theta} - \theta) I(\tilde{\theta} > \theta) + (\theta - \tilde{\theta}) I(\theta > \tilde{\theta}),$$

  and a standard computation shows that this is minimised at $\tilde{\theta} = \theta_{1/2} = B 2^{1/A}$.

Comments on Bayesian inference

- ☐ Often Bayesian models are formulated using a judgement that some variables/observations are exchangeable, as de Finetti theorems then imply that we can write

$$Y_1, \ldots, Y_n \mid \theta \overset{\text{iid}}{\sim} f(y; \theta), \qquad \theta \sim f(\theta).$$

- ☐ In general, 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.
- ☐ Objectively specifying prior 'ignorance' is problematic and can lead to paradoxes, especially in high dimensions.
- ☐ (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	$\theta + l_j$	$\theta + r_j$

- We observe either $(\theta + l_j, r_j)$ or $(l_j, r_j + \theta)$ so the difference D_j of B and A for boy j is $\theta + l_j - r_j$ or $\theta + r_j - l_j$. These are equally likely, so we can write $D_j = \theta + I_j c_j$, where
  - θ 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 $(\theta + l_j, r_j)$ for boy j, then we cannot observe $(l_j, \theta + r_j)$, 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

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

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

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

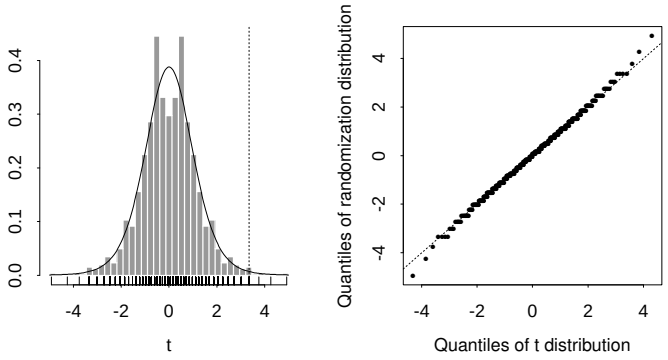
- Hence $\mathrm{E}(\overline{D}) = \theta$ and $\mathrm{var}(\overline{D}) = m^{-2}\sum_{j=1}^{m} c_j^2$, which is unknown because the c_j are unknown, is estimated by (exercise)

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

- $\overline{D}$ and S^2 can be computed from the observed data, so the standardized quantity $Z = (\overline{D} - \theta)/S$ is an approximate pivot.
- If there was no difference between B and A (i.e., $\theta = 0$), 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 $\theta = 0$, 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 θ.
- If $\theta = 0$, then the observed value of $\overline{D}$ is highly unlikely: just 3 values of $\overline{D}$ exceed $\overline{d} = 0.41$, so if $\theta = 0$ then **exact calculation** gives

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

  which seems unlikely enough to suggest that $\theta > 0$.
- Normal distribution theory suggests that $Z \mathrel{\dot\sim} t_9$, 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.

Big picture summary

- 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 59

2.1 Likelihood slide 60

Likelihood

- We now suppose that the data are provisionally believed to come from a parametric model $f_Y(y;\theta)$ for which θ lies in $\Theta \subset \mathbb{R}^d$.
- 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 formula for $\ell(\theta)$ 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 27 (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 27

☐ 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 prefer inferences to be invariant 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 $\varphi(\theta)$ will give

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

  where the tilde denotes the density expressed using φ. Clearly

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

  so the maximum likelihood estimates satisfy $\widehat{\varphi} = \varphi(\widehat{\theta})$. This implies that we can optimise ℓ in a numerically convenient parametrisation, φ, say, and then transform to θ.

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

  but in applications this cannot be taken for granted.
- ☐ Ideally we can ensure random sampling and full measurement of observations from a well-specified population, but if not, possible complications include:
  - selection of observations based on their values;
  - 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

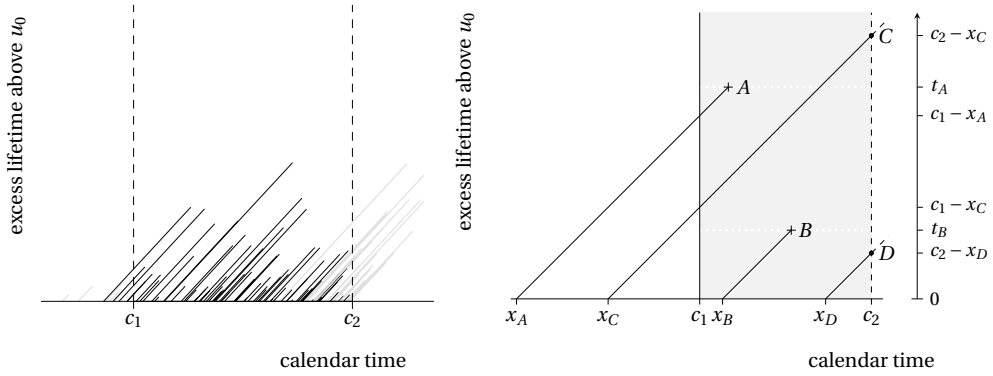
$$\prod_{j=1}^{n} f(y_j \mid y_j \in \mathcal{I}_j; \theta).$$

Example 28 *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*

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

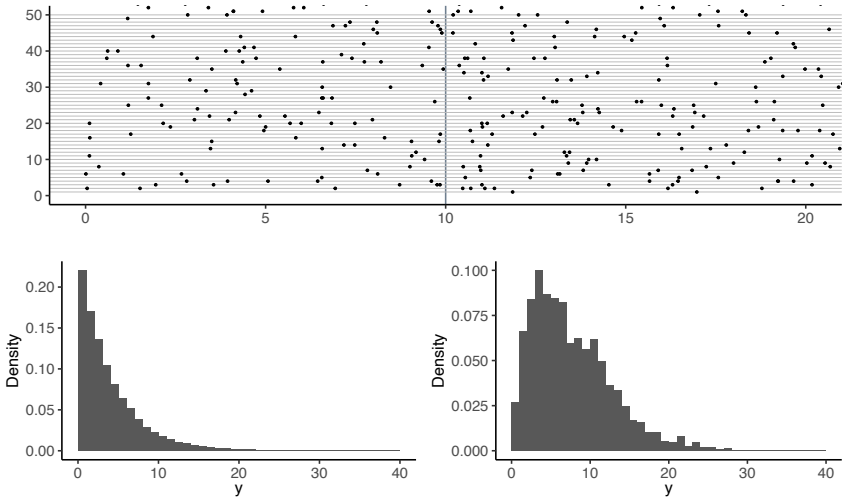
where x is the calendar date at which they reach age u_0. See the next page.

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.

Length-biased sampling



Top: we select the intervals that contain time $y = 10$.
Lower left: histogram of all the intervals
Lower right: histogram of the selected intervals.

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 $\mathrm{E}(Y) = \mu$ 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, in sampling networks, and in other contexts.

Censoring

- Selection and truncation determine which observations appear in a sample, whereas censoring reduces the information available.
- **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 certain intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$ contains Y.
- Here the interval limits may be random, for simplicity are often taken to be independent of 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 29 (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 29

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

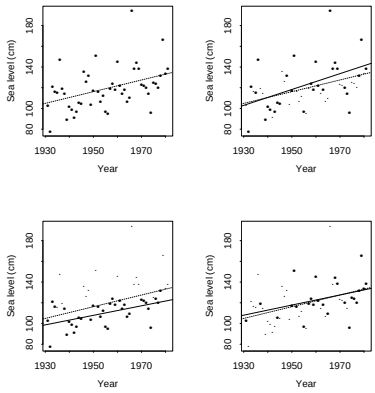
$$\mathrm{P}(I = 1 \mid x, y) f(y \mid x; \theta) f(x; \theta), \qquad \int \mathrm{P}(I = 0 \mid x, y) f(y \mid x; \theta) f(x; \theta) \, \mathrm{d}y,$$

  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 them, especially in terms of non- and semi-parametric estimation.
- Selection (especially self-selection!) can totally undermine analysis 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.

Sufficiency

- When can a lot of data be reduced to a few relevant quantities without 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 $f_{Y|S}(y \mid s)$,

  and 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 30 (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 30

- 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 to compute 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.

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. Hence S is not unique.
- ☐ To deal with this we define a **minimal sufficient statistic** to be a function of any other sufficient statistic. This gives a 'maximal data reduction' and is unique up to 1–1 maps.
- ☐ To formalise this, 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 31 (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 32 *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 31

- ☐ 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 of notation 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 32

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

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

  because $y'_{s(y)}$ is a function of $s(y)$. This factorisation shows that $S = s(Y)$ is sufficient.
- ☐ To show minimality, if $T = t(Y)$ is any other sufficient statistic the factorisation theorem gives

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

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

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

  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 33 (Uniform model) *Discuss minimal sufficiency when* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0,\theta)$.

Example 34 (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 33

☐ We already saw in Example 30 that $M = \max(Y_1, \ldots, Y_n)$ is sufficient, so if $U = \min(Y_1, \ldots, Y_n)$ then clearly $S = (U, M)$ 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 a 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 this 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.

☐ 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.

Note to Example 34

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

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

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

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

  where $\mathcal{Y}_s$ is the set of permutations of $(y_1, \ldots, y_n)$ with order statistics s; clearly $|\mathcal{Y}_s| = n!$, because there are no ties.
- 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 35 (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 36 (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 35

☐ 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 36

☐ The log joint density is

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

so $s = \sum s(y_j)$ is sufficient. It is also minimal, because

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

does not depend on θ iff $\sum s(y_j) = \sum s(z_j)$ (and $n = m$).

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

Theorem 38 (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 37

- 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 39 *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 40 *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 41 (2×2 table) *Apply Lemma 40 to the 2×2 table*

	Success	*Failure*	*Total*
Treated	R_1	m_1-R_1	m_1
Control	R_0	m_0-R_0	m_0
Total	R_1+R_0	$m_0+m_1-R_1-R_0$	m_1+m_0

where $R_0\sim B(m_0,\pi_o)$ and $R_1\sim B(m_1,\pi_1)$ are taken to be independent.

Note to Lemma 39

If ψ is regarded as fixed, then we can write

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

where the rightmost term is free of λ, with logarithm

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

Note to Lemma 40

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

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

so

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

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

Note to Example 41

- 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, $R_1 \sim B(m_1, \pi_1)$. Thus m_0 and m_1 are considered to be fixed, and R_0 and R_1 as random.
- A number of parameters might be of interest, but most commonly ψ is taken to be the difference in log odds of success and λ the log odds of success in the control group, i.e.,

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

giving

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

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

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

- Lemma 40 implies that conditioning on $W = R_0 + R_1$ will eliminate λ. Now

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

where $r_- = \max(0, w-m_0)$, $r_+ = \min(w, m_1)$, so 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.
- An ancillary statistic is distribution-constant, but the converse may not be true.

Example 42 (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 43 (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 44 (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 45 (Basu) *A complete minimal sufficient statistic is independent of any distribution-constant statistic.*

Note to Example 44

☐ 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 45

☐ 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 $F(T;\psi_0)$ uniform?

- 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.
- This argument holds 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**

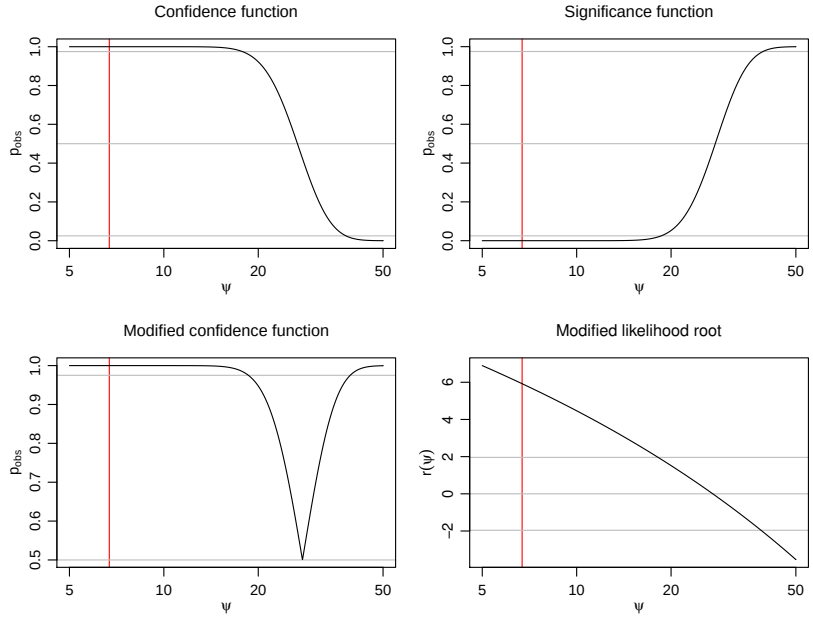
$$p(\psi) = \mathrm{P}(T \ge t^o;\psi) = 1 - F(t^o;\psi) \quad \text{against} \quad \psi.$$

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

Example 47 (Uniform sample) *Apply the recipe above to inference for the upper limit of a uniform sample.*

Example 48 (2×2 table) *Apply the recipe above to the 2×2 table.*

Note to Example 46

- ☐ 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 47

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 48

☐ 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 87 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 92

3.1 Basic Results

slide 93

Motivation

- ☐ Likelihood
  - provides a general paradigm for inference on parametric models, with many generalisations and variants;
  - uses only minimal sufficient statistics;
  - 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 $f \equiv g$ (Jensen).
- ☐ In a parametric setting f belongs to a parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, 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 large-sample fit of f_θ to g.
- ☐ In an ideal case $g \in \mathcal{F}$, so $g = f_{\theta_g}$, but the theory does not require this (yet).
- ☐ The natural estimator of θ_g is the **maximum likelihood estimator**

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

  but we need conditions on $\overline{\ell}$ to ensure that $\widehat{\theta} \xrightarrow{P} \theta_g$ or (better) $\widehat{\theta} \xrightarrow{\text{a.s.}} \theta_g$ as $n \to \infty$.

Regular models

- Notation: $\nabla h(\theta) = \partial h(\theta)/\partial\theta$ and $\nabla^2 h(\theta) = \nabla\nabla^{\mathrm{T}} h(\theta) = \partial^2 h(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$.
- The asymptotic properties of the MLE rely on **regularity conditions**:

(C1) θ_g is unique and 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 $\hbar_1(\theta_g) = \imath_1(\theta_g)$.

Regularity conditions

- (C1) ensures that $\widehat{\theta}$ can be 'on all sides' of θ_g in the limit — if it fails, then any limiting distribution cannot be normal;
- (C2) is essential for consistency, otherwise $\widehat{\theta}$ might not converge — it often fails in mixture models, for which care is needed;
- (C3) is needed to bound terms of a Taylor series — can be replaced by other conditions, see 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 49 *If $Y_1, \ldots, Y_n \sim g$ and $n \to \infty$, then under (C1) and (C2) 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 49

- ☐ 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 50 *If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, then under (C1)–(C4) the consistent sequence of maximum likelihood estimators $\widehat{\theta}$ satisfies*

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

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 we can use the approximation

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

  where $\imath(\theta) = n\imath_1(\theta)$ and $\hbar(\theta) = n\hbar_1(\theta)$ correspond to a random 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 50: A (fairly) simple argument

☐ Write

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

and note that $U_n = n^{1/2}\nabla\overline{\ell}(\theta_g) \xrightarrow{D} U \sim \mathcal{N}_d\{0, \hbar_1(\theta_g)\}$, so writing $Z_n = n^{1/2}(\widehat{\theta} - \theta_g)$ we have

$$\imath_1(\theta_g)^{-1}U_n = \imath_1(\theta_g)^{-1}\left\{-\int_0^1 \nabla^2\overline{\ell}(\theta_g + tn^{-1/2}Z_n)\,\mathrm{d}t\right\}Z_n = \imath_1(\theta_g)^{-1}J_n^*Z_n,$$

say, and as $n \to \infty$, $J_n^* \doteq -\int_0^1 \nabla^2\overline{\ell}(\theta_g)\,\mathrm{d}t \xrightarrow{P} \imath_1(\theta_g)$ and thus $\imath_1(\theta_g)^{-1}J_n^* \xrightarrow{P} I_d$. Hence

$$\imath_1(\theta_g)^{-1}J_n^*Z_n = \imath_1(\theta_g)^{-1}U_n \xrightarrow{D} \imath_1(\theta_g)^{-1}U \sim \mathcal{N}_d\{0, \imath_1(\theta_g)^{-1}\hbar_1(\theta_g)\imath_1(\theta_g)^{-1}\}.$$

☐ For a more careful treatment of the integral, we need a *uniform law of large numbers (ULLN)*, which requires that $J_n(\theta) = -\nabla^2\overline{\ell}(\theta)$ is measurable and continuous in θ within a compact subset $\mathcal{N}'$ of $\mathcal{N}$, for almost all y, and that there exists a function $d(Y)$ whose expectation is finite and for which $\|J_n(\theta)\| < d(Y)$ for all $\theta \in \mathcal{N}'$, where $\|\cdot\|$ is a matrix norm. Then $\mathrm{E}\{J_n(\theta)\} = \imath_1(\theta)$ is continuous in θ and

$$\sup_{\theta\in\mathcal{N}'} \|J_n(\theta) - \imath_1(\theta)\| \xrightarrow{P} 0, \quad n \to \infty.$$

☐ Let $\delta > 0$ be small enough that $B_\delta = \{\theta : |\theta - \theta_g| \le \delta\} \subset \mathcal{N}'$ and let $A_n = \{|n^{-1/2}Z_n| \le \delta\}$ and $C_n = \|J_n^* - \imath_1(\theta_g)\|$. Then for $\varepsilon > 0$ we have

$$\mathrm{P}(C_n > \varepsilon) = \mathrm{P}(\{C_n > \varepsilon\} \cap A_n) + \mathrm{P}(\{C_n > \varepsilon\} \cap A_n^c) \le \mathrm{P}(\{C_n > \varepsilon\} \cap A_n) + \mathrm{P}(A_n^c),$$

where the last term tends to zero because $n^{-1/2}Z_n = \widehat{\theta} - \theta_g \xrightarrow{P} 0$. Now if A_n holds, then $\theta_g + tn^{-1/2}Z_n \in B_\delta$ when $0 \le t \le 1$, so

$$\begin{aligned}
C_n &= \left\|\int_0^1 \left\{J_n(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g + tn^{-1/2}Z_n) + \imath_1(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g)\right\}\,\mathrm{d}t\right\| \\
&\le \int_0^1 \left\|J_n(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g + tn^{-1/2}Z_n)\right\|\,\mathrm{d}t + \int_0^1 \left\|\imath_1(\theta_g + tn^{-1/2}Z_n) - \imath_1(\theta_g)\right\|\,\mathrm{d}t \\
&\le \sup_{\theta\in B_\delta} \|J_n(\theta) - \imath_1(\theta)\| + \sup_{\theta\in B_\delta} \|\imath_1(\theta) - \imath_1(\theta_g)\| \\
&= D_n + E_n,
\end{aligned}$$

say. If $C_n > \varepsilon$ then at least one of D_n and E_n must exceed $\varepsilon/2$, so

$$\begin{aligned}
\mathrm{P}(\{C_n > \varepsilon\} \cap A_n) &\le \mathrm{P}\left(\{\{D_n > \varepsilon/2\} \cup \{E_n \ge \varepsilon/2\}\} \cap A_n\right) \\
&\le \mathrm{P}(\{D_n > \varepsilon/2\} \cap A_n) + \mathrm{P}(\{E_n \ge \varepsilon/2\} \cap A_n) \\
&\le \mathrm{P}(D_n > \varepsilon/2) + \mathrm{P}(E_n > \varepsilon/2).
\end{aligned}$$

Now $D_n \xrightarrow{P} 0$ using the ULLN, and the continuity of $\imath_1(\theta)$ at θ_g implies that E_n can be made smaller than $\varepsilon/2$ by a suitable choice of $\delta > 0$, in which case

$$\begin{aligned}
\mathrm{P}(C_n > \varepsilon) &\le \mathrm{P}(\{C_n > \varepsilon\} \cap A_n) + \mathrm{P}(A_n^c) \\
&\le \mathrm{P}\left(D_n > \varepsilon/2\right) + \mathrm{P}\left(E_n > \varepsilon/2\right) + \mathrm{P}(A_n^c) \\
&\to 0, \quad n \to \infty,
\end{aligned}$$

which implies that $J_n^* \xrightarrow{P} \imath_1(\theta_g)$ and therefore that $\imath_1(\theta_g)^{-1}J_n^* \xrightarrow{P} I_d$, as required.

Note to Theorem 50: Another approach

☐ 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 $\theta^* \xrightarrow{P} \theta_g$). Hence

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

☐ 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 (3) 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 51 shows that $R_n \xrightarrow{P} 0$, so the denominator of (3) 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: A useful lemma

Lemma 51 *Under the conditions of Theorem 50,* $R_n = (\widehat{\theta} - \theta_g)\nabla^3\overline{\ell}(\theta^*) \xrightarrow{P} 0$ *as* $n \to \infty$.

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

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

  where the first term tends to zero because the sequence $\widehat{\theta}$ is consistent.
- If $|\widehat{\theta} - \theta_g| < \delta$, then (C3) implies that

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

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

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

  and for $\eta > 0$ this equals

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

  which is bounded by

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

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

Classical asymptotics

- The true model is supposed to lie in the candidate family, i.e., $g \in \mathcal{F}$, so $\theta_g \in \Theta$.
- We saw on slide 38 that the moments of the $d \times 1$ **score vector** $U(\theta) = \nabla\ell(\theta)$ are given under mild conditions by the Bartlett identities, i.e.,

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

- Hence $\imath(\theta) = \hbar(\theta)$, and $\imath(\theta) = n\imath_1(\theta) = n\hbar_1(\theta)$ when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{\theta_g}$.
- 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 largely irrelevant if model-checking suggests that $f_{\widehat{\theta}_g}$ 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.

Note: Stable interpretation of a parameter

☐ To put some mathematical flesh on the discussion, suppose that $g(y) = f(y; \theta, \gamma)$ and the assumed model is $f(y; \theta, 0)$. Then for small γ, $\theta_g \equiv \theta_\gamma$ satisfies

$$\begin{aligned} 0 &= \int \nabla_\theta \log f(y; \theta_\gamma, 0) f(y; \theta, \gamma) \, \mathrm{d}y \\ &= \int \left\{ \nabla_\theta \log f(y; \theta, \gamma) + \nabla_\theta^2 \log f(y; \theta, \gamma)(\theta_\gamma - \theta) + \nabla_\gamma^{\mathrm{T}} \nabla_\theta \log f(y; \theta, \gamma)(0 - \gamma) + \cdots \right\} f(y; \theta, \gamma) \, \mathrm{d}y \\ &= 0 - \imath_{\theta\theta}(\theta, \gamma)(\theta_\gamma - \theta) + \imath_{\theta\gamma}(\theta, \gamma)\gamma + o(\gamma), \end{aligned}$$

which implies that the effect of incorrectly assuming that $\gamma = 0$ is that $\widehat{\theta}$ converges to

$$\theta_\gamma = \theta + \imath_{\theta\theta}^{-1}(\theta, \gamma)\imath_{\theta\gamma}(\theta, \gamma)\gamma + o(\gamma).$$

☐ It is also easy to check that $\hbar_{\theta\theta}(\theta, 0) = \imath_{\theta\theta}(\theta, 0) + O(\gamma)$, so the two matrices become equal if $\gamma \to 0$, in which case $\imath_1(\theta, \gamma)^{-1} \hbar_1(\theta, \gamma) \imath_1(\theta, \gamma)^{-1} \to \imath_1(\theta, \gamma)^{-1}$, which implies that for small γ we have

$$n^{1/2}(\widehat{\theta} - \theta) = n^{1/2}(\widehat{\theta} - \theta_\gamma) + n^{1/2}(\theta_\gamma - \theta) \dot{\sim} \mathcal{N}_d\{0, \imath_1(\theta, \gamma)^{-1}\} + n^{1/2}(\theta_\gamma - \theta).$$

☐ Now if $\gamma = n^{-a}\delta$ for some $a > 0$, then

$$n^{1/2}(\theta_\gamma - \theta) = n^{1/2-a} \imath_{\theta\theta}^{-1}(\theta, \gamma)\imath_{\theta\gamma}(\theta, \gamma)\delta,$$

which will tend to infinity if $a < 1/2$ (should be obvious asymptotically), to zero if $a > 1/2$ (can be ignored asymptotically) and to a constant if $a = 1/2$. Hence there is an asymptotic bias for $\widehat{\theta}$ if there is misspecification, $\delta \neq 0$, unless $\imath_{\theta\gamma}(\theta, \gamma) = 0$, i.e., the information matrix covariance for the scores for θ and γ is zero. This is known as orthogonality of θ and γ; see later.

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

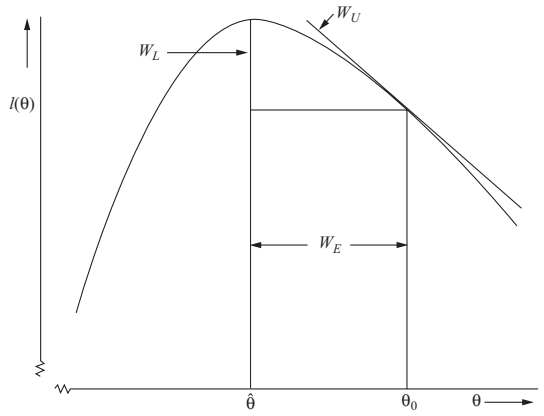


Figure 6.2. Three asymptotically equivalent ways, all based on the log likelihood function of testing null hypothesis $\theta = \theta_0$: W_E, horizontal distance; W_L vertical distance; W_U slope at null point.

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

Related statistics

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

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

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

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

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

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

leads to $(1-\alpha)$ **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 $\imath(\theta_g)$;

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

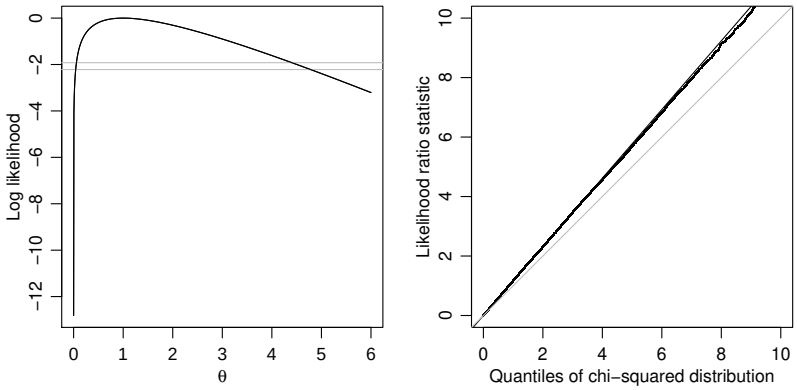
Note to Example 52

- ☐ 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

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

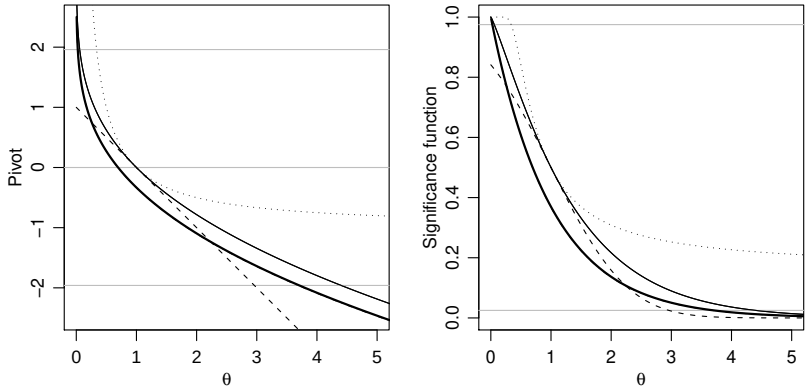
- ☐ The exact pivot is $\theta\sum Y_j$ whose distribution is gamma with unit scale and shape parameter n.
- ☐ Consider an exponential sample with $n = 1$ and $\overline{y} = 1$; then $\widehat{\jmath} = 1$. The log likelihood $\ell(\theta)$, shown in the left-hand panel of the figure, is unimodal but 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 Bartlett 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 53 *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 53

☐ In this case $1 = \int f(y;\theta)\,\mathrm{d}y = \int_0^\theta \theta^{-1}\,\mathrm{d}y$, and differentiation with respect to θ gives

$$0 = 1/\theta + \int_0^\theta (-\theta^{-2})\,\mathrm{d}y,$$

so the first Bartlett identity is not satisfied (because the support depends on θ, and $f(\theta;\theta) \neq 0$.

☐ Owing to the independence,

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

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

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

Now

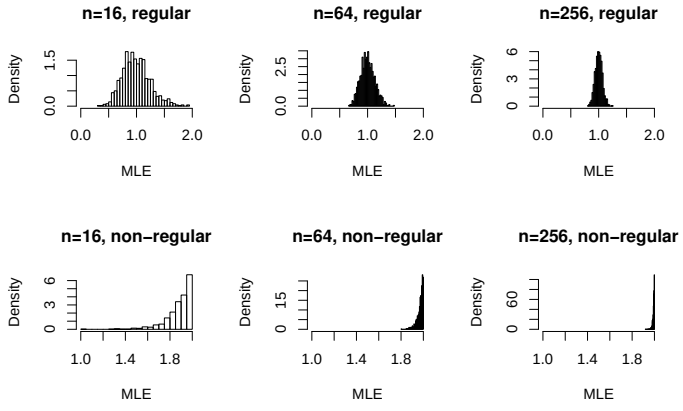
$$\mathrm{P}\left\{n(\theta - \widehat{\theta})/\theta \leq x\right\} = \mathrm{P}(\widehat{\theta} \geq \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n \to 1 - \exp(-x),$$

as required. Note that:

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

Uniform example

Comparison of the distributions of $\widehat{\theta}$ in a regular case (panels above, with standard deviation $\propto n^{-1/2}$) and in a nonregular case (Example 53, 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.



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 $\widehat{\lambda}_\psi = \max_\lambda \ell(\psi, \lambda)$, $\tilde{\theta} = \widehat{\theta}_\psi = (\psi, \widehat{\lambda}_\psi)$,

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

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

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

  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** is defined as

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

- 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

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

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

☐ To lighten the notation we let ℓ, $\tilde{\ell}$ and $\widehat{\ell}$ denote $\ell(\psi_g, \lambda_g)$, $\ell(\psi_g, \widehat{\lambda}_{\psi_g})$ and $\ell(\widehat{\psi}, \widehat{\lambda})$, and likewise with derivatives such as $\ell_\theta = \partial\ell(\theta)/\partial\theta|_{\theta=\theta_g}$, $\ell_{\lambda\psi} = \partial^2\ell(\theta)/\partial\lambda\partial\psi^{\mathrm{T}}|_{\theta=\theta_g}$. We shall also replace matrices such as $\ell_{\theta\theta}$ by large-sample approximations such as $-\imath_{\theta\theta}$; this can be justified by dividing both sides by n and noting that $-\overline{\ell}_{\theta\theta}(\theta_g) \xrightarrow{P} \imath_1(\theta_g)$.

☐ 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 $\cdots$ denotes terms of smaller order containing third derivatives of ℓ. The λ component of this equation is

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

Likewise

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

Equating the expressions for ℓ_λ from the last two displays gives

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

so

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

☐ 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

$$\begin{aligned} 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} &\doteq (\widehat{\theta} - \theta_g)^{\mathrm{T}}\imath_{\theta\theta}(\widehat{\theta} - \theta_g) \\ &= (\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\psi}(\widehat{\psi} - \psi_g) + 2(\widehat{\psi} - \psi_g)^{\mathrm{T}}\imath_{\psi\lambda}(\widehat{\lambda} - \lambda_g) + (\widehat{\lambda} - \lambda_g)^{\mathrm{T}}\imath_{\lambda\lambda}(\widehat{\lambda} - \lambda_g), \end{aligned}$$

and likewise

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

Subtracting the two quadratic forms gives

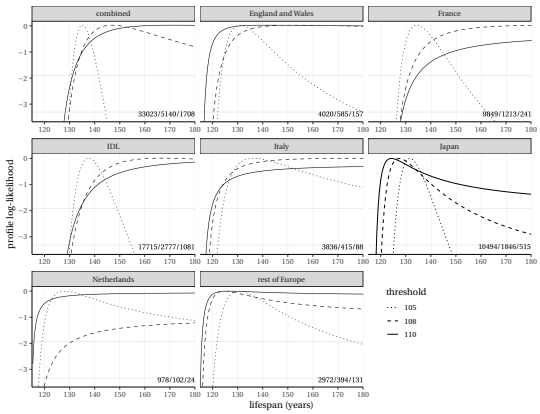
$$\begin{aligned} w_{\mathrm{p}}(\psi_g) &= 2\{\ell(\widehat{\theta}) - \ell(\theta_g)\} - 2\{\ell(\widehat{\theta}_\psi) - \ell(\theta_g)\} \\ &\doteq (\widehat{\psi} - \psi_g)^{\mathrm{T}}(\imath_{\psi\psi} - \imath_{\psi\lambda}\imath_{\lambda\lambda}^{-1}\imath_{\lambda\psi})(\widehat{\psi} - \psi_g),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

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

☐ Here we are under classical asymptotics, whereby the dimensions of ψ and λ are fixed and $n \to \infty$, and arguments along the lines of Theorem 50 show that the terms $\cdots$ all tend in probability to zero, and thus do not affect the limiting distribution.

Example: Human lifespan

Example 54 *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 $-\frac{1}{2}\chi_1^2(0.95) = -1.92$ 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

☐ As

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

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 Taylor expansion

$$\begin{aligned} 2\ell^+(\widehat{\theta}) &= 2\ell^+(\widehat{\theta}^+) + 2(\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\ell_\theta^+(\widehat{\theta}^+) + (\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\ell_{\theta\theta}^+(\widehat{\theta}^+)(\widehat{\theta}-\widehat{\theta}^+) + \cdots \\ &= 2\ell^+(\widehat{\theta}^+) - \mathrm{tr}\left\{(\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)(\widehat{\theta}-\widehat{\theta}^+)\right\} + \cdots \\ &= 2\ell^+(\widehat{\theta}^+) - \mathrm{tr}\left\{(\widehat{\theta}-\widehat{\theta}^+)(\widehat{\theta}-\widehat{\theta}^+)^{\mathrm{T}}\imath_{\theta\theta}(\theta_g)\right\} + \cdots \end{aligned}$$

where $\widehat{\theta}^+$ maximises $\ell^+(\theta)$, $\widehat{\theta}$ maximises $\ell(\theta)$, we have replaced $-\ell_{\theta\theta}^+(\widehat{\theta}^+)$ by its large-sample limit $\imath_{\theta\theta}(\theta_g)$ and neglected terms that are $o_p(1)$. Recall that θ_g is the large-sample limit of $\widehat{\theta}$ when data are sampled from g.

☐ Now $\widehat{\theta}^+$ and $\widehat{\theta}$ are independent and approximately $\mathcal{N}_d(\theta_g, V)$, where $V = \imath_{\theta\theta}^{-1}(\theta_g)\hbar(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)$, so $\widehat{\theta}^+ - \widehat{\theta} \mathrel{\dot\sim} \mathcal{N}_d(0, 2V)$, giving

$$\begin{aligned} -2\mathrm{E}_g\mathrm{E}_g^+\left\{\ell^+(\widehat{\theta})\right\} &\doteq -2\mathrm{E}_g\mathrm{E}_g^+\left\{\ell(\widehat{\theta})\right\} + \mathrm{tr}\left\{2V\imath_{\theta\theta}(\theta_g)\right\} + o(1) \\ &= 2\left[\mathrm{tr}\left\{\hbar(\theta_g)\imath_{\theta\theta}^{-1}(\theta_g)\right\} - \mathrm{E}_g\mathrm{E}_g^+\left\{\ell(\widehat{\theta})\right\}\right] + o(1). \end{aligned}$$

☐ 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* $\mathrm{NIC} = 2\{\mathrm{tr}(\widehat{\hbar}\widehat{\jmath}^{-1}) - \ell(\widehat{\theta})\}$.

☐ Neither AIC or NIC gives consistent selection of the true model, which would require the penalty to grow with n.

☐ The calculations above use generic large-sample likelihood approximations, and can be improved in specific cases (e.g., with normal errors).

3.3 Nuisance Parameters

slide 114

Effect of nuisance parameters

Example 55 (Neyman–Scott) *Find the profile log likelihood for σ^2 when $(y_{j1}, y_{j2}) \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, 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 55

☐ The overall log likelihood is

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

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

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

☐ This is maximised at $\widehat{\sigma}_p^2 = (4n)^{-1}\sum_{j=1}^{n}(y_{j1}-y_{j2})^2$, but as $Y_{j1}-Y_{j2} \overset{\text{iid}}{\sim} \mathcal{N}(0, 2\sigma^2)$, we see that $\sigma_p^2 \xrightarrow{P} \sigma^2/2$ 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 39 and 40) — 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

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

  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

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

  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

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

  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 52).

. . . with nuisance parameters

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

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

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

- In a general exponential family $\varphi(\theta)$ is the canonical parameter, and in a linear exponential family,

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

- In general for independent continuous observations we write

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

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

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 56 (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 56

- 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

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

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 57 (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 57

☐ 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.
- Weaknesses of likelihood approach:
  - requires 'parametric' model for data;
  - can fail in high-dimensional settings;
  - not all models are regular.

4 Hypothesis Testing

slide 126

4.1 Pure Significance Tests

slide 127

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:

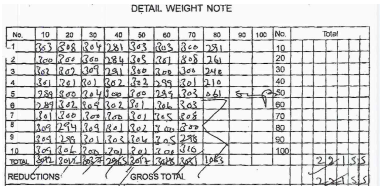
  H_0: the top quark does not exist.

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

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

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

Industrial fraud?


DETAIL WEIGHT NOTE

- $n = 92$ 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 58 *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.
- Massive evidence for non-uniformity (and 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.
- p_{obs} is a realisation of a **P-variable** P, which is $U(0,1)$ under H_0 (if T is continuous), so

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

- 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

- $P \sim U(0,1)$ under H_0, 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 59 *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 60 *Find an exact test on a canonical parameter in a logistic regression model.*

Note to Example 61

- $\overline{Y}$ 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 60

- 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 40 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

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

  where $\mathcal{S}^{\mathrm{o}} = \{(y_1', \ldots, y_n') : X_2^{\mathrm{T}}y' = w^{\mathrm{o}}\}$, with $w^{\mathrm{o}} = X_2^{\mathrm{T}}y^{\mathrm{o}}$ and y^{o} respectively the observed value of W and the observed data.
- 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 a hypothesis 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 can 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

- Fisher regarded a P-value as a **measure of the evidence** against H_0.
- Neyman and Pearson formulated testing as **making a decision** between two hypotheses:
  - 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 losses due to wrong decisions are not taken into account.

Example 61 *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 61

- 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.

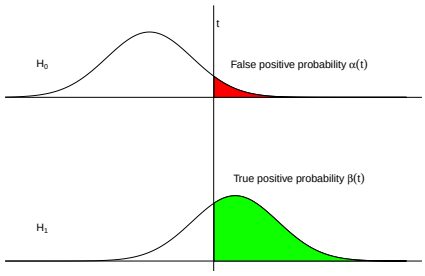
True and false positives: Example

- 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 61 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 62 *The* **receiver operating characteristic (ROC) curve** *of a test plots* $\beta(t)$ *against* $\alpha(t)$ *as* t *varies, i.e., it shows the graph* $(x, y) = (\mathrm{P}_0(T \geq t), \mathrm{P}_1(T > t))$*, 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 then the curve lies on the diagonal, and the hypotheses cannot be distinguished.
- One summary measure of the overall quality of a test is the **area under the curve**,

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

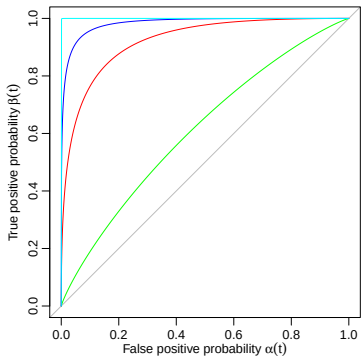
  which ranges between 0.5 for a useless test and 1.0 for a perfect test.

Example

☐ In Example 61 $\alpha(t) = 1 - \Phi\{n^{1/2}(t - \mu_0)/\sigma\}$ 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 $\delta = 2$ (in red). Also shown are curves for $\delta = 0, 0.4, 3, 6$. Which is which?



Neyman–Pearson lemma

Definition 63 *A* **simple hypothesis** *entirely fixes the distribution of the data Y, whereas a* **composite hypothesis** *does not fix the distribution of Y.*

Definition 64 *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 as large as possible provided $\mathrm{P}_0(Y \in \mathcal{Y}_1) \leq \alpha$ (with equality in continuous problems).

Lemma 65 (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 66 *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 65

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. Note that $\mathcal{Y}_0 \cup \mathcal{Y}_1 = \mathcal{Y}_0' \cup \mathcal{Y}_1' = \mathcal{Y}$. If we write $F(\mathcal{C}) = \int_{\mathcal{C}} f(y)\,\mathrm{d}y$ for any density f with corresponding distribution F, then we aim to show that $F_1(\mathcal{Y}_1) \geq F(\mathcal{Y}_1')$. Now

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

equals

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

If $F = F_0$, then (5) is non-negative, because $\alpha = F_0(\mathcal{Y}_1) \geq F_0(\mathcal{Y}_1')$, 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 66

- The likelihood ratio is

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

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

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

  and if $\varphi_1 > \varphi_0$ then

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

  say. 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)

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

  Example 61 shows an example for normal data with $\varphi_1 = \mu_1/\sigma^2 > \varphi_0 = \mu_0/\sigma^2$ and known σ^2.
- If $\varphi_1 < \varphi_0$, then division by $\varphi_1 - \varphi_0 < 0$ leads to (say),

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

- 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 $\varphi_1 < \varphi_0$.

Power

- The NP lemma applies to simple hypotheses, but sometimes (e.g., Example 66) gives **uniformly most powerful (UMP) tests** against composite alternatives, i.e., a single critical region $\mathcal{Y}_1$ is most powerful against $\theta = \theta_1$ for all $\theta_1 > \theta_0$ or for all $\theta_1 < \theta_0$.
- If there is no UMP region, we might compare tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ by
  - comparing them at some (arbitrary) 'typical' alternative;
  - averaging power over some suitable set of alternatives; or
  - looking at local alternatives, i.e., when $\theta_1 = \theta_0 + \delta$ for small δ.
- For local alternatives, note that with scalar θ and mild regularity of the log likelihood,

$$\log\left\{\frac{f(y;\theta_0+\delta)}{f(y;\theta_0)}\right\} = \ell(\theta_0+\delta) - \ell(\theta_0) = \delta\frac{\mathrm{d}\ell(\theta_0)}{\mathrm{d}\theta} + o(\delta) = \delta\ell_\theta(\theta_0) + o(\delta).$$

- Hence the **locally most powerful critical region** for $\delta > 0$ is obtained from large values of the score statistic, and conversely for $\delta < 0$.
- When $\theta = (\psi, \lambda)$ and we test the composite hypothesis $H_0 : \psi = \psi_0$ against $H_0 : \psi > \psi_0$, without constraints on λ, the optimal local test for each λ will be based on the score $\ell_\psi(\theta) = \partial\ell(\psi,\lambda)/\partial\psi$ evaluated at (ψ_0, λ), which unless λ can somehow be eliminated is often replaced in practice by $(\psi_0, \widehat{\lambda}_{\psi_0})$.

Aside: Score testing

- Score tests can be useful when maximising a full likelihood is difficult or not worthwhile.
- Suppose we want to test $H_0 : \theta = \theta_0$ for scalar θ. Under H_0 and classical asymptotics,

$$\ell_\theta(\theta_0) \mathrel{\dot\sim} \mathcal{N}(0, \imath(\theta_0)\} \quad \Longrightarrow \quad \ell_\theta(\theta_0)/\sqrt{\imath(\theta_0)} \mathrel{\dot\sim} \mathcal{N}(0,1),$$

  which gives a basis for the test.
- When $\theta = (\psi, \lambda)$ and $H_0 : \psi = \psi_0$, then

$$\ell_\psi(\widehat{\theta}_0) \mathrel{\dot\sim} \mathcal{N}(0, \imath^{\psi\psi}(\widehat{\theta}_0)^{-1}\} \quad \Longrightarrow \quad \ell_\psi(\widehat{\theta}_0)^{\mathrm{T}}\imath^{\psi\psi}(\widehat{\theta}_0)\ell_\psi(\widehat{\theta}_0) \mathrel{\dot\sim} \chi^2_{\dim\psi},$$

  where $\widehat{\theta}_0 = (\psi_0, \widehat{\lambda}_{\psi_0})$ and

$$\imath^{\psi\psi}(\theta)^{-1} = \imath_{\psi\psi}(\theta) - \imath_{\psi\lambda}(\theta)\imath_{\lambda\lambda}(\theta)^{-1}\imath_{\lambda\psi}(\theta).$$

  If ψ is scalar, then $\ell_\psi(\widehat{\theta}_0)\{\imath^{\psi\psi}(\widehat{\theta}_0)\}^{1/2} \mathrel{\dot\sim} \mathcal{N}(0,1)$.
- In both cases
  - any maximisation is needed only on H_0, and
  - if the expected information is difficult to compute, it can be replaced by the corresponding observed information (if this is positive).

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

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 (e.g., score test for gamma shape);
  - 'dividing hypothesis' used to partition the parameter space into subsets with sharply different interpretations;
  - as a technical device for generating confidence intervals;
  - to flag which of many similar null hypotheses might be false.
- ☐ Hence arguing that testing should be abolished is unreasonable (as well as unrealistic).

Example 67 *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 (for example) when the distribution is fitted to data on supercentenarians (finite or infinite limit to human life?).

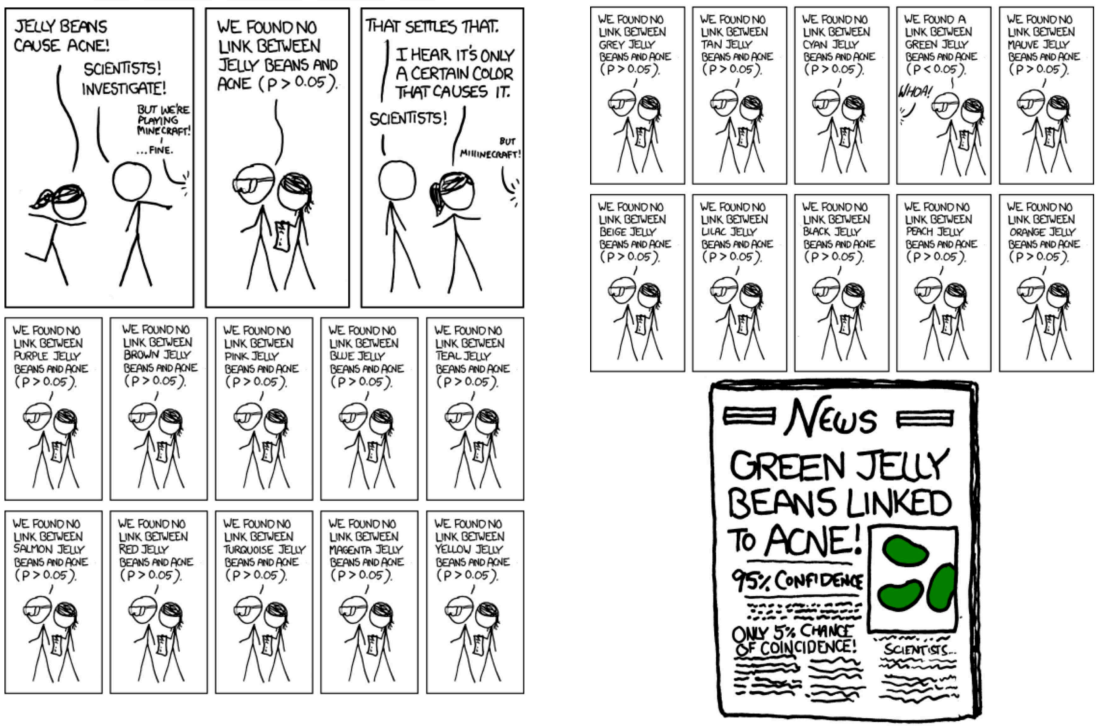
Motivation

- ☐ 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 on the next slide 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

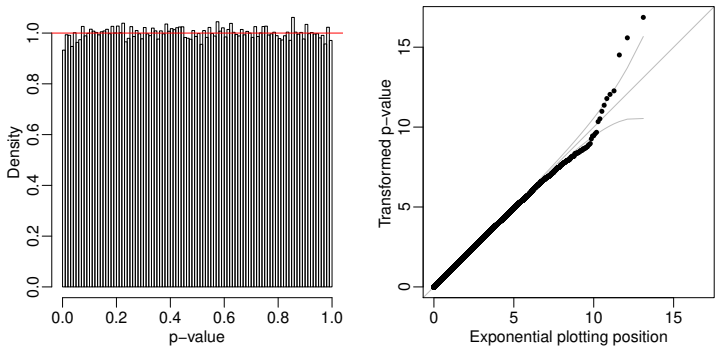
$$p_{\text{obs},j} = \text{P}_0(T_j > t_{\text{obs},j}) = 1 - \text{P}_0(-t_{\text{obs},j} \leq Y_j \leq t_{\text{obs},j}) \doteq 2\Phi(-t_{\text{obs},j}),$$

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 with individual levels all equal to 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 $P_j < \alpha_j$, then **Boole's inequality** (the first **Bonferroni inequality**, aka the **union bound**) gives

$$\mathrm{FWER} = \mathrm{P}(\mathrm{FP} \geq 1) = \mathrm{P}\left(\bigcup_{j=1}^{m_0} \{P_j \leq \alpha_j\}\right) \leq \sum_{j=1}^{m_0} \mathrm{P}\left(P_j \leq \alpha_j\right) = \sum_{j=1}^{m_0} \alpha_j,$$

  so even if the tests are dependent we have strong control of FWER if $\sum_{j=1}^{m} \alpha_j \leq \alpha$.
- Usually we set $\alpha_j \equiv \alpha/m$, so $\sum_{j=1}^{m_0} \alpha_j = m_0\alpha/m \leq \alpha$.
- 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 still gives strong control but is more powerful than the basic Bonferroni procedure, because it uses higher rejection thresholds. Hence the basic procedure should not be used.

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 $\mathrm{FP} \geq 1$, 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 $\mathrm{FP} \geq 1$ 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

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

- 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 aim is to bound the proportion of false positives among the rejections.

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

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

- The **Benjamini–Hochberg procedure** gives strong control for independent tests: 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

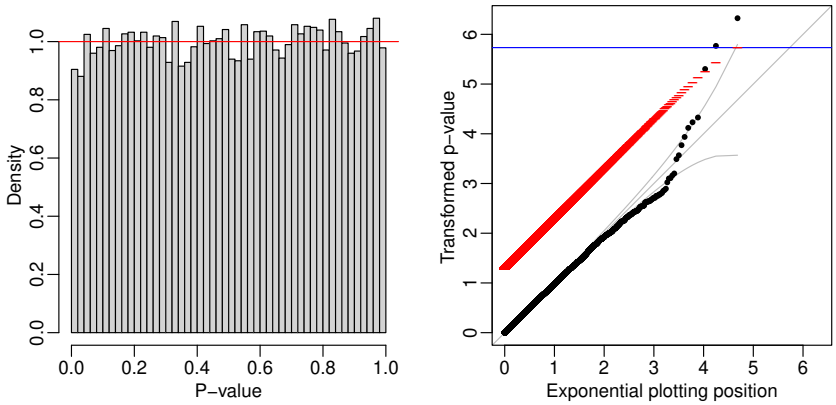
- if $m_0 \ll m$, then the last inequality may be very unequal, so possibly FDR $\ll \alpha$.
- if the P-values are dependent in such a way that

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

then the result also holds.

GWAS, II

- Left: 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, and the $U(0, 1)$ density (red).
- Right: exponential Q-Q plot of $Z_j = -\log_{10} Q_j$, 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

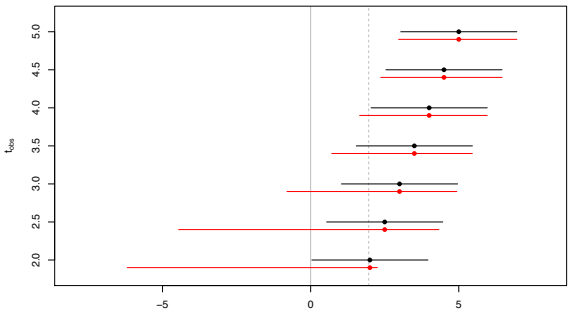
- Contrast
  - **exploratory analysis**, where we study data with no strong prior hypotheses, aiming to find something 'interesting' for future study, and
  - **confirmatory analysis**, where we specify an analysis protocol (hypotheses/tests/...) in advance and stick to it.
- Most statistical procedures 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 $\mathcal{S}$ that is *based on the same data*, and compute

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

  we are effectively pretending that $\mathcal{S}$ did not exist.

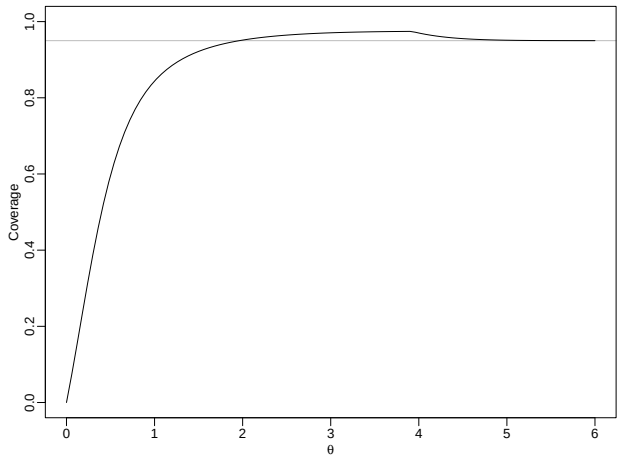
Simple example

Example 68 *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 $\mathcal{I}_{95}$ around the observed t_{obs} if we reject H_0. Compare the resulting confidence intervals when we do and do not allow for selection. What is the coverage of $\mathcal{I}_{95}$ conditional on $\mathcal{S}$?*



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

Simple example II



Conditional coverage $\mathrm{P}(\theta \in \mathcal{I}_{95} \mid \mathcal{S})$ of $\mathcal{I}_{95}$ as a function of θ.

Note to Example 68

- 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 α, $1 - \alpha$ 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 $\mathcal{I}_{95}$ with limits $t_{\mathrm{obs}} \pm 1.96$ when $\alpha = 0.025$.

- If we condition on the selection event $\mathcal{S}_R = \{T > z_{1-\beta}\}$ and, compute the 95% confidence interval for θ if this event occurs, 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 to

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

- If we set $\beta = \alpha = 0.025$, 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.

- The second graph shows that if we ignore the selection and just use the interval $\mathcal{I}_{95}$ after observing the event $\mathcal{S} = \{|T| > z_{0.975}\}$, then the true coverage varies from 0 when $\theta = 0$ to 0.95 when $\theta \to \infty$, but does not pass its nominal value until $\theta > 2$.

Allowing for selection

- Lots of work in last decade, in two main categories:
- methods for specific algorithms (e.g., the lasso) with a selection event $\mathcal{S}$ of a specified form and for which $f(\mathcal{Y} \mid \mathcal{S})$ is tractable;
- more general approaches, including
  - methods that allow for all possible selection procedures, and hence are hyper-conservative (e.g., so-called universal inference, e-values, ...);
  - splitting the data into two or more groups (below);
  - adding noise (less general, since strictly applicable only to certain setttings).
- Garcia Rasines and Young (2023, *Biometrika*) have a good discussion and more references.

Sample splitting

- **Sample splitting** is a standard approach to dealing with selection.
- Partition (independent) original data $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ at random into subsets $\mathcal{Y}_0$ and $\mathcal{Y}_1$, of respective sizes $n_0 = pn$ and $n_1 = (1-p)n$, use $\mathcal{Y}_0$ for selection, and then perform inferences using $\mathcal{Y}_1$.
- As $\mathcal{Y}_1 \perp\!\!\!\perp \mathcal{Y}_0$ and $\mathcal{S}$ depends only on $\mathcal{Y}_0$, we have

$$f(\mathcal{Y}) = f(\mathcal{Y} \mid \mathcal{S}) f(\mathcal{S}) = f(\mathcal{Y}_0, \mathcal{Y}_1 \mid \mathcal{S}) f(\mathcal{S}) = f(\mathcal{Y}_1) f(\mathcal{Y}_0 \mid \mathcal{S}) f(\mathcal{S}),$$

  so any inference based on $\mathcal{Y}_1$ is unaffected by the selection.
- This approach is simple and widely applicable (at least for random samples), but
  - if the split is random, selections and inferences may be different for different splits;
  - there is a loss of power, both for finding any effects (using $\mathcal{Y}_0$) and for inference for them (using $\mathcal{Y}_1$);
  - if the data are not a random sample (e.g., in a regression setup, (y, x), with x treated as constant), then we should aim for similar information contents in $\mathcal{Y}_0$ and $\mathcal{Y}_1$ (more formally, ancillary statistics should be similar for both parts), and it may be hard to achieve this, particularly in high dimensions.

Randomisation

- ☐ Data (Y, X), with X (if present) treated as constant
- ☐ Have random variable W, maybe dependent on X, and base selection on $U = u(Y, W)$, e.g., setting selection variable $S = s(U)$ equal to s.
- ☐ Then base inference on $Y \mid U$, which is conditionally independent of S.
- ☐ If $Y \mapsto (U, V) = (u(Y, W), v(Y, W))$, where (U, V) are jointly sufficient for model and $U \perp\!\!\!\perp V$, then inference from $Y \mid U$ is equivalent to inference from V.
- ☐ Simple example: $Y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$ with σ^2 known, $U = Y + \sigma p W$ and $V = Y - \sigma p^{-1} W$, where $W \sim \mathcal{N}_n(0, I_n)$ for some $p \in (0, 1)$:
  - if $p \approx 0$, then $U \approx Y$ and the selection will be nearly the same as with the original data, but the inference will be poor because $V \not\approx Y$;
  - if $p \approx 1$, then $V \approx Y$ and the inference will be good but $U \not\approx Y$ so the selection may be very different from that based on Y.
  - Implies context-based trade-off between selection and inference.
- ☐ It can be shown that this beats sample splitting, at least in some special cases.

Example 69 *Discuss randomisation in Example 68.*

Note to Example 69

- ☐ Here $T \sim \mathcal{N}(\theta, 1)$, so we would take $U = T + pW$, where $W \sim \mathcal{N}(0, 1)$ independent of T. Note that if we set $V = T - W/p$, then

$$U \sim \mathcal{N}(\theta, 1 + p^2), \quad V \sim \mathcal{N}(\theta, 1 + 1/p^2), \quad \operatorname{cov}(U, V) = 0,$$

so $U \perp\!\!\!\perp V$, and we can write

$$T = \frac{U + p^2 V}{1 + p^2}.$$

Hence

$$T \mid U = u \stackrel{\mathrm{D}}{=} \frac{u + p^2 V}{1 + p^2},$$

which is equivalent to using the normal distribution of V for inference, as p and u are known.

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'.
- ☐ Very active area of research, likely to keep on changing in next few years.
- ☐ At present it looks like randomisation is a good approach in cases with simple sufficient statistics . . . and asymptotically when σ^2 can be estimated reasonably well.

5 Bootstrap Inference

slide 166

5.1 Basic Notions

slide 167

Parameters and functionals

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

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

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

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

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

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

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

Algorithmic approach

Example 70 *Give general definitions of the median and the parameter obtained from a maximum likelihood fit of a density $f(y;\theta)$. What are the corresponding estimators (a) under a fitted exponential model, and (b) a nonparametric model?*

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

Note to Example 70

☐ The usual definition of the p quantile is

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

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

☐ The maximum likelihood estimator is defined as

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

which we earlier called θ_g.

☐ Under an exponential model

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

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

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

note that f is not necessarily exponential.

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

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

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

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

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

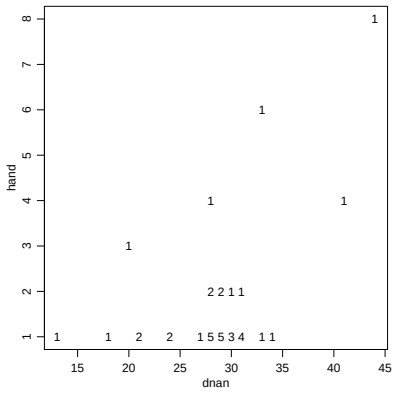
Example: Handedness data

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

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

Example: Handedness data

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



Example: Handedness data

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

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

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

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

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

Bootstrap simulation

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

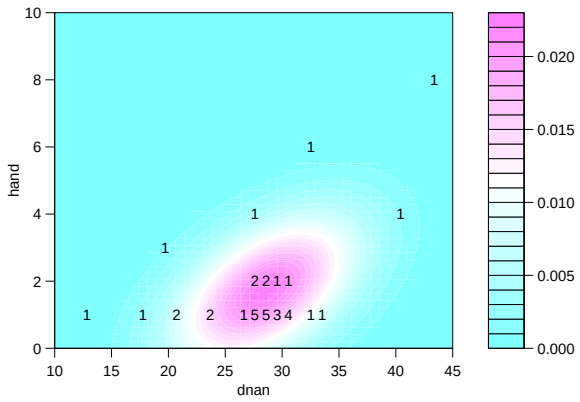
    so the output is a set of **bootstrap replicates**,

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

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

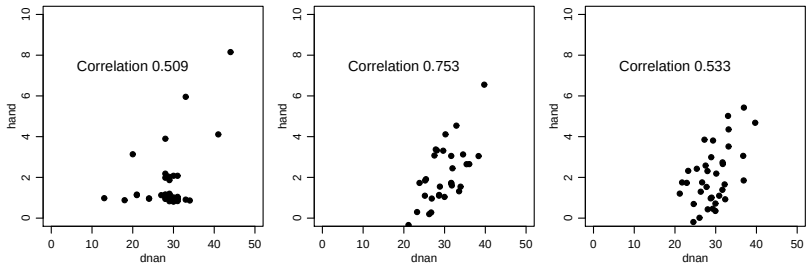
Handedness data: Fitted bivariate normal model

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



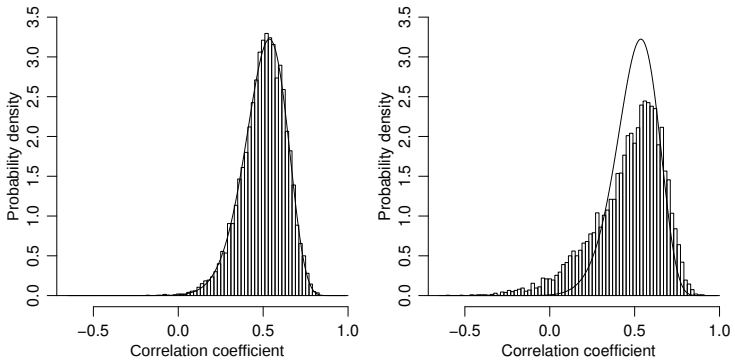
Handedness data: Parametric bootstrap samples

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



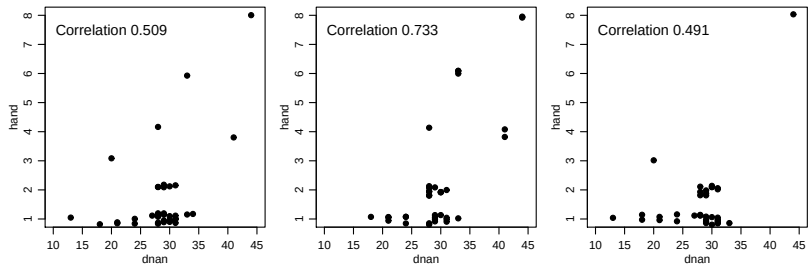
Handedness data: Correlation coefficient

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



Handedness data: Bootstrap samples

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



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

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

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

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

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

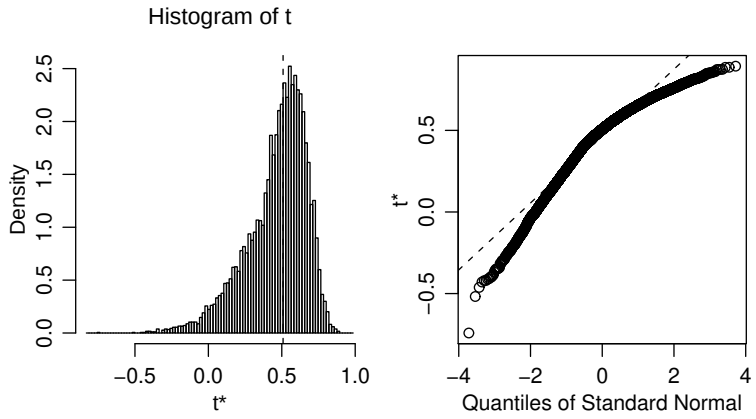
- The Monte Carlo approximations to $\beta(\widehat{G})$ and $\nu(\widehat{G})$ are

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

  For the handedness data, $R = 10^4$ and $b = -0.046$, $v = 0.043 = 0.205^2$.
- We estimate the p **quantile** of $\widehat{\theta}$ using the p quantile of $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$, i.e., $\widehat{\theta}^*_{((R+1)p)}$.

Handedness data

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



Common questions

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

How big should n be?

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

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

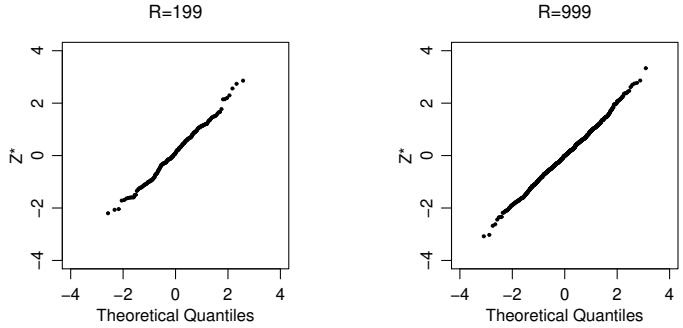
  the most probable of which has probability $p_n = n!/n^n$.
  For $n > 12$, we have $m_n > 10^6$ and $p_n < 6 \times 10^{-5}$.
- Bootstrapping of smooth statistics like the average will often work OK provided $n > 20$.
- For the **median** of a sample of size $n = 2m + 1$, the possible distinct values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

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

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

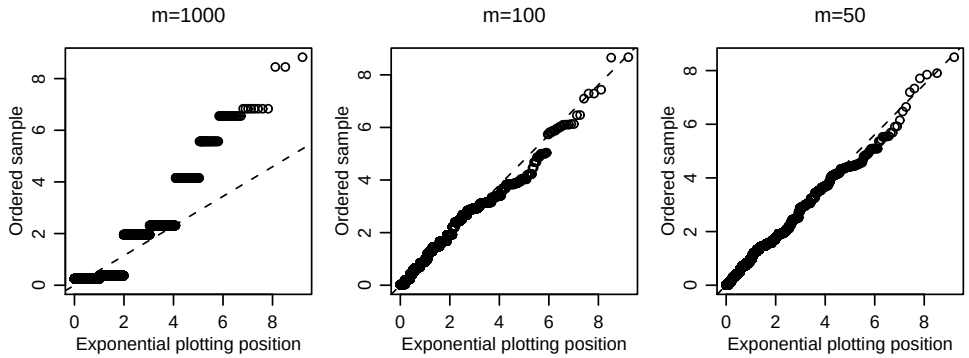
How many bootstraps?

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



Resamples of size n?

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



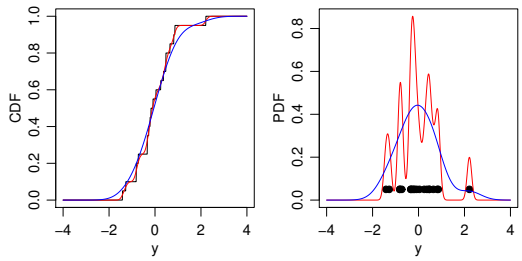
Variants of $\widehat{G}$?

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

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

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

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



When does the bootstrap work?

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

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

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

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

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

Regularity conditions

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

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

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

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

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

Summary

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

Bootstrap confidence Intervals: Desiderata

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

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

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

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

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

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

and the corresponding two-sided interval satisfies

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

Normal confidence intervals

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

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

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

- We replace β, ν by the bootstrap estimates

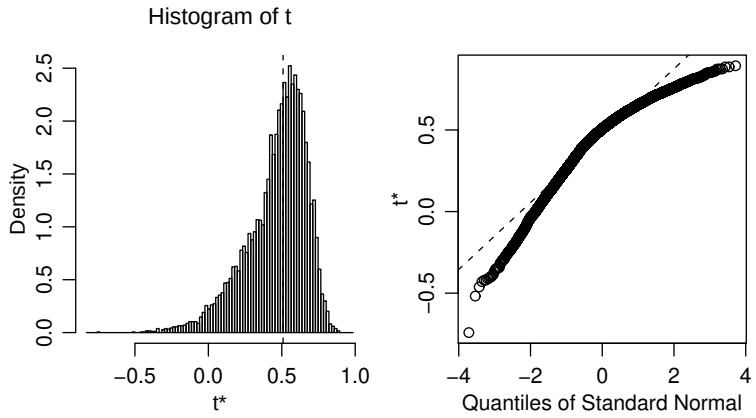
$$\begin{aligned} \beta(G) &\doteq \beta(\widehat{G}) \doteq b = \overline{\widehat{\theta}^*} - \widehat{\theta}, \\ \nu(G) &\doteq \nu(\widehat{G}) \doteq v = (R-1)^{-1} \sum_r (\widehat{\theta}_r^* - \overline{\widehat{\theta}^*})^2, \end{aligned}$$

to get the $(1-2\alpha)$ interval with limits $\widehat{\theta} - b \pm z_\alpha v^{1/2}$.

- For the handedness data we have $R = 10,000$, $b = -0.046$, $v = 0.205^2$, $\alpha = 0.025$, $z_\alpha = -1.96$, so 95% CI is $(0.147, 0.963)$
- We can use the $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ to check the quality of the normal approximation, and perhaps to suggest transformations.

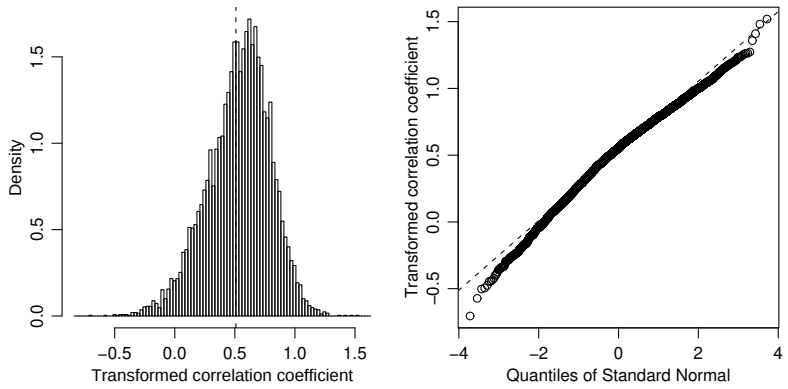
Handedness data

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



Handedness data: Transformed scale?

Plots for $\widehat{\psi}^* = \frac{1}{2}\log\{(1+\widehat{\theta}^*)/(1-\widehat{\theta}^*)\}$:



Normal confidence intervals

- Correlation coefficient: try Fisher's z transformation:

$$\widehat{\psi}^* = \psi(\widehat{\theta}^*) = \tfrac{1}{2}\log\{(1+\widehat{\theta}^*)/(1-\widehat{\theta}^*)\}$$

with bias and variance estimates

$$b_\psi = R^{-1}\sum_{r=1}^{R}\widehat{\psi}_r^* - \widehat{\psi}, \quad v_\psi = \frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{\psi}_r^* - \overline{\widehat{\psi}^*}\right)^2,$$

- Then the $(1-2\alpha)$ confidence interval for θ is

$$\psi^{-1}\left\{\widehat{\psi} - b_\psi - z_{1-\alpha}v_\psi^{1/2}\right\}, \quad \psi^{-1}\left\{\widehat{\psi} - b_\psi - z_\alpha v_\psi^{1/2}\right\}$$

- For handedness data, get $(0.074, 0.804)$ ... but how do we choose a transformation in general?

Pivots

- Assume properties of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$ mimic effect of sampling from original model (plug-in principle) — false in general, but more nearly true for pivots.
- **Pivot** is combination of data and parameter whose distribution is independent of underlying model, such as t statistic

$$Z = \frac{\overline{Y} - \mu}{(S^2/n)^{1/2}} \quad \sim \quad t_{n-1},$$

when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

- Exact pivot generally unavailable in nonparametric case, but if we can estimate the variance of $\widehat{\theta}^*$ using V, we use

$$Z = \frac{\widehat{\theta} - \theta}{V^{1/2}}$$

- If the quantiles z_α of Z known, then

$$\mathrm{P}\left(z_\alpha \leq Z \leq z_{1-\alpha}\right) = \mathrm{P}\left(z_\alpha \leq \frac{\widehat{\theta} - \theta}{V^{1/2}} \leq z_{1-\alpha}\right) = 1 - 2\alpha$$

(z_α no longer denotes a normal quantile!) gives $(1-2\alpha)$ CI $(\widehat{\theta} - V^{1/2}z_{1-\alpha}, \widehat{\theta} - V^{1/2}z_\alpha)$

Studentized statistic

- Bootstrap sample gives $(\widehat{\theta}^*, V^*)$ and hence

$$Z^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{V^{*1/2}}.$$

- We bootstrap to get R copies of $(\widehat{\theta}, V)$, i.e.,

$$(\widehat{\theta}^*_1, V^*_1), \quad (\widehat{\theta}^*_2, V^*_2), \quad \ldots, \quad (\widehat{\theta}^*_R, V^*_R),$$

  and the corresponding

$$z^*_1 = \frac{\widehat{\theta}^*_1 - \widehat{\theta}}{V^{*1/2}_1}, \quad z^*_2 = \frac{\widehat{\theta}^*_2 - \widehat{\theta}}{V^{*1/2}_2}, \quad \ldots, \quad z^*_R = \frac{\widehat{\theta}^*_R - \widehat{\theta}}{V^{*1/2}_R},$$

  then order these to estimate quantiles of Z, with z_p estimated by $z^*_{(p(R+1))}$.
- Get $(1-2\alpha)$ **Studentized bootstrap confidence interval**

$$\widehat{\theta} - V^{1/2} z^*_{((1-\alpha)(R+1))}, \quad \widehat{\theta} - V^{1/2} z^*_{(\alpha(R+1))}.$$

- This is not invariant to transformation and needs an estimated variance V^*_r for each $\widehat{\theta}^*_r$.

Why Studentize?

- If we Studentize, then $Z \xrightarrow{D} N(0,1)$ as $n \to \infty$, and we can use Edgeworth series to write

$$\mathrm{P}_G(Z \le z) = \Phi(z) + n^{-1/2} a(z)\phi(z) + O(n^{-1}),$$

  where $a(\cdot)$ is an even quadratic polynomial.
- For example, if we use $\widehat{\theta} = \overline{Y}$ and $V = n^{-1}S^2$ to compute Z for data with skewness γ, then $a(x) = \gamma(2x^2+1)/6$ and (next slide) $a'(x) = -\gamma(x^2-1)/6$.
- The corresponding expansion for Z^* is

$$\mathrm{P}_{\widehat{G}}(Z^* \le z) = \Phi(z) + n^{-1/2} \widehat{a}(z)\phi(z) + O_p(n^{-1}).$$

- Typically $\widehat{a}(z) = a(z) + O_p(n^{-1/2})$, so

$$\mathrm{P}_{\widehat{G}}(Z^* \le z) - \mathrm{P}_G(Z \le z) = O_p(n^{-1}),$$

  so the order of error is n^{-1}.

Why Studentize? II

- Without Studentization, $Z = n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, \nu')$, and then

$$\mathrm{P}_G(Z \leq z) = \Phi\left(\frac{z}{\nu'^{1/2}}\right) + n^{-1/2} a' \left(\frac{z}{\nu'^{1/2}}\right) \phi\left(\frac{z}{\nu'^{1/2}}\right) + O(n^{-1})$$

  and

$$\mathrm{P}_{\widehat{G}}(Z^* \leq z) = \Phi\left(\frac{z}{\widehat{\nu}'^{1/2}}\right) + n^{-1/2} \widehat{a}' \left(\frac{z}{\widehat{\nu}'^{1/2}}\right) \phi\left(\frac{z}{\widehat{\nu}'^{1/2}}\right) + O_p(n^{-1}).$$

- Typically $\widehat{\nu}' = \nu' + O_p(n^{-1/2})$, giving

$$\mathrm{P}_{\widehat{G}}(Z^* \leq z) - \mathrm{P}_G(Z \leq z) = O_p(n^{-1/2}),$$

  and the difference in the leading terms means that the overall error is of order $n^{-1/2}$.
- Thus Studentizing reduces error from $O_p(n^{-1/2})$ to $O_p(n^{-1})$: better than using large-sample asymptotics, for which error is usually $O_p(n^{-1/2})$.

Other confidence intervals

- Simpler approaches:
  - **Basic bootstrap** interval: treat $\widehat{\theta} - \theta$ as pivot, get

$$\widehat{\theta} - (\widehat{\theta}^*_{((R+1)(1-\alpha))} - \widehat{\theta}), \quad \widehat{\theta} - (\widehat{\theta}^*_{((R+1)\alpha)} - \widehat{\theta}).$$

  - **Percentile interval**: use empirical quantiles of $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$:

$$\widehat{\theta}^*_{((R+1)\alpha)}, \quad \widehat{\theta}^*_{((R+1)(1-\alpha))}.$$

- The percentile interval is transformation-invariant, not the basic bootstrap interval.
- **Bias-corrected and accelerated (BC_a)** intervals replace percentile interval with $(\widehat{\theta}^*_{((R+1)\alpha')}, \widehat{\theta}^*_{((R+1)(1-\alpha''))})$, where

$$\alpha' = \Phi\left\{w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right\}, \quad w = \Phi^{-1}\left\{\widehat{G}^*(\widehat{\theta})\right\}, \quad a = \frac{1}{6}\frac{\sum_{j=1}^n l_j^3}{\left(\sum_{j=1}^n l_j^2\right)^{3/2}},$$

  with $\widehat{G}^*$ the EDF of the $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_R$, and $l_1, \ldots, l_n$ the empirical influence values (soon).
- If the **bias** $w = 0$, then $\widehat{G}^*(\widehat{\theta}) = \frac{1}{2}$, so $\widehat{\theta}$ is at the median of the EDF of $\widehat{\theta}^*$
- If the **acceleration** $a = 0$, then the effect of the data $y_1, \ldots, y_n$ on $\widehat{\theta}$ is symmetric.

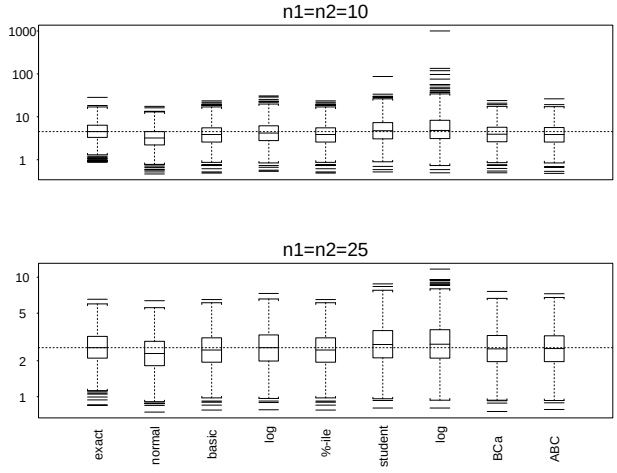
Comparisons

Table 2: Empirical error rates (%) for nonparametric bootstrap confidence limits in ratio estimation: rates for sample sizes $n_1 = n_2 = 10$ are given above those for sample sizes $n_1 = n_2 = 25$. $R = 999$ for all bootstrap methods. 10,000 data sets generated from Gamma distributions.

Method	Nominal error rate							
	Lower limit				Upper limit			
	1	2.5	5	10	10	5	2.5	1
Exact	1.0	2.8	5.5	10.5	9.8	4.8	2.6	1.0
	1.0	2.3	4.8	9.9	10.2	4.9	2.5	1.1
Normal approximation	0.1	0.5	1.7	6.3	20.6	15.7	12.5	9.6
	0.1	0.5	2.1	6.4	16.3	11.5	8.2	5.5
Basic bootstrap	0.0	0.0	0.2	1.8	24.4	21.0	18.6	16.4
	0.0	0.1	0.4	3.0	19.2	15.0	12.5	10.3
Basic bootstrap, log scale	2.6	4.9	8.1	12.9	13.1	7.5	4.8	2.5
	1.6	3.2	6.0	11.4	11.5	6.3	3.3	1.7
Studentized bootstrap	0.6	2.1	4.6	9.9	11.9	6.7	4.0	2.0
	0.8	2.3	4.6	9.9	10.9	5.9	3.0	1.4
Studentized bootstrap, log scale	1.1	2.8	5.6	10.7	11.6	6.3	3.5	1.7
	1.1	2.5	5.0	10.1	10.8	5.7	2.9	1.3
Bootstrap percentile	1.8	3.6	6.5	11.6	14.6	8.9	5.9	3.3
	1.2	2.6	5.1	10.1	12.6	7.1	4.2	2.1
BC_a	1.9	4.0	6.9	12.3	14.0	8.3	5.3	3.0
	1.4	3.0	5.6	10.9	11.8	6.8	3.8	1.9
ABC	1.9	4.2	7.4	12.7	14.6	8.7	5.5	3.1
	1.3	3.0	5.7	11.0	12.1	6.8	3.7	1.9

Confidence interval lengths

Lengths of 95% confidence intervals for the first 1000 simulated samples in the numerical experiment with Gamma data.



Discussion

- ☐ Bootstrap confidence intervals usually under-cover (i.e., are too short).
- ☐ Normal, basic, and studentized intervals depend on scale.
- ☐ Percentile interval often too short but is transformation-invariant.
- ☐ Studentized intervals give best coverage overall, but
  - they depend on scale, can be sensitive to V;
  - their lengths can be very variable;
  - they are best when V is approximately constant.
- ☐ Improved percentile intervals have same asymptotic error as Studentized intervals, but often are shorter, so give lower coverage probabilities.
- ☐ Caution: Edgeworth theory OK for smooth statistics, but beware rough statistics: must check output.
- ☐ Typically need $R > 1000$ for reliable estimation of quantiles.

5.3 Nonparametric Delta Method

Nonparametric delta method

- ☐ The **delta method** (Theorem 11) gives variance formulae for functions of averages.
- ☐ More generally we use the **nonparametric delta method**, which is based on the linear functional expansion

$$t(F) \doteq t(G) + \int L_t(x; G)\, \mathrm{d}F(x),$$

where L_t, the first derivative of $t(\cdot)$ at G, is defined by

$$L_t(y; G) = \lim_{\varepsilon \to 0} \frac{t\{(1-\varepsilon)G + \varepsilon H_y\} - t(G)}{\varepsilon} = \left. \frac{\partial t\,\{(1-\varepsilon)G + \varepsilon H_y\}}{\partial \varepsilon} \right|_{\varepsilon=0},$$

with $H_y(u) \equiv H(u - y)$ the Heaviside function jumping from 0 to 1 at $u = y$.
- ☐ The **influence function value** $L_t(y; G)$ for the statistical functional t for an observation at y when the background distribution is G, satisfies $\mathrm{E}_G\{L_t(Y; G)\} = 0$.
- ☐ If $\widehat{G}$ is based on a random sample $y_1, \ldots, y_n$, then the jth **empirical influence value** is

$$l_j = L_t(y_j; \widehat{G}),$$

and $\mathrm{E}_{\widehat{G}}\{L_t(Y; \widehat{G})\} = n^{-1} \sum_j l_j = 0$.
- ☐ The influence function also plays an important role in robust statistics.

Nonparametric delta method II

☐ If we replace F by the EDF $\widehat{G}$ for a random sample $Y_1, \ldots, Y_n$, then

$$t(\widehat{G}) \doteq t(G) + \int L_t(x; G)\, \mathrm{d}\widehat{G}(x) = t(G) + \frac{1}{n}\sum_{j=1}^{n} L_t(Y_j; G),$$

has variance

$$\mathrm{var}\{t(\widehat{G})\} \doteq \frac{1}{n^2}\sum_{j=1}^{n} L_t^2(Y_j; G) = V_L,$$

say, which we estimate based on a sample $y_1, \ldots, y_n$ by $v_L = n^{-2}\sum l_j^2$.

Example 71 *Apply the nonparametric delta method to the average $\overline{Y}$.*

Example 72 *Apply the nonparametric delta method to a statistic defined by an estimating equation, and hence find the variance of the ratio $\overline{V}/\overline{U}$ for data pairs $Y = (U, V)$.*

Note to Example 71

☐ The population mean and its empirical version are

$$\theta = t(G) = \int x\, dG(x), \quad \widehat{\theta} = t(\widehat{G}) = \int x\, d\widehat{G}(x) = n^{-1}\sum_{j=1}^{n} Y_j = \overline{Y}.$$

☐ If H_y puts unit mass at y, its 'density' is a Dirac delta function $\delta_y(x)$, and

$$\begin{aligned} \theta\{(1-\varepsilon)G + \varepsilon H_y\} &= \int x\, \mathrm{d}\{(1-\varepsilon)G + \varepsilon H_y\}(x) \\ &= (1-\varepsilon)\int x\, \mathrm{d}G(x) + \varepsilon \int x\, \mathrm{d}H_y(x) = (1-\varepsilon)\theta(G) + \varepsilon y \end{aligned}$$

and therefore

$$L(y; G) = \lim_{\varepsilon \to 0} \frac{\theta\{(1-\varepsilon)G + \varepsilon H_y\} - \theta(G)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\theta(G) + \varepsilon y - \theta(G)}{\varepsilon} = y - \theta(G),$$

☐ Hence the empirical influence values and variance estimate are

$$l_j = L(y_j; \widehat{G}) = y_j - \overline{y}, \qquad v_L = \frac{1}{n^2}\sum (y_j - \overline{y})^2 = \frac{n-1}{n} n^{-1} s^2.$$

Note to Example 72

☐ The scalar parameter $\theta = t(G)$ is determined implicitly through the estimating equation

$$\int a(x;\theta)\,\mathrm{d}G(x) = \int a\{x;t(G)\}\,\mathrm{d}G(x) = 0.$$

We replace G by $G_\varepsilon = (1-\varepsilon)G + \varepsilon H_y$ and see that

$$\begin{aligned} 0 &= \int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G_\varepsilon(x) \\ &= (1-\varepsilon)\int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G(x) + \varepsilon\int a\{x;t(G_\varepsilon)\}\,\mathrm{d}H_y(x) \\ &= (1-\varepsilon)\int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G(x) + \varepsilon a\{y;t(G_\varepsilon)\}, \end{aligned}$$

and differentiation using the chain rule gives

$$0 = a\{y;t(G_\varepsilon)\} - \int a\{x;t(G_\varepsilon)\}\,\mathrm{d}G(x) + \varepsilon a_\theta\{y;t(G_\varepsilon)\}\frac{\partial t(G_\varepsilon)}{\partial\varepsilon} + (1-\varepsilon)\int a_\theta\{x;t(G_\varepsilon)\}\frac{\partial t(G_\varepsilon)}{\partial\varepsilon}\,\mathrm{d}G(x),$$

which reduces to

$$0 = a\{y;t(G)\} + \int a_\theta\{x;t(G)\}\,\mathrm{d}G(x)\,\frac{\partial t(G)}{\partial\varepsilon}\bigg|_{\varepsilon=0}$$

on setting $\varepsilon = 0$. Hence

$$L_t(y;G) = \frac{\partial t(G_\varepsilon)}{\partial\varepsilon}\bigg|_{\varepsilon=0} = \frac{a(y;\theta)}{-\int a_\theta(x;\theta)\,\mathrm{d}G(x)}, \quad \text{where} \quad a_\theta(x;\theta) = \frac{\partial a(x;\theta)}{\partial\theta}.$$

☐ In the case of the ratio and with $y = (u,v)$, we take $a(y;\theta) = v - \theta u$, so

$$\theta = \theta(G) = \int v\,\mathrm{d}G(u,v) \Big/ \int u\,\mathrm{d}G(u,v), \quad \widehat{\theta} = \overline{v}/\overline{u},$$

and $a_\theta = -u$, so $l_j = (x_j - \widehat{\theta}u_j)/\overline{u}$, giving

$$v_L = \frac{1}{n^2}\sum\left(\frac{x_j - \widehat{\theta}u_j}{\overline{u}}\right)^2.$$

Comments

- For statistics involving only averages (ratio, correlation coefficient, ...), the nonparametric delta method retrieves the delta method.
- For example, the correlation coefficient may be written as a function of $\overline{xu} = n^{-1} \sum x_j u_j$, etc.:

$$\widehat{\theta} = \frac{\overline{xu} - \overline{x}\,\overline{u}}{\left\{(\overline{x^2} - \overline{x}^2)(\overline{u^2} - \overline{u}^2)\right\}^{1/2}},$$

  from which empirical influence values l_j can be derived, giving $v_L = 0.029$ for the handedness data, to be compared with $v = 0.043$ obtained by bootstrapping.
- v_L typically underestimates $\mathrm{var}(\widehat{\theta})$!
- The l_j can also be obtained by numerical differentiation if $t(\widehat{G})$ is coded appropriately, or approximated using a jackknife method.