apts.ac.uk
Academy for PhD Training in Statistics

APTS module *Statistical Inference*

Cambridge, January 2011

Rough plan:

1. Introduction: Some basic ideas of inference
2. Sufficiency and conditioning
3. Maximum likelihood
4. Some non-standard situations
5. Frequentist and Bayesian approaches: some similarities and differences
6. Some non-likelihood approaches

Skeletal notes are attached for much of parts 1, 2, 3 and 6; this should allow us to progress more quickly in those parts.

Students are advised to make their own notes on the parts where skeleton notes are not provided here, and to supplement the lectures by broader reading. Some suggestions for further reading are made at the end.

David Firth
January 2011

Part 1
Introduction: some basic ideas of inference

1 Role of (a) theory of inference

Objective: provide concepts and methods helpful for science, technology, public affairs, etc.

Variety of problems ⇒ various approaches

(Statistics unlike many other branches of science — less specialized.)

Ultimate criteria for any analysis:

- informativeness
- relevance
  - to the scientific question
  - to the data at hand

An idealized scheme:

- research question or questions
- measurement issues
- study design
- data collection

- preliminary analysis / data editing
- more formal (probabilistic) analysis
  - specific methods
  - general principles
- conclusions and interpretation
- (usually) more questions

Formal theory of inference:

- brings a systematic approach to what might otherwise be fragmentary
- provides a basis for studying new problems

[Here we emphasize the sequence

questions — data — analysis

In 'data mining' applications the sequence may be closer to

data — analysis — questions

The same methods/principles may still be useful; but care needed with any probabilistic interpretation.]

2 Formulation: statistical model

- Data $\{y; x\}$
- Model: y is the value of a random vector Y with pdf/pmf $f_Y(y; x, \theta)$, where θ is unknown.
- Choice of split between y and x
- Interpretation of model (especially of probability)

- Types of model (parametric, semi-parametric, non-parametric); e.g., for simple regression,

  $$Y_k = \alpha + \beta z_k + e_k \qquad\qquad e_k \sim N(0, \sigma^2)$$

  or

  $$Y_k = w(z_k) + e_k \qquad\qquad w \text{ monotonic}, e_k \text{ as before}$$

  or

  $$Y_k = w(z_k) + e_k \qquad\qquad w \text{ monotonic}, e_k \text{ has median } 0$$

  etc.

We will concentrate here on *parametric* models.

Often drop x/z, and just write $f_Y(y;\theta)$.

Typically

$$\theta = (\psi, \lambda)$$

where ψ is of primary interest, while λ represents 'incidental' or 'nuisance' parameters.

Objective: what can we learn about ψ (and hence about interesting aspects of the real world or assumed data generating process) from y?

Some other possible objectives: decision, prediction.

Model specification is crucial — translates a subject-matter question into a statistical one.

(Sometimes model represents d.g.p.; sometimes it's just empirically descriptive.)

Parameter θ aims to capture features of the system under study.

Model is an *idealized representation* of variation in the physical, biological, social, ..., world. Probabilities represent limiting frequencies under (often hypothetical) repetition.

Specific possible objectives:

(i) Assuming that the model is a sound basis for inference:

- set(s) of values within which ψ is likely to lie?
- consistency of y with a particular ψ_0?
- 'predict' as yet unobserved values of y from the same system
- use the data to make a choice among possible *decisions* (requires consequences of decisions)

(ii) Use the data to criticize the model.

Here we focus mainly on the first two items in (i); we'll also look more briefly at (ii).

3 Two broad approaches

- 'Frequentist' (sampling theory) — probability is constrained to mean a (usually hypothetical) frequency
- 'Bayesian' (inverse probability) — notion of probability is extended to cover assessment of *any* uncertain event or proposition.

Some contrasts:

Frequentist: θ is an unknown *constant.* Calibrate procedures for assessing evidence on θ by how they would perform if used repeatedly — just like other kinds of measuring instrument.

Requires notion of a 'long run' of similar situations where a given procedure is used. Important to ensure that the long run is *relevant* to the specific instance. (role of conditioning, ancillarity)

(cf. calibration of an instrument to operate well at a *specific* temperature)

Bayesian: θ is a random variable, with a specified (prior) distribution. Inference is via application of Bayes theorem:

$$f_{\Theta|y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)f_{\Theta}(\theta)}{\int f_{Y|\Theta}(y|\theta')f_{\Theta}(\theta')d\theta'}$$

— the posterior density for Θ.

Inference on the parameter(s) Ψ of specific interest is then by integration.

In a Bayesian analysis:

- all calculations are by the laws of probability
- *any* uncertain event has a probability, *and* these probabilities follow the standard rules
- a *prior* distribution must be specified
- relevance to the observed y is *assured* (by 'automatic' conditioning).

Example: normal with known variance

$$Y_i|M \sim N(M, \sigma_0^2) \qquad (i = 1, \ldots, n;\ \text{i.i.d.})$$

$$M \sim N(m, v) \qquad \text{(conjugate prior)}$$

Apply Bayes theorem (exercise) to get

$$M|Y \sim N\left[\frac{\bar{y}/(\sigma_0^2/n) + m/v}{1/(\sigma_0^2/n) + 1/v}, \frac{1}{1/(\sigma_0^2/n) + 1/v}\right]$$

A useful definition: *pivotal quantity*

Suppose that $p(y, \psi)$ is monotonic in ψ, for any fixed y.

Then

(i) p is a *frequentist pivot* if for every fixed θ the distribution of $p(Y; \psi)$ is fixed and known (e.g., $N(0, 1)$, t_m, etc.)

(ii) p is a *Bayesian pivot* if for every y the distribution of $p(y; \Psi)$ in the posterior distribution, given $Y = y$, is fixed and known.

Use of a pivot: in either case, use a pivot to determine upper (or lower) limits on ψ (or Ψ) in terms of data y, at any specified level.

Example continued: (frequentist)

$$p(Y, \mu) = \frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1).$$

Hence

$$\mathrm{pr}(\mu < \bar{Y} + k^*_\epsilon \sigma_0/\sqrt{n}) = 1 - \epsilon$$

by choice of k^*_ϵ.

Substitute the observed $\bar{y}$ to get an upper confidence limit.

(NB: this 'looks like' a probability statement about μ — but it cannot be manipulated as such, so it is not.)

Example continued: (Bayesian)

Bayesian pivot in this example is

$$\left[M - \frac{\bar{y}/(\sigma_0^2/n) + m/v}{1/(\sigma_0^2/n) + 1/v}\right] \sqrt{1/(\sigma_0^2/n) + 1/v}$$

from which we *can* obtain a probability statement about M.

If v is large, in the limit the f- and B-pivots are formally the same, *viz.*

$$\frac{\mu - \bar{Y}}{\sigma_0/\sqrt{n}}, \quad \frac{M - \bar{y}}{\sigma_0/\sqrt{n}}.$$

BUT: the corresponding limiting prior is flat, uniform over the whole real line, so is improper.

(More later on the role of 'flat' priors.)

4 'Relevant'

Various senses:

Frequentist:

- ensure that the 'long run' is relevant to the particular data y

Bayesian:

- is (my) prior f_Θ relevant to someone else's inference?

Both:

- is the model relevant to the scientific questions?
- is our inference robust to failure of (parts of) the model?

(In essence, these points form the agenda for most of this APTS module.)

A note on interpretation of confidence intervals

Example: Uniform with known range. A confidence interval can sometimes be valid but highly questionable in terms of its relevance!

Part 2
Sufficiency and conditioning

1 Use of a minimal sufficient statistic: some principles

Here ‘sufficient statistic’ will always mean *minimal* sufficient statistic.

Notation:

- random vector Y
- parameter (usually vector) θ
- sometimes $\theta = (\psi, \lambda)$, with ψ of interest and λ nuisance
- symbol f used for pdf, pmf — conditional or marginal as indicated by context (and sometimes explicitly by subscripts).

1.1 Inference on θ

Sufficient statistic S:

$$f_Y(y;\theta) = f_S(s(y);\theta) f_{Y|S}(y|s)$$

where the second factor does not involve θ.

Implications:

- inference for θ based on $f_S(s;\theta)$

- $f_{Y|S}(y|s)$ eliminates θ, and provides a basis for model checking.

Idea here is that S is a substantial reduction of Y.

(At the other extreme, if the minimal sufficient statistic is $S = Y$, the second factor above is degenerate and this route to model-checking is not available.)

1.2 Inference on ψ (free of λ)

Often $\theta = (\psi, \lambda)$, where ψ is the parameter (scalar or vector) of interest, and λ represents one or more nuisance parameters.

Ideal situation: there exists statistic S_λ — a function of the minimal sufficient statistic S — such that, for every fixed value of ψ, S_λ is sufficient for λ. For then we can write

$$f(y;\psi,\lambda) = f_{Y|S_\lambda}(y|s_\lambda;\psi) f_{S_\lambda}(s_\lambda;\psi,\lambda),$$

and inference on ψ can be based on the first factor above.

This kind of factorization is not always possible. But:

- exponential families — exact;
- more generally — approximations.

1.3 Inference on model adequacy (free of θ)

How well does the assumed model $f_Y(y;\theta)$ fit the data?

Now θ is the 'nuisance' quantity to be eliminated.

Suppose that statistic T is designed to measure lack of fit. Ideally, T has a distribution that does not involve θ: a significant value of T relative to that distribution then represents evidence against the model (i.e., against the *family* of distributions $f_Y(y;\theta)$).

Condition on the minimal sufficient statistic for θ: refer T to its conditional distribution $f_{T|S}(t|s)$, which does not depend on θ.

2 Ancillarity

3 Exponential families

Introduced here as the cleanest/simplest class of models in which to explore and exemplify the above principles.

3.1 Introduction: some special types of model

Many (complicated) statistical models used in practice are built upon one or more of these three types of family:

- transformation family;
- mixture family;
- exponential family.

Transformation families and exponential families are excellent models for the purpose of studying general principles. (Mixture families tend to be messier, inferentially speaking.)

Our main focus in the rest of this lecture will be on exponential families. The other two types will be introduced briefly for completeness.

3.1.1 Transformation families

Prime examples of a transformation model are

- *location model*

$$f(y;\theta) = g(y-\theta)$$

- *scale model*

$$f(y;\theta) = \theta^{-1} g(y/\theta)$$

- *location-scale model*

$$f(y;\mu,\tau) = \tau^{-1} g\{(y-\mu)/\tau\}$$

where in each case $g(.)$ is a fixed function (not depending on θ).

Each such model is characterized by a specified group of transformations.

3.1.2 Mixture families

Simplest case: 2-component mixture

$$f(y;\theta) = (1-\theta)f(y;0) + \theta f(y;1) \qquad (0 \leq \theta \leq 1),$$

where $f(y;0)$ and $f(y;1)$ are the specified 'component' distributions.

More generally: any number of components (possibly infinite), with θ indexing a suitable 'mixing' distribution.

Summation of components makes life easy in some respects (normalization is automatic), but much harder in other ways (no factorization of the likelihood).

3.1.3 Exponential families

When the parameter is the canonical parameter of an exponential family (EF), we will call it ϕ instead of θ (merely to remind ourselves).

An EF interpolates between (and extrapolates beyond) component distributions on the scale of $\log f$ (cf. mixtures; interpolation on the scale of f itself). For example, a one-parameter EF constructed from two known components is $f(y;\theta)$ such that

$$\begin{aligned}\log f(y;\phi) &= (1-\phi)\log f(y;0) + \phi \log f(y;1) - k(\phi) \\ &= \phi \log \frac{f(y;1)}{f(y;0)} + \log f(y;0) - k(\phi),\end{aligned}$$

where the $k(\phi)$ is needed in order to normalize the distribution. This is an instance of the general form for an EF (see the preliminary material)

$$f(y;\phi) = m(y)\exp[s^T(y)\phi - k(\phi)].$$

Some EFs are also transformation models [but not many! — indeed, it can be shown that among univariate models there are just *two* families in both classes, namely $N(\mu,\sigma^2)$ (a location-scale family) and the Gamma family with known 'shape' parameter α (a scale family)].

3.2 Canonical parameters, sufficient statistic

Consider a d-dimensional full EF, with canonical parameter vector $\phi = (\phi_1,\ldots,\phi_d)$, and sufficient statistic $S = (S_1,\ldots,S_d)$.

Clearly (from the definition of EF) the components of ϕ and of S are in one-one correspondence.

Suppose now that $\phi = (\psi,\lambda)$, and that the corresponding partition of S is $S = (S_\psi, S_\lambda)$.

It is then immediate that, for each fixed value of ψ, S_λ is sufficient for λ. This is the 'ideal situation' mentioned in 1.2 above.

More specifically:

1. the distribution of S is a full EF with canonical parameter vector ϕ;
2. the conditional distribution of S_ψ, given that $S_\lambda = s_\lambda$, is a full EF with canonical parameter vector ψ.

3.3 Conditional inference on parameter of interest

The key property, of the two just stated, is the second one: the conditional distribution of S_ψ given S_λ is free of λ. This allows 'exact' testing of a hypothesis of the form $\psi = \psi_0$, since the null distribution of any test statistic is (in principle) known — it does not involve the unspecified λ.

Tests → confidence sets.

Note that the canonical parameter vector ϕ can be linearly transformed to $\phi' = L\phi$, say, with L a fixed, invertible $d \times d$ matrix, without disturbing the EF property:

$$s^T\phi = [(L^{-1})^T s]^T (L\phi),$$

so the sufficient statistic after such a re-parameterization is $(L^{-1})^T S = S'$, say. This allows the parameter of interest ψ to be specified as any linear combination, or vector of linear combinations, of $\phi_1, \ldots, \phi_d$.

3.3.1 Example: 2 by 2 table of counts

Counts R_{ij} in cells of a table indexed by two binary variables:

$$\begin{array}{cc|c}
R_{00} & R_{01} & R_{0+} \\
R_{10} & R_{11} & R_{1+} \\
\hline
R_{+0} & R_{+1} & R_{++} = n
\end{array}$$

Several possible sampling mechanisms for this:

- Individuals counted into the four cells as result of random events over a fixed time-period. Model: $R_{ij} \sim \text{Poisson}(\mu_{ij})$ independently. [No totals fixed in the model.]

- *Fixed* number n of individuals counted into the four cells. Model: $(R_{00}, R_{01}, R_{10}, R_{11}) \sim \text{Multinomial}(n; \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$. [Grand total, n, fixed in the model]

- Row variable is treatment (present/absent), column variable is binary response. Numbers treated and untreated are fixed ($R_{0+} = n_0$, $R_{1+} = n_1$, say). Model: $R_{i0} \sim \text{Binomial}(n_i; \pi_i)$ $(i = 0, 1)$. [Row totals fixed in the model]

In each case the model is a full EF. Take the (canonical) parameter of interest to be

$$\psi = \log \frac{\mu_{11}\mu_{00}}{\mu_{10}\mu_{01}},$$

where $\mu_{ij} = E(R_{ij})$. In the pair-of-binomials model this is the log odds ratio.

In each case the relevant conditional distribution for inference on ψ turns out to be the same. It can be expressed as the distribution of R_{11}, say, conditional upon the observed values of all four marginal totals $M = \{R_{0+}, R_{1+}, R_{+0}, R_{+1}\}$:

$$\text{pr}(R_{11} = r_{11} | M) = \frac{\binom{r_{0+}}{r_{01}}\binom{r_{1+}}{r_{11}} \exp(r_{11}\psi)}{\sum \binom{r_{0+}}{r_{+1}-w}\binom{r_{1+}}{w} \exp(w\psi)}$$

— a *generalized hypergeometric* distribution.

When $\psi = 0$, this reduces to the ordinary hypergeometric distribution, and the test of $\psi = 0$ based on that distribution is known as *Fisher's exact test.*

The practical outcome (condition on all four marginal totals for inference on ψ) is thus the same for all 3 sampling mechanisms. But there are two distinct sources of conditioning at work:

Conditioning by model formulation: the multinomial model conditions on n; the pair-of-binomials model conditions on $r_{0+} = n_0, r_{1+} = n_1$.

'Technical' conditioning (to eliminate nuisance parameters) applies in all 3 models; the numbers of nuisance parameters eliminated are 3, 2 and 1 respectively.

3.3.2 Example: Several 2 by 2 tables

(The Mantel-Haenszel procedure)

Extend the previous example: m independent 2×2 tables, with assumed common log odds ratio ψ.

Pair-of-binomials model for each table: canonical parameters (log odds) for table k are

$$\phi_{k0} = \alpha_k, \qquad \phi_{k1} = \alpha_k + \psi.$$

Parameters $\alpha_1, \ldots, \alpha_m$ are nuisance. Eliminate by (technical) conditioning on all of the individual column totals, as well as conditioning (as part of the model formulation) on all the row totals.

Resulting conditional distribution is the distribution of $S_\psi = \sum R_{k.11}$ conditional upon all row and column totals — the convolution of m generalized hypergeometric distributions.

In practice (justified by asymptotic arguments), the 'exact' conditional distribution for testing $\psi = 0$ — the convolution of m hypergeometrics — is usually approximated by the normal with matching mean and variance.

3.3.3 Example: binary matched pairs

Extreme case of previous example: row totals $r_{k.0+}, r_{k.1+}$ are all 1.

Each table is a pair of independent *binary* observations (e.g., binary response before and after treatment).

Conditional upon column totals: only ‘mixed’ pairs k, with $r_{k.+0} = r_{k.+1} = 1$, carry any information at all.

Conditional distribution for inference on ψ is binomial. (see exercises)

This is an example where conditional inference is a *big* improvement upon use of the unconditional likelihood: e.g., the unconditional MLE $\hat{\psi}$ is inconsistent as $m \to \infty$, its limit in probability being 2ψ rather than ψ.

3.4 Conditional test of model adequacy

The principle: refer any proposed lack-of-fit statistic to its distribution conditional upon the minimal sufficient statistic for the model parameter(s).

We mention here just a couple of fairly simple examples, to illustrate the principle in action.

3.4.1 Example: Fit of Poisson model for counts

(Fisher, 1950)

Testing fit of a Poisson model.

Conditional distribution of lack-of-fit statistic given MLE (which is minimal sufficient since the model is a full EF).

Calculation quite complicated but ‘do-able’ in this simple example.

3.4.2 Example: Fit of a binary logistic regression model

A standard lack-of-fit statistic in generalized linear models is the *deviance*, which is twice the log likelihood difference between the fitted model and a ‘saturated’ model.

In the case of independent binary responses y_i the deviance statistic for a logistic regression with maximum-likelihood fitted probabilities $\hat{\pi}_i$ is

$$\begin{aligned} D &= 2\sum\left\{y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right\} \\ &= 2\sum\left\{y_i \log y_i + (1-y_i)\log(1-y_i) - y_i \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) - \log(1-\hat{\pi}_i)\right\} \end{aligned}$$

Since y is 0 or 1, the first two terms are both zero. Since the fitted log odds is $\log\{\hat{\pi}_i/(1-\hat{\pi}_i)\} = x_i^T\hat{\beta}$, the deviance can be written as

$$\begin{aligned} D &= -2\hat{\beta}^T X^T Y - 2\sum \log(1-\hat{\pi}_i) \\ &= -2\hat{\beta}^T X^T \hat{\pi} - 2\sum \log(1-\hat{\pi}_i), \end{aligned}$$

since the MLE solves $X^T Y = X^T \hat{\pi}$.

Hence D in this (binary-response) case is a function of $\hat{\beta}$, which is equivalent to the minimal sufficient statistic.

The required conditional distribution of D is thus *degenerate*. The deviance statistic carries *no information at all* regarding lack of fit of the model.

The same applies, not much less severely, to other general-purpose lack of fit statistics such as the 'Pearson chi-squared' statistic $X^2 = \sum(y_i - \hat{\pi})^2/\{\hat{\pi}_i(1-\hat{\pi}_i)\}$.

This is a common source of error in applications.

The above (i.e., the case of binary response) is an extreme situation. In logistic regressions where the binary responses are *grouped*, the lack-of-fit statistics usually have non-degenerate distributions; but when the groups are small it will be important to use (at least an approximation to) the conditional distribution given $\hat{\beta}$, to avoid a potentially misleading result.

EXERCISE

For the binary matched pairs model, derive the conditional binomial distribution for inference on the common log odds ratio ψ. Discuss whether it is reasonable to discard all the data from ‘non-mixed’ pairs.

Part 3
Maximum likelihood

Scalar parameter

Score function:

$$U = \frac{\partial l(\theta; Y)}{\partial \theta}$$

— a random function of θ.

The score has mean zero at the true value of θ (subject to regularity condition).

Regularity: can validly differentiate under the integral sign the normalizing condition

$$\int f_Y(y; \theta) dy = 1$$

so that

$$\int U(\theta; y) f_Y(y; \theta) dy = 0,$$

i.e.,

$$E[U(\theta; Y); \theta] = 0.$$

MLE

Maximum likelihood estimator (MLE): taken here to be $\hat{\theta}$ which solves

$$U(\hat{\theta}; Y) = 0,$$

(or the solution giving largest l if there is more than one)

— a random variable.

We will not discuss (here) situations where the value of θ that maximizes the likelihood is not a solution of the *score equation* as above.

Observed information

Observed information measures curvature (as a function of θ) of the log likelihood:

$$j(\theta) = -\frac{\partial U}{\partial \theta} = -\frac{\partial^2 l}{\partial \theta^2}$$

— the [in general, random] curvature of $l(\theta; Y)$ at θ.

High curvature $\hat{j} = j(\hat{\theta})$ indicates a well-determined MLE.

Expected information

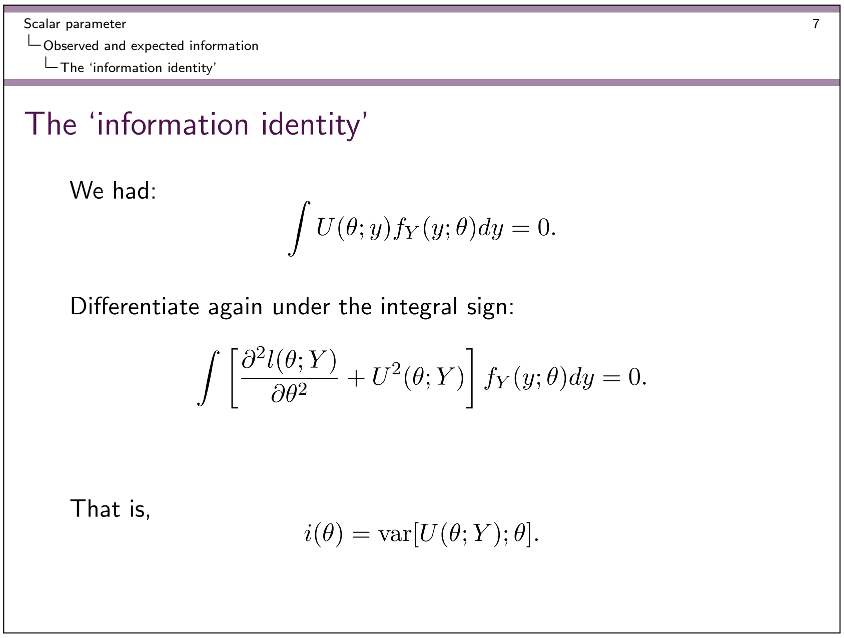
In most models, $j(\theta)$ is random — a function of Y.

The *expected* information is

$$\begin{aligned} i(\theta) &= E[j(\theta); \theta] \\ &= E\left[-\frac{\partial^2 l}{\partial \theta^2};\ \theta\right] \end{aligned}$$

— a repeated-sampling property of the likelihood for θ; important in asymptotic approximations.

Expected information is also known as *Fisher information*.



The 'information identity'

We had:

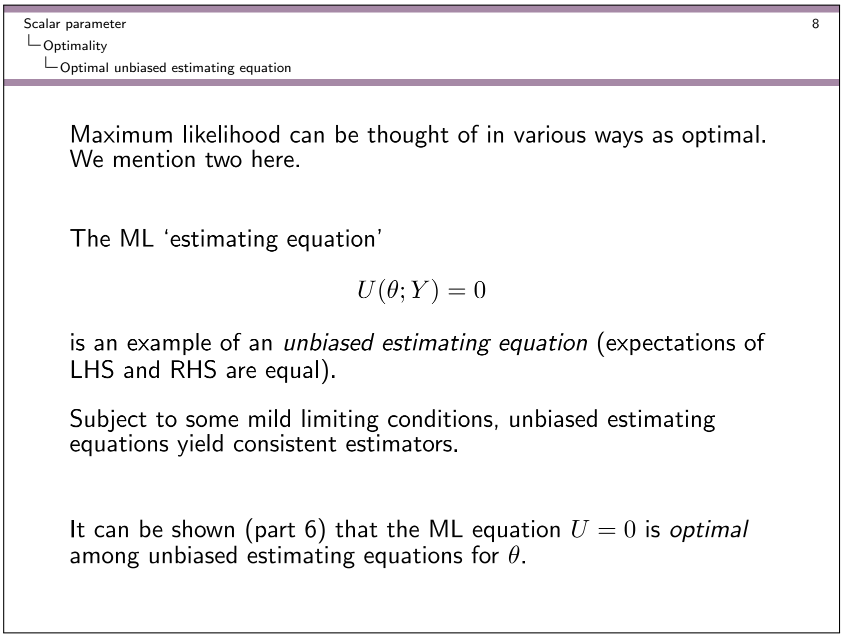
$$\int U(\theta;y)f_Y(y;\theta)dy = 0.$$

Differentiate again under the integral sign:

$$\int \left[\frac{\partial^2 l(\theta;Y)}{\partial\theta^2} + U^2(\theta;Y)\right] f_Y(y;\theta)dy = 0.$$

That is,

$$i(\theta) = \mathrm{var}[U(\theta;Y);\theta].$$



Maximum likelihood can be thought of in various ways as optimal. We mention two here.

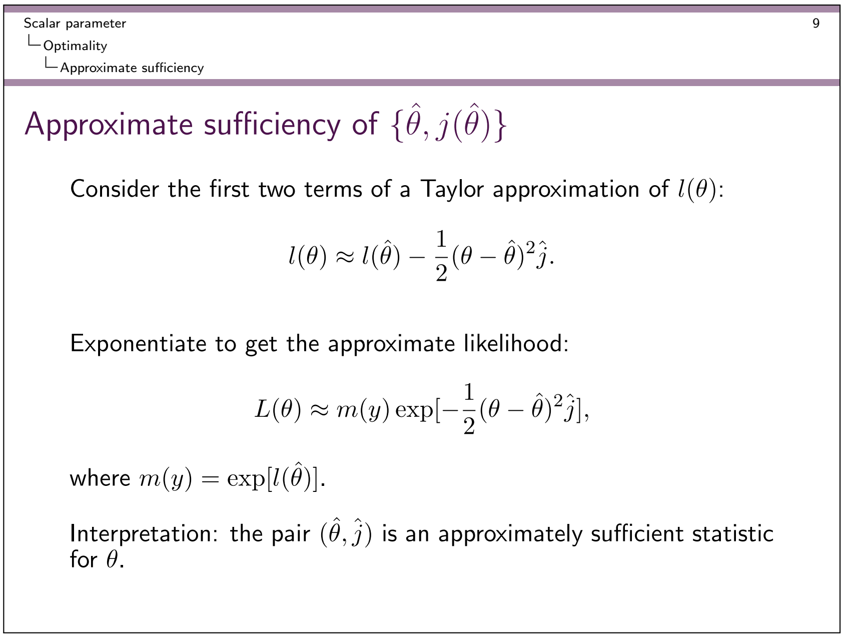
The ML 'estimating equation'

$$U(\theta;Y) = 0$$

is an example of an *unbiased estimating equation* (expectations of LHS and RHS are equal).

Subject to some mild limiting conditions, unbiased estimating equations yield consistent estimators.

It can be shown (part 6) that the ML equation $U = 0$ is *optimal* among unbiased estimating equations for θ.



Approximate sufficiency of $\{\hat\theta, j(\hat\theta)\}$

Consider the first two terms of a Taylor approximation of $l(\theta)$:

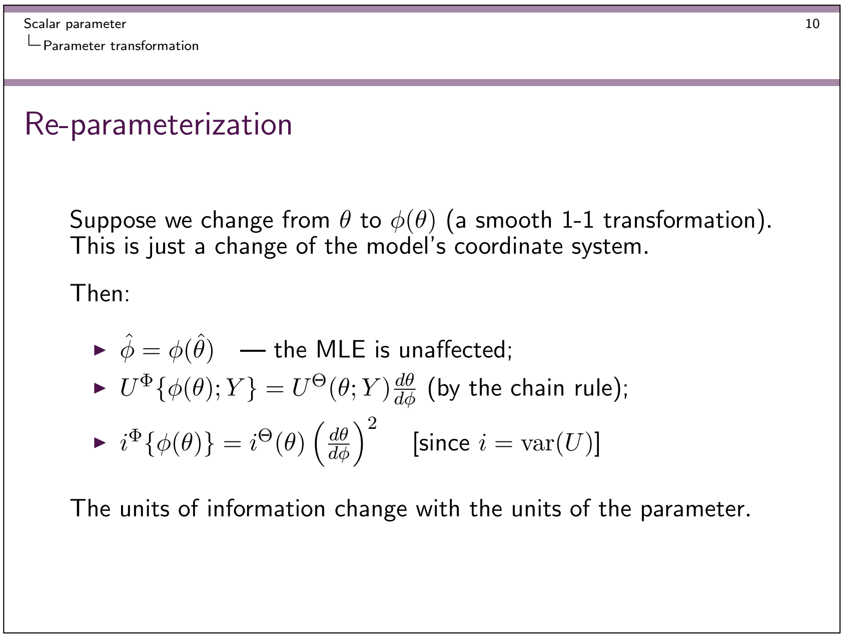
$$l(\theta) \approx l(\hat\theta) - \frac{1}{2}(\theta - \hat\theta)^2\hat{\jmath}.$$

Exponentiate to get the approximate likelihood:

$$L(\theta) \approx m(y)\exp[-\frac{1}{2}(\theta - \hat\theta)^2\hat{\jmath}],$$

where $m(y) = \exp[l(\hat\theta)]$.

Interpretation: the pair $(\hat\theta, \hat{\jmath})$ is an approximately sufficient statistic for θ.



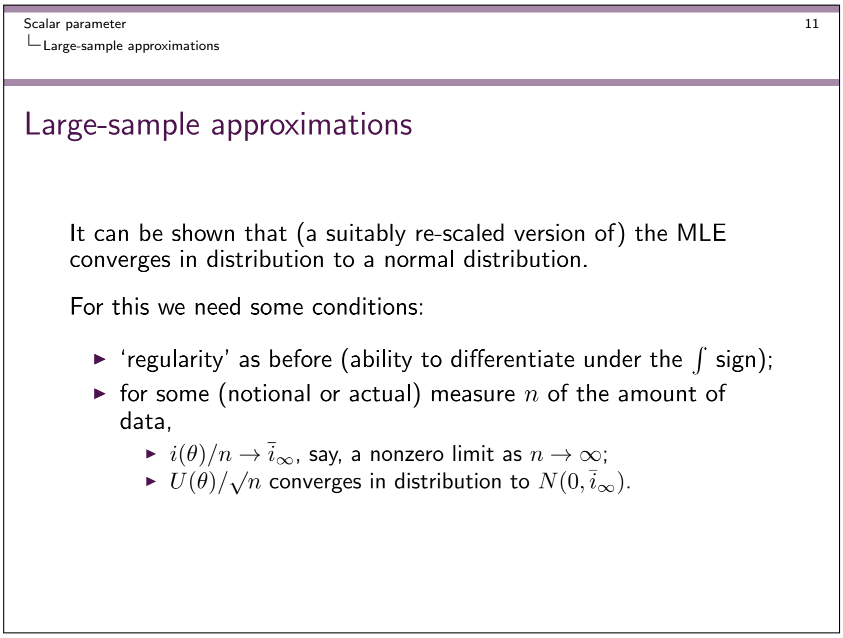
Re-parameterization

Suppose we change from θ to $\phi(\theta)$ (a smooth 1-1 transformation). This is just a change of the model's coordinate system.

Then:

- $\hat{\phi} = \phi(\hat{\theta})$ — the MLE is unaffected;
- $U^{\Phi}\{\phi(\theta); Y\} = U^{\Theta}(\theta; Y)\frac{d\theta}{d\phi}$ (by the chain rule);
- $i^{\Phi}\{\phi(\theta)\} = i^{\Theta}(\theta)\left(\frac{d\theta}{d\phi}\right)^2$ [since $i = \mathrm{var}(U)$]

The units of information change with the units of the parameter.

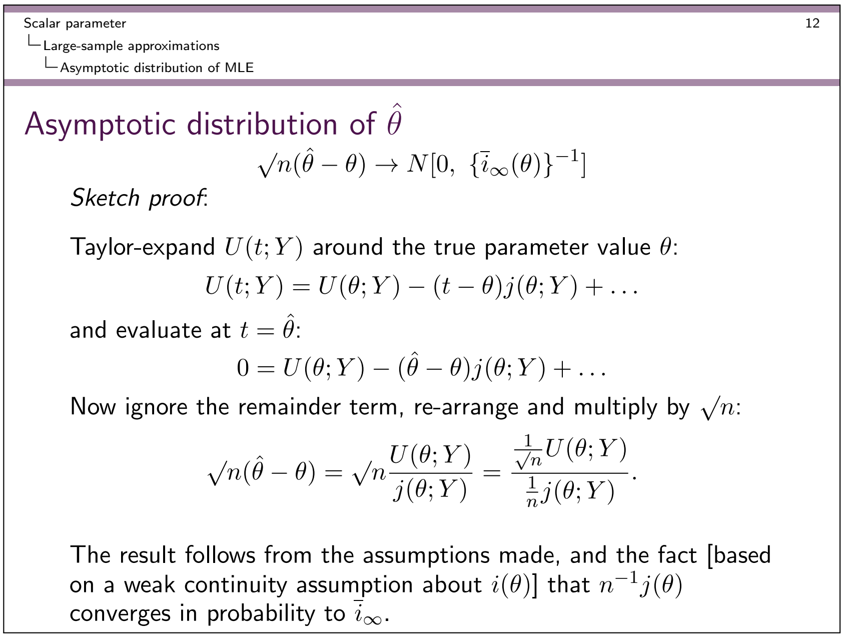


Large-sample approximations

It can be shown that (a suitably re-scaled version of) the MLE converges in distribution to a normal distribution.

For this we need some conditions:

- 'regularity' as before (ability to differentiate under the $\int$ sign);
- for some (notional or actual) measure n of the amount of data,
  - $i(\theta)/n \to \bar{i}_\infty$, say, a nonzero limit as $n \to \infty$;
  - $U(\theta)/\sqrt{n}$ converges in distribution to $N(0, \bar{i}_\infty)$.



Asymptotic distribution of $\hat{\theta}$

$$\sqrt{n}(\hat{\theta} - \theta) \to N[0,\ \{\bar{i}_\infty(\theta)\}^{-1}]$$

Sketch proof:

Taylor-expand $U(t; Y)$ around the true parameter value θ:

$$U(t; Y) = U(\theta; Y) - (t - \theta)j(\theta; Y) + \dots$$

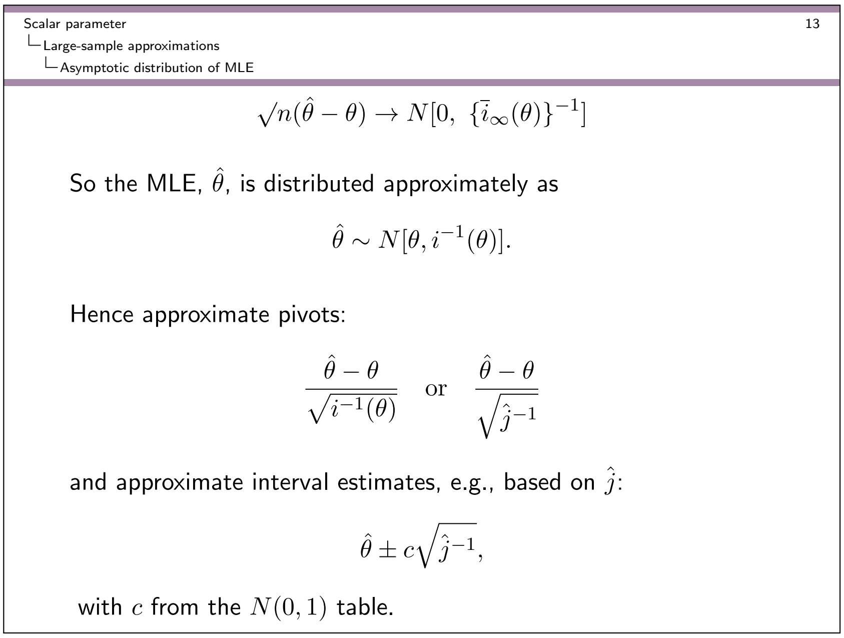
and evaluate at $t = \hat{\theta}$:

$$0 = U(\theta; Y) - (\hat{\theta} - \theta)j(\theta; Y) + \dots$$

Now ignore the remainder term, re-arrange and multiply by $\sqrt{n}$:

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\frac{U(\theta; Y)}{j(\theta; Y)} = \frac{\frac{1}{\sqrt{n}}U(\theta; Y)}{\frac{1}{n}j(\theta; Y)}.$$

The result follows from the assumptions made, and the fact [based on a weak continuity assumption about $i(\theta)$] that $n^{-1}j(\theta)$ converges in probability to $\bar{i}_\infty$.



$$\sqrt{n}(\hat{\theta}-\theta) \to N[0,\ \{\bar{i}_\infty(\theta)\}^{-1}]$$

So the MLE, $\hat{\theta}$, is distributed approximately as

$$\hat{\theta} \sim N[\theta, i^{-1}(\theta)].$$

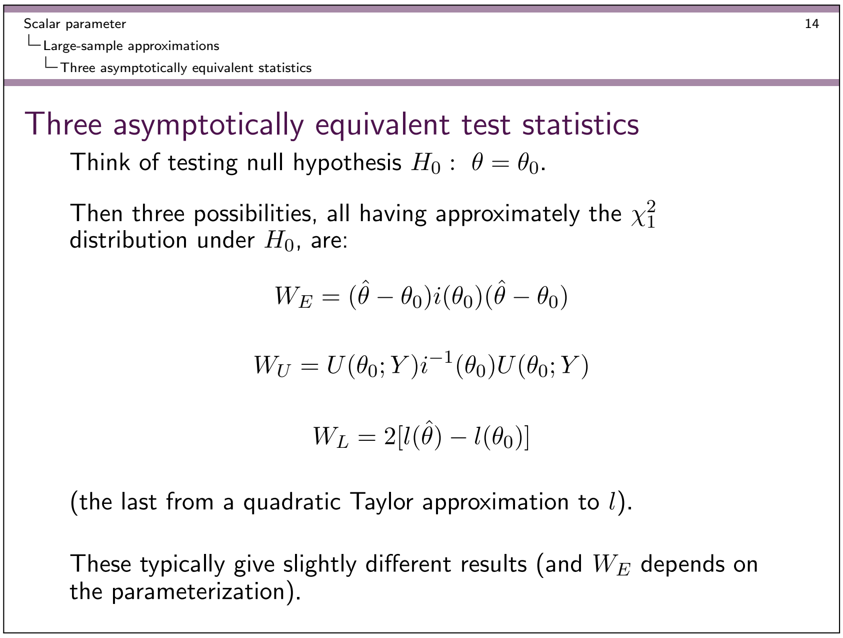
Hence approximate pivots:

$$\frac{\hat{\theta}-\theta}{\sqrt{i^{-1}(\theta)}} \quad \text{or} \quad \frac{\hat{\theta}-\theta}{\sqrt{\hat{j}^{-1}}}$$

and approximate interval estimates, e.g., based on $\hat{j}$:

$$\hat{\theta} \pm c\sqrt{\hat{j}^{-1}},$$

with c from the $N(0,1)$ table.



Three asymptotically equivalent test statistics

Think of testing null hypothesis $H_0:\ \theta=\theta_0$.

Then three possibilities, all having approximately the χ^2_1 distribution under H_0, are:

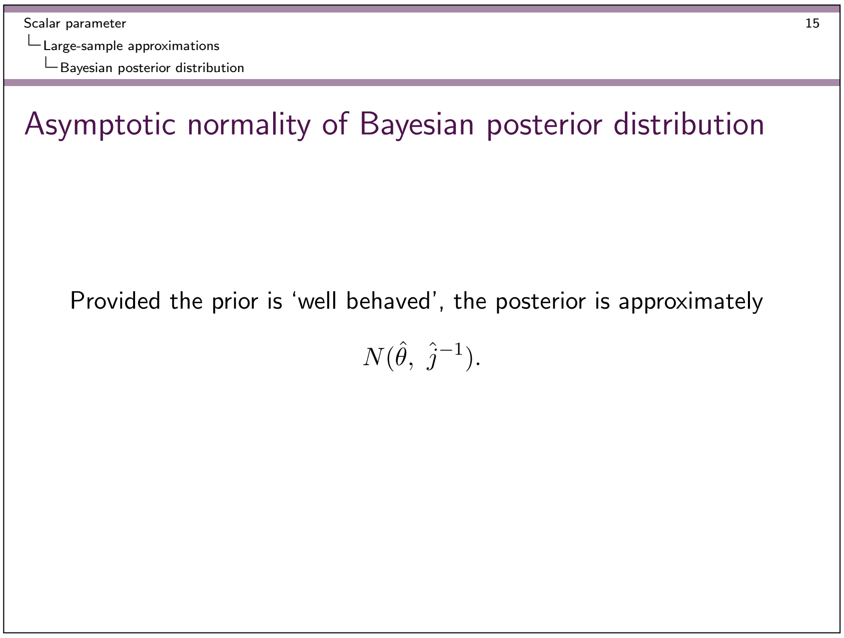
$$W_E = (\hat{\theta}-\theta_0)i(\theta_0)(\hat{\theta}-\theta_0)$$

$$W_U = U(\theta_0;Y)i^{-1}(\theta_0)U(\theta_0;Y)$$

$$W_L = 2[l(\hat{\theta})-l(\theta_0)]$$

(the last from a quadratic Taylor approximation to l).

These typically give slightly different results (and W_E depends on the parameterization).



Asymptotic normality of Bayesian posterior distribution

Provided the prior is 'well behaved', the posterior is approximately

$$N(\hat{\theta},\ \hat{j}^{-1}).$$

Multidimensional parameter θ

All of the above results extend straightforwardly. Score is a *vector*, and information is a *matrix*.

Write

$$U(\theta; Y) = \nabla l(\theta; Y).$$

Then

$$E(U) = 0$$

$$\operatorname{cov}(U) = E(-\nabla\nabla^T l) = i(\theta).$$

The extension of the asymptotic normality argument yields

- a *multivariate* normal approximation for $\hat{\theta}$, with variance-covariance matrix $i^{-1}(\theta)$
- test statistics which straightforwardly extend W_E, W_U and W_L.

The information matrix transforms between parameterizations as

$$i^{\Phi}(\phi) = \left(\frac{\partial\theta}{\partial\phi}\right)^T i^{\Theta}(\theta) \left(\frac{\partial\theta}{\partial\phi}\right)$$

and its inverse transforms as

$$\left[i^{\Phi}(\phi)\right]^{-1} = \left(\frac{\partial\phi}{\partial\theta}\right)^T \left[i^{\Theta}(\theta)\right]^{-1} \left(\frac{\partial\phi}{\partial\theta}\right).$$

Nuisance parameters

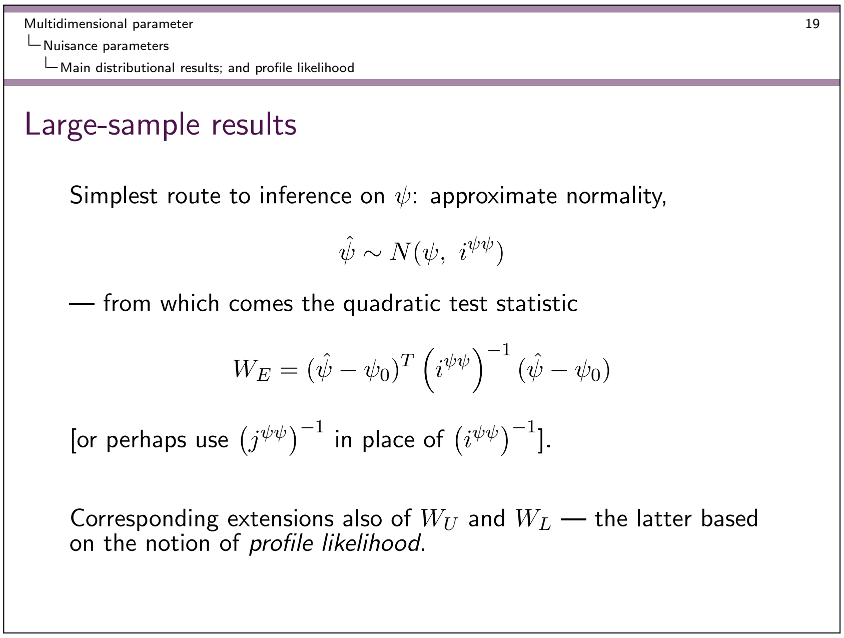
Suppose $\theta = (\psi, \lambda)$, with ψ of interest.

Then partition vector U into (U_ψ, U_λ), and information matrix (and its inverse) correspondingly:

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}$$

$$i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}$$

(and similarly for observed information j)



Large-sample results

Simplest route to inference on ψ: approximate normality,

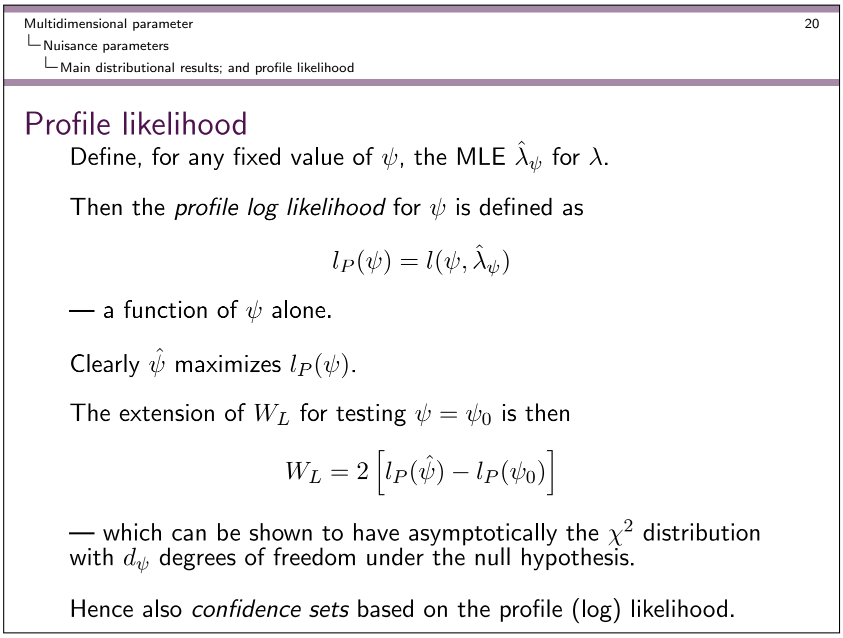
$$\hat{\psi} \sim N(\psi,\ i^{\psi\psi})$$

— from which comes the quadratic test statistic

$$W_E = (\hat{\psi} - \psi_0)^T \left(i^{\psi\psi}\right)^{-1} (\hat{\psi} - \psi_0)$$

[or perhaps use $\left(j^{\psi\psi}\right)^{-1}$ in place of $\left(i^{\psi\psi}\right)^{-1}$].

Corresponding extensions also of W_U and W_L — the latter based on the notion of *profile likelihood*.



Profile likelihood

Define, for any fixed value of ψ, the MLE $\hat{\lambda}_\psi$ for λ.

Then the *profile log likelihood* for ψ is defined as

$$l_P(\psi) = l(\psi, \hat{\lambda}_\psi)$$

— a function of ψ alone.

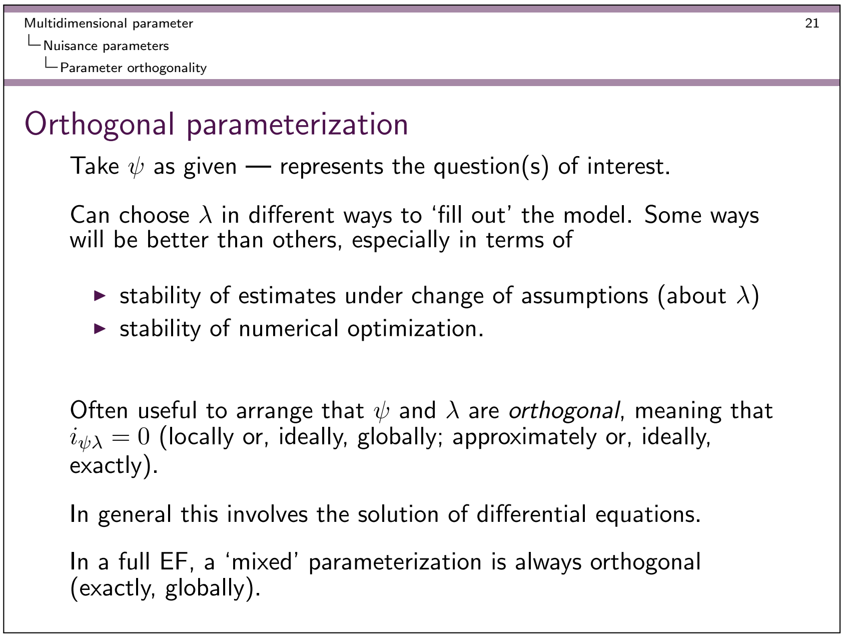
Clearly $\hat{\psi}$ maximizes $l_P(\psi)$.

The extension of W_L for testing $\psi = \psi_0$ is then

$$W_L = 2\left[l_P(\hat{\psi}) - l_P(\psi_0)\right]$$

— which can be shown to have asymptotically the χ^2 distribution with d_ψ degrees of freedom under the null hypothesis.

Hence also *confidence sets* based on the profile (log) likelihood.



Orthogonal parameterization

Take ψ as given — represents the question(s) of interest.

Can choose λ in different ways to ‘fill out’ the model. Some ways will be better than others, especially in terms of

- stability of estimates under change of assumptions (about λ)
- stability of numerical optimization.

Often useful to arrange that ψ and λ are *orthogonal*, meaning that $i_{\psi\lambda} = 0$ (locally or, ideally, globally; approximately or, ideally, exactly).

In general this involves the solution of differential equations.

In a full EF, a ‘mixed’ parameterization is always orthogonal (exactly, globally).

Information in a full EF

Information on the canonical parameters does not depend on Y:

$$i(\phi) = j(\phi) = \nabla\nabla^T k(\phi).$$

So in a full EF model it does not matter whether we use *observed* or *expected* information for inference on ϕ: they are the same.

Full EF: Orthogonality of mixed parameterization

If $\phi = (\phi_1, \phi_2)$ and the parameter (possibly vector) of interest is $\psi = \phi_1$, then choosing

$$\lambda = \eta_2 = E[s_2(Y)]$$

makes the interest and nuisance parameters (ϕ_1, η_2) orthogonal.

This follows straight from the transformation rule, for re-parameterization $(\phi_1, \phi_2) \to (\phi_1, \eta_2)$.

Example: The model $Y \sim N(\mu, \sigma^2)$ is a full 2-parameter EF, with $\phi_1 = 1/(2\sigma^2), \phi_2 = -\mu/\sigma^2$ and $(s_1, s_2) = (y^2, y)$. Hence $\mu = E[s_2(Y)]$ is orthogonal to ϕ_1 (and thus orthogonal to σ^2).

Exercise

Let $Y_1, \ldots, Y_n$ have independent Poisson distributions with mean μ. Obtain the maximum likelihood estimate of μ and its variance

(a) from first principles

(b) by the general results of asymptotic theory.

Suppose now that it is observed only whether each observation is zero or non-zero.

- What now are the maximum likelihood estimate of μ and its asymptotic variance?
- At what value of μ is the ratio of the latter to the former variance minimized?
- In what practical context might these results be relevant?

Part 4
Some non-standard situations

There are various situations in which the standard asymptotic theory (frequentist or Bayesian) has problems. Some of these are mentioned here.

1 Non-regular model

Failure of 'differentiation under the integral sign'.

2 Large number of nuisance parameters

The binary matched pairs example from Part 2 is a fairly extreme example.

Another simple case to study is the 'Neyman-Scott problem', with pairs of normally distributed measurements.

3 Parameter on the boundary of parameter space

Often encountered in mixture models, for example.

4 Multi-modal likelihood

Often hard to diagnose, especially in multi-parameter problems. Bot a potentially serious problem when it occurs.

5 Monotone likelihood

Can happen even in full exponential family models: a prime example is 'complete separation' in logistic regression.

6 Wrong model

MLE minimizes Kullback-Leibler discrepancy; 'sandwich' variance-covariance matrix.

Part 5
Frequentist and Bayesian approaches: some similarities and differences

1 Similarities

2 Differences

3 Hypothesis testing and confidence/credible sets

4 Prior specification

4.1 Subjective (or personal) degree of belief

4.2 'Non-informative' priors

Part 6
Estimating equations

Non-likelihood inference

Sometimes inference based on likelihood is not possible (e.g., for computational reasons, or because a full probability model cannot be specified).

Sometimes inference based on likelihood may be regarded as not desirable (e.g., worries about impact of failure of tentative ‘secondary’ assumptions).

Various non-likelihood approaches, including

- ‘pseudo likelihoods’ — typically designed either for computational simplicity or robustness to failure of (some) assumptions
- ‘estimating equations’ approaches (includes ‘quasi likelihood’)

Estimating equations

Consider scalar θ.

Define estimator θ^* as solution to

$$g(\theta^*; Y) = 0$$

— an *estimating equation*, with the ‘estimating function’ g chosen to that the equation is *unbiased*:

$$E[g(\theta; Y); \theta] = 0$$

for all possible values of θ. (cf. score equation for MLE)

Unbiasedness of the estimating equation results (subject to limiting conditions) in a consistent estimator θ^*.

Examples

Two extremes:

1. Model is fully parametric, $Y \sim f_Y(y;\theta)$. Then the choice $g(\theta;Y) = U(\theta;Y)$ results in an unbiased estimating equation. There may be many others (e.g., based on moments).

2. Model is 'semi-parametric' perhaps specified in terms of some moments. For example, the specification

$$E(Y) = m(\theta)$$

for some given function m may be all that is available, or all that is regarded as reliable: in particular, the full distribution of Y is not determined by θ.

In this case, with Y a scalar rv, the equation

$$g(\theta;Y) = Y - m(\theta) = 0$$

is (essentially) the only unbiased estimating equation available.

Properties

Assume 'standard' limiting conditions. (as for MLE)

Then a similar asymptotic argument to the one used for the MLE yields the large-sample normal approximation

$$\theta^* \sim N\left(\theta,\ \frac{E(g^2)}{[E(g')]^2}\right).$$

Note that the asymptotic variance is invariant to trivial scaling $g(\theta;Y) \to ag(\theta;Y)$ for constant a — as it should be, since θ^* is invariant.

Lower bound on achievable variance

(Godambe, 1960)

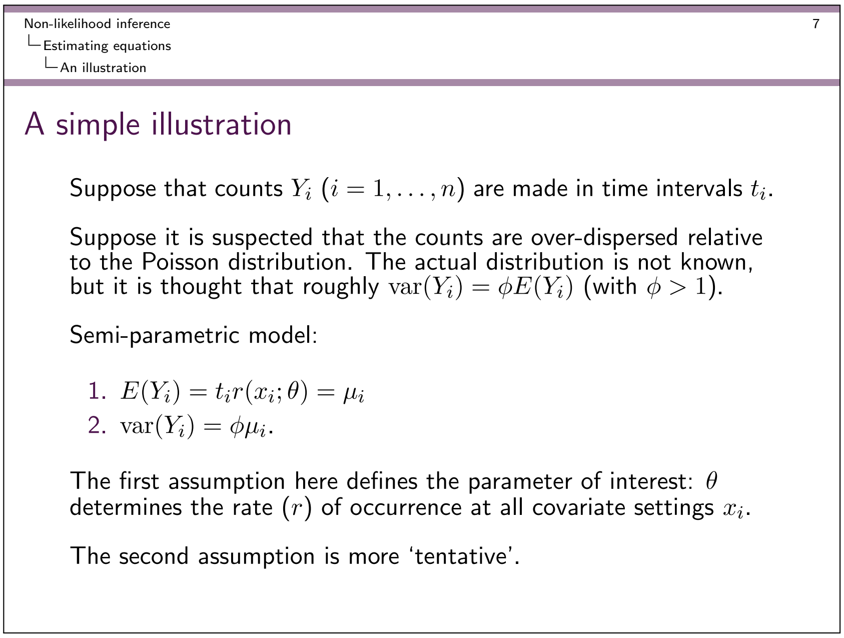
For unbiased estimating equation $g = 0$,

$$\frac{E(g^2)}{[E(g')]^2} \geq \frac{1}{E(U^2)} = i^{-1}(\theta),$$

where $U = \partial \log f / \partial\theta$.

Equality if $g = U$.

This comes from the Cauchy-Schwarz inequality; it generalizes the Cramér-Rao lower bound for the variance of an unbiased estimator.



A simple illustration

Suppose that counts Y_i $(i = 1, \dots, n)$ are made in time intervals t_i.

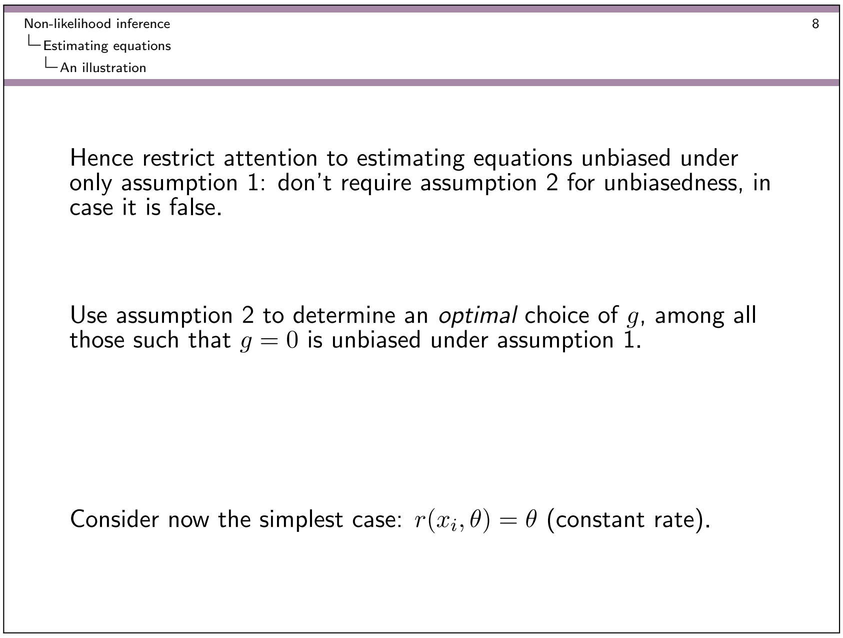
Suppose it is suspected that the counts are over-dispersed relative to the Poisson distribution. The actual distribution is not known, but it is thought that roughly $\operatorname{var}(Y_i) = \phi E(Y_i)$ (with $\phi > 1$).

Semi-parametric model:

1. $E(Y_i) = t_i r(x_i; \theta) = \mu_i$
2. $\operatorname{var}(Y_i) = \phi \mu_i$.

The first assumption here defines the parameter of interest: θ determines the rate (r) of occurrence at all covariate settings x_i.

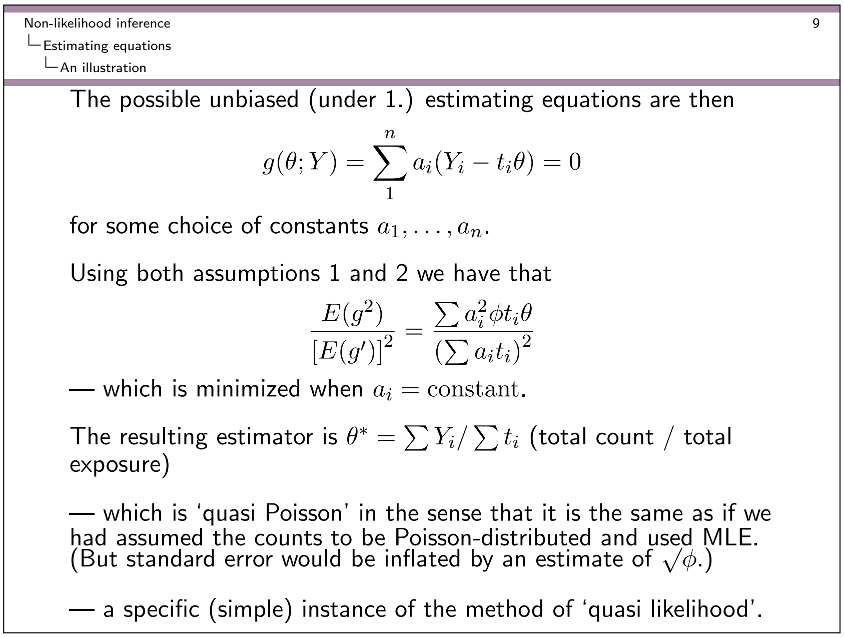
The second assumption is more 'tentative'.



Hence restrict attention to estimating equations unbiased under only assumption 1: don't require assumption 2 for unbiasedness, in case it is false.

Use assumption 2 to determine an *optimal* choice of g, among all those such that $g = 0$ is unbiased under assumption 1.

Consider now the simplest case: $r(x_i, \theta) = \theta$ (constant rate).



The possible unbiased (under 1.) estimating equations are then

$$g(\theta; Y) = \sum_1^n a_i (Y_i - t_i \theta) = 0$$

for some choice of constants $a_1, \dots, a_n$.

Using both assumptions 1 and 2 we have that

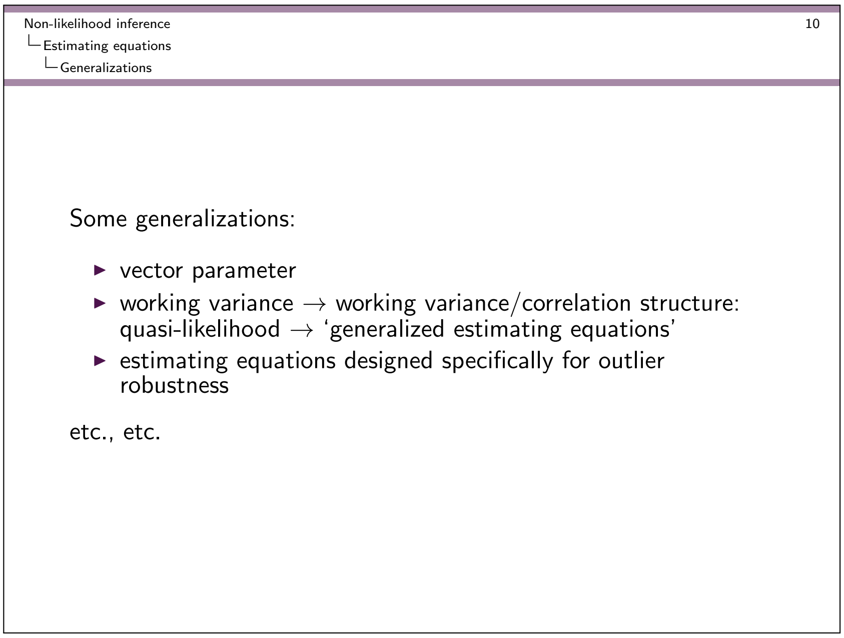
$$\frac{E(g^2)}{[E(g')]^2} = \frac{\sum a_i^2 \phi t_i \theta}{\left(\sum a_i t_i\right)^2}$$

— which is minimized when $a_i = \text{constant}$.

The resulting estimator is $\theta^* = \sum Y_i / \sum t_i$ (total count / total exposure)

— which is 'quasi Poisson' in the sense that it is the same as if we had assumed the counts to be Poisson-distributed and used MLE. (But standard error would be inflated by an estimate of $\sqrt{\phi}$.)

— a specific (simple) instance of the method of 'quasi likelihood'.



Non-likelihood inference
└ Estimating equations
└ Generalizations

10

Some generalizations:

- vector parameter
- working variance → working variance/correlation structure: quasi-likelihood → ‘generalized estimating equations’
- estimating equations designed specifically for outlier robustness

etc., etc.

APTS module *Statistical Inference*

D. Firth, January 2011

Some suggested reading/reference material

Books

Cox, D. R. (2006). *Principles of Statistical Inference.* CUP.
Closest book to the APTS lectures.

Cox, D. R and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall.
An older and more detailed account of similar material.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference.* CUP.
A broad, concise introduction, including some decision analysis.

Lehmann, E. L. and Romano, J. P. (2004, 3rd ed). *Testing Statistical Hypotheses.* Springer.
Lehmann, E. L. and Casella, G. C. (2001, 2nd ed). *Theory of Point Estimation.* Springer.
Much detailed mathematical material on estimation and testing.

Jeffreys, H. (1961, 3rd ed.). *Theory of Probability.* OUP.
A detailed development of objective Bayesian theory.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley.
Pioneering account of the personalistic Bayesian view.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Inference.* Arnold.
Thorough treatment of current Bayesian approaches.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* Wiley.
A good discussion of the principles and methods of decision analysis.

Edwards, A. W. F. (1972). *Likelihood.* CUP.
Statistical theory based solely on likelihood.

Davison, A. C. (2003). *Statistical Models.* CUP.
Especially chapters 3, 4, 7, 11, 12. Recommended book also for APTS module *Statistical Modelling.*

McCullagh, P. and Nelder, J. A. (1989, 2nd ed.). *Generalized Linear Models.* Chapman and Hall.
An authoritative account of the generalized linear model (links with the APTS module *Statistical Modelling*).

Papers

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics* **6**, 17-24.
Conditional test of lack of fit: as in part 2.

Cox, D. R (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-372.
A discussion of conditioning and the relations between various approaches.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Stat. Soc.* B **34**, 1-41.
Influential paper on hierarchical Bayesian analysis.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Stat. Soc.* B **49**, 1-39.
Orthogonality (as in part 3).

McCullagh, P. (1991). Quasi-likelihood and estimating functions. Pages 265-286 in: Hinkley DV, Reid N and Snell EJ (eds.), *Statistical Theory and Modelling, in Honour of Sir David Cox FRS.* Chapman and Hall.
Review of the theory of estimating equations.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5-42.

APTS module *Statistical Inference*

January 2011

Assessment material

The work provided here is intended to take students up to half a week to complete. Students should talk to their supervisors to find out whether their department requires this work as part of any formal accreditation process (APTS itself has no resources to assess or certify students). It is anticipated that departments will decide on the appropriate *level* of assessment locally, and may choose to drop some (or indeed all) of the items, accordingly.

Students should also take their supervisor's advice on how much total time to devote to these problems. A reasonable target would be to do at least two of the questions from Section A and one from Section B.

Section A: consolidation of the APTS-week material

1. [From Part 2. If you get stuck, any good book on the analysis of binary/categorical data should have some discussion of this, e.g., Cox & Snell (1989) *Analysis of Binary Data*.] For the binary matched pairs model, derive the conditional binomial distribution for inference on the common log odds ratio ψ. Discuss whether it is reasonable to discard all the data from 'non-mixed' pairs.

2. [From Part 3.] Let $Y_1, \dots, Y_n$ have independent Poisson distributions with mean μ. Obtain the maximum likelihood estimator of μ. Obtain that estimator's variance,

   (a) from first principles;

   (b) by the general results of asymptotic theory.

   Suppose now that it is observed only whether each observation is zero or non-zero.

   (c) What now are the maximum likelihood estimate of μ and its asymptotic variance?

   (d) At what value of μ is the ratio of the latter to the former variance minimized?

   (e) In what practical context might these results be relevant?

3. [From Part 3.] Suppose that $Y_1, \dots, Y_n$ are independent, with $Y_i \sim N(\lambda + \psi x_i, \sigma^2)$ and σ^2 known.

   (a) Calculate the expected information matrix $i(\psi, \lambda)$, and relate this to what you know about least squares.

   (b) Find a new parameterization, (ψ, τ) say, in which τ is orthogonal to ψ. What are the advantages of orthogonality?

Section B: extension of the APTS-week material

1. [Extends Part 4.] Let $Y_1, \dots, Y_n$ be independently binomially distributed each corresponding to ν trials with probability of success θ. Both ν and θ are unknown. Construct simple (inefficient) estimates of the parameters, for example by considering

   - the mean and variance of the sample
   - or the proportions of values equal to zero and one

   On the basis of one or both of these preliminary estimates for what combinations of ν, θ would you expect the maximum likelihood estimate of ν to be at infinity with appreciable probability? Simulate one of these situations and either

   - examine the shape of the likelihood surface for say 10 simulation runs

   or (more advanced)

   - study how the proportion of formally infinite estimates depends on the underlying parameters
   - or study the properties of profile-likelihood based confidence limits in such a situation
   - or set up a Bayesian formulation.

   The situation described is a simplified version of a model for the estimation of the number of bugs in a complex piece of computer software. (e.g., Joe and Reid, 1985, *JASA* **80**, 222–226)

2. [Extends Part 5.] Write a short summary (2 pages or so) of the uses and limitations of *empirical Bayes* methods. [Not covered in this week's APTS lectures; see for example the books by Cox (2006; sec 5.12), Davison (2003; sec 11.5) or O'Hagan and Forster (2004; sec 5.25–5.27) for some discussion and further references.]