2 Some Basic Concepts

2.1 Likelihood

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

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

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

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

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

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

  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 $\ldots$;
  - 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.

Overview

- In theoretical discussion we glibly write something like

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

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

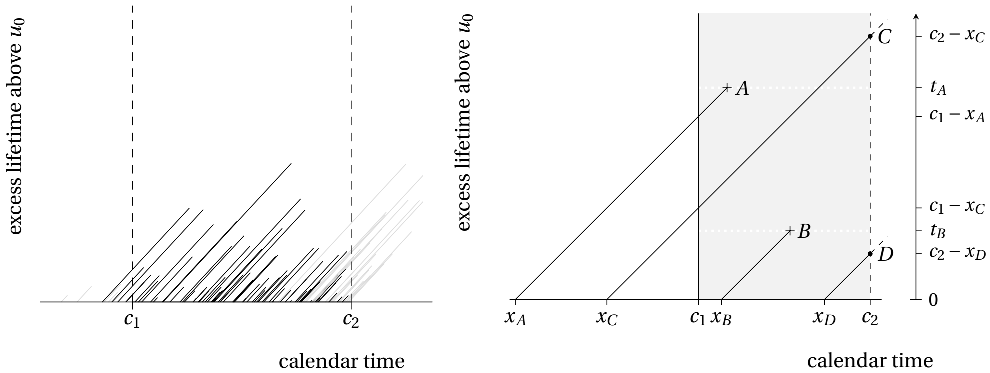
$$\prod_{j=1}^{n} f(y_j \mid y_j \in \mathcal{I}_j; \theta).$$

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*

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

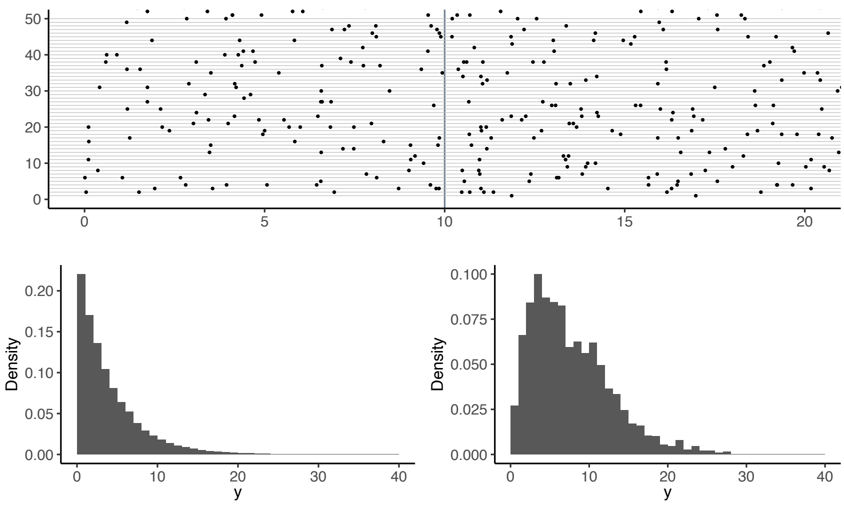
where x is the calendar date at which they reach age u_0. See the next page.

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 $(T = \min(Y, b), D = I(Y \le b))$ for some b;
  - **left-censoring** results in $(T = \max(Y, a), D = I(Y > a))$ for some a;
  - **interval-censoring** results in $(Y, I(a < Y \le b))$, $(a, I(Y \le a))$ 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 $Y_0 \sim \mathrm{Poiss}(\theta)$ and $Y_0,\ldots,Y_n$ are such that $Y_{j+1} \mid Y_0 = y_0,\ldots,Y_j = y_j \sim \mathrm{Poiss}(\theta y_j)$.*

Note to Example 29

Here

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

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

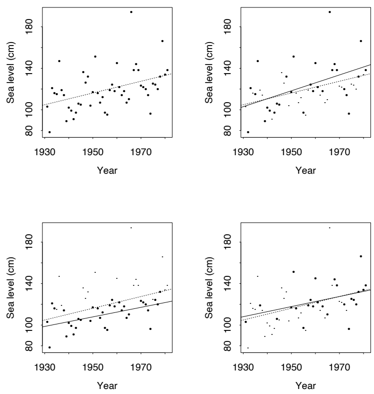
$$\mathrm{P}(I = 1 \mid x, y) f(y \mid x; \theta) f(x; \theta), \qquad \int \mathrm{P}(I = 0 \mid x, y) f(y \mid x; \theta) f(x; \theta)\,\mathrm{d}y,$$

  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 $\mathrm{P}(M \le m) = (m/\theta)^n$ 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. Such a 'smallest sufficient statistic' 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 $s(z) = s(z)$, 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 $\mathrm{P}(U \le u) = 1 - (1 - u/\theta)^n$, 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 $\mathrm{var}(T) \le \mathrm{var}(\tilde\theta)$.*

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

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

so the Jacobian for the transformation is

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

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

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

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**

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

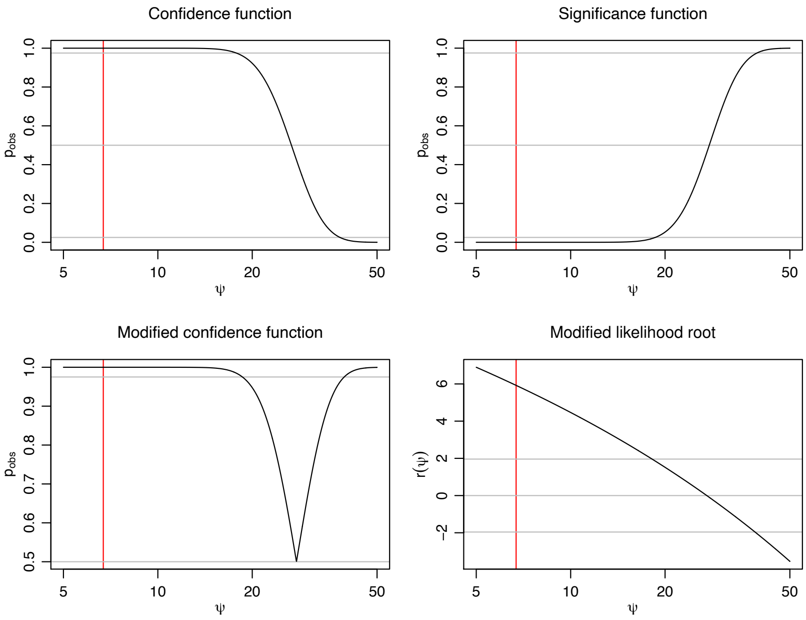
- As $F_0(T) \sim U(0,1)$ when $\psi = \psi_0$, we regard values of ψ for which $p(\psi)$ is too extreme as incompatible with t^o, leading to the (two-sided) $(1-\alpha)$ confidence set

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

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 $Y_1, \ldots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\psi, 1)$. 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 $\mathrm{P}(M \leq x) = (x/\theta)^n$, 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

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

  and we can vary ψ to (numerically) solve

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

  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**

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

  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.