Likelihood

Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch

October 7, 2025

- 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 $n^{-1}\ell(\theta)$.

- the formula for $\ell(\theta)$ is readily extended — for example, if $y_1, \ldots, y_n$ are in time order, then

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

- 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**

$$\nabla \ell(\theta) = \left. \frac{\partial \ell(\theta)}{\partial \theta} \right|_{\theta=\widehat{\theta}} = 0,$$

  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.

- 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.

- 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 (Coin Tossing)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(\theta)$, and $T(\boldsymbol{Y}) = \sum_{i=1}^n Y_i$. For $\boldsymbol{y} \in \{0,1\}^n$,

$$\begin{aligned}
\mathbb{P}[\boldsymbol{Y} = \boldsymbol{y} | T = t] &= \frac{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}, T=t]}{\mathbb{P}[T=t]} = \frac{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}[T=t]} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\} \\
&= \frac{\theta^{\Sigma_{i=1}^n y_i}(1-\theta)^{n-\Sigma_{i=1}^n y_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\} \\
&= \binom{n}{t}^{-1} \mathbf{1}\{\Sigma_{i=1}^n y_i = t\}.
\end{aligned}$$

- T is sufficient for $\theta \rightarrow$ Given # of tosses that came heads, knowing *which tosses* came heads is irrelevant in deciding the probability of heads:

0 0 1 1 1 0 1 VS 1 0 0 0 1 1 1 VS 1 0 1 0 1 0 1

If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(\theta)$, find a sufficient statistic for θ. 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 $M = \max(y_1, \ldots, y_n)$ (since $\prod_j I(0 < y_j < \theta) = I(0 < M < \theta)$). 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 know that $\Pr(M \leq m) = (m/\theta)^n$ 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.

- Definition of sufficient statistics is hard to verify (especially for continuous variables)
- Definition does not allow easy identification of sufficient statistics

Theorem (Fisher-Neyman Factorization Theorem)

Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ *has a joint density or frequency function* $f(\boldsymbol{y}; \theta)$, $\theta \in \Theta$. *A statistic* $T = T(\boldsymbol{Y})$ *is sufficient for* θ *if and only if*

$$f(\boldsymbol{y}; \theta) = g(T(\boldsymbol{y}), \theta) h(\boldsymbol{y}).$$

Example

Let $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$ with pdf $f(y; \theta) = \mathbf{1}\{y \in [0, \theta]\}/\theta$. Then,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\theta^n} \mathbf{1}\{\boldsymbol{y} \in [0, \theta]^n\} = \frac{\mathbf{1}\{\max[y_1, \ldots, y_n] \leq \theta\} \mathbf{1}\{\min[y_1, \ldots, y_n] \geq 0\}}{\theta^n}$$

Therefore $T(\boldsymbol{Y}) = Y_{(n)} = \max[Y_1, ..., Y_n]$ is sufficient for θ.

Proof of Neyman-Fisher Theorem - Discrete Real Statistic.

Suppose first that T is sufficient. Then

$$\begin{aligned} f(y;\theta) &= \mathbb{P}_\theta[\boldsymbol{Y}=\boldsymbol{y}] = \sum_t \mathbb{P}_\theta[\boldsymbol{Y}=\boldsymbol{y}, T=t] \\ &= \mathbb{P}_\theta[\boldsymbol{Y}=\boldsymbol{y}, T=T(\boldsymbol{y})] = \mathbb{P}_\theta[T=T(\boldsymbol{y})]\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|T=T(\boldsymbol{y})] \end{aligned}$$

Since T is sufficient, $\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|T=T(\boldsymbol{y})]$ is independent of θ and so $f(y;\theta) = g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})$. Now suppose that $f(y;\theta) = g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})$. Then if $T(\boldsymbol{y}) = t$,

$$\begin{aligned} \mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|T=t] &= \frac{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y},T=t]}{\mathbb{P}[T=t]} = \frac{\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}]}{\mathbb{P}[T=t]}\mathbf{1}\{T(\boldsymbol{y})=t\} \\ &= \frac{g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})\mathbf{1}\{T(\boldsymbol{y})=t\}}{\sum_{\boldsymbol{z}:T(\boldsymbol{z})=t} g(T(\boldsymbol{z});\theta)h(\boldsymbol{z})} = \frac{h(\boldsymbol{y})\mathbf{1}\{T(\boldsymbol{y})=t\}}{\sum_{T(\boldsymbol{z})=t} h(\boldsymbol{z})}. \end{aligned}$$

which does not depend on θ. □

Example (Sufficient statistics for i.i.d. normal samples)

Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Recall that we can write

$$f(y; \mu, \sigma^2) = \frac{e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} = \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}$$

and so

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} y_i - \frac{n}{2}\log(2\pi\sigma^2) - \frac{n\mu^2}{2\sigma^2}\right\}.$$

Consequently, Fisher-Neyman factorisation implies that the statistic

$$S(\boldsymbol{Y}) = \left(S_1(\boldsymbol{Y}), S_2(\boldsymbol{Y})\right)^\top = \left(\sum_{i=1}^{n} Y_i\,, \sum_{i=1}^{n} Y_i^2\right)^\top = \left(\overline{Y}\,, \sum_{i=1}^{n} Y_i^2\right)^\top$$

is sufficient for the parameter (μ, σ^2) and so is the statistic

$$T(\boldsymbol{Y}) = \left(T_1(\boldsymbol{Y}), T_2(\boldsymbol{Y})\right)^\top = \left(n^{-1}\sum_{i=1}^{n} Y_i\,,\, n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)^\top$$

since T and S are 1-1 functions of each other.

- 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.

Lemma

If T and S are minimally sufficient statistics for a parameter θ, then there exists injective functions g and h such that $S = g(T)$ and $T = h(S)$.

Theorem

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ have joint density or frequency function $f(\boldsymbol{y}; \theta)$ and $T = T(\boldsymbol{Y})$ be a statistic. If $f(\boldsymbol{y}; \theta)/f(\boldsymbol{z}; \theta) \perp \theta \iff T(\boldsymbol{y}) = T(\boldsymbol{z})$. Then T is minimally sufficient for θ.

Assume for simplicity that $f(\boldsymbol{y}; \theta) > 0$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and $\theta \in \Theta$. Let $\mathcal{T} = \{T(\boldsymbol{u}) : u \in \mathbb{R}^n\}$ be the image of $\mathbb{R}^n$ under T and let A_t be the level sets of T. For each t, choose a representative element $\boldsymbol{w}_t \in A_t$. Notice that for any $\boldsymbol{y}$, $\boldsymbol{w}_{T(\boldsymbol{y})}$ is in the same level set as $\boldsymbol{y}$, so that

$$f(\boldsymbol{y}; \theta)/f(\boldsymbol{w}_{T(\boldsymbol{y})}; \theta)$$

does not depend on θ by assumption. Let $g(t, \theta) := f(\boldsymbol{w}_t; \theta)$ and notice

$$f(\boldsymbol{y}; \theta) = \frac{f(\boldsymbol{w}_{T(\boldsymbol{y})}; \theta) f(\boldsymbol{y}; \theta)}{f(\boldsymbol{w}_{T(\boldsymbol{y})}; \theta)} = g(T(\boldsymbol{y}), \theta) h(\boldsymbol{y})$$

and sufficiency follows from the Fisher-Neyman factorization theorem.
[minimality part] Suppose that T' is another sufficient statistic. By the factorization thm: $\exists g', h' : \quad f(\boldsymbol{y}; \theta) = g'(T'(\boldsymbol{y}); \theta) h'(\boldsymbol{y})$. Let $\boldsymbol{y}, \boldsymbol{z}$ be such that $T'(\boldsymbol{y}) = T'(\boldsymbol{z})$. Then

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

Since ratio does not depend on θ, we have by assumption $T'(\boldsymbol{y}) = T'(\boldsymbol{z})$. Hence T is a function of T'; so is minimal by arbitrary choice of T'.

Example (Bernoulli Trials)

Let $Y_1, ..., Y_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Let $\boldsymbol{z}, \boldsymbol{y} \in \{0,1\}^n$ be two possible outcomes. Then

$$\frac{f(\boldsymbol{z};\theta)}{f(\boldsymbol{y};\theta)} = \frac{\theta^{\Sigma z_i}(1-\theta)^{n-\Sigma z_i}}{\theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}}$$

which is constant if and only if $T(\boldsymbol{z}) = \sum z_i = \sum y_i = T(\boldsymbol{y})$, so that T is minimally sufficient.

$$f(y) = \exp\left\{\sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}$$

where:

1. $\phi = (\phi_1, ..., \phi_k)$ is a k-dimensional parameter in $\Phi \subseteq \mathbb{R}^k$;
2. $T_i : \mathcal{Y} \to \mathbb{R}$, $i = 1, ..., k$, $S : \mathcal{Y} \to \mathbb{R}$, and $\gamma : \mathbb{R}^k \to \mathbb{R}$, are real-valued;
3. The support $\mathcal{Y}$ of f does not depend on ϕ.

"Natural" is from the mathematics point of view – usual parameter $\theta = \eta^{-1}(\phi)$ often different.

Natural vs Usual Parametrization

$$\exp\left\{\sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\phi) + S(y)\right\} = \exp\left\{\sum_{i=1}^{k} \eta_i(\theta) T_i(y) - d(\theta) + S(y)\right\}.$$

where $\eta : \mathbb{R}^k \to \mathbb{R}^k$ is a C^2 map such that

$$\phi = \eta(\theta)$$

and so $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, for $d = \gamma \circ \eta$.

Example (Minimal sufficiency for k-parameter exponential families)

An i.i.d. sample $(Y_1, ..., Y_n)^\top$ from an exponential family has joint distribution

$$f_{Y_1,...,Y_n}(y_1, \ldots, y_n) = \exp\left\{\sum_{j=1}^{k} \phi_j \tau_j(y_1, \ldots, y_n) - n\gamma(\phi_1, ..., \phi_n) + \sum_{i=1}^{n} S(y_i)\right\}$$

where $\tau_j(y_1, \ldots, y_n) = \sum_{i=1}^{n} T_j(y_i)$. If the $\{T_j\}_{j=1}^{k}$ are non-trivial, the ratio $f(\boldsymbol{y})/f(\boldsymbol{z})$ will be constant with respect to $(\phi_1, \ldots, \phi_k)$ if and only if as $(\phi_1, \ldots, \phi_k)$ varies, the quantity below remains constant.

$$\sum_{j=1}^{k} \phi_j \big(\tau_j(y_1, \ldots, y_n) - \tau_j(z_1, \ldots, z_n)\big)$$

So if $(\phi_1, ..., \phi_k)$ range over an open parameter space of dimension k, this must imply that

$$\tau_j(y_1, \ldots, y_n) = \tau_j(z_1, \ldots, z_n).$$

Conversely, when the latter is true, the density ratio is clearly independent of the parameters, and so the statistic $\tau(\boldsymbol{y}) = (\tau_1(\boldsymbol{y}), ..., \tau_k(\boldsymbol{y}))$ is minimally sufficient for $(\phi_1, ..., \phi_k)$.

Theorem (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 $\mathrm{var}(T) \leq \mathrm{var}(\tilde{\theta})$.

Comments:

- Throwing away irrelevant aspects of the data improves estimation quality.
- These irrelevant aspects contribute to the variation of the estimator (as they have sampling variation of their own), but without furnishing any useful information on the parameter
- $\tilde{\theta}^* = \mathbb{E}[\tilde{\theta}|S]$ is called a "Rao-Blackwellised" version of $\tilde{\theta}$.
- 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).

Proof.

Since S is sufficient for θ, $\mathbb{E}[\tilde{\theta}|S = s] = h(s)$ is independent of θ, so that $\tilde{\theta}^*$ is well-defined as a statistic (depends only on $\boldsymbol{Y}$ and not θ). Then,

$$\mathbb{E}[\tilde{\theta}^*] = \mathbb{E}[\mathbb{E}[\tilde{\theta}|S]] = \mathbb{E}[\tilde{\theta}] = \theta.$$

Furthermore, from the law of total variance, we have

$$\mathrm{var}(\tilde{\theta}) = \mathrm{var}[\mathbb{E}(\tilde{\theta}|S)] + \mathbb{E}[\mathrm{var}(\tilde{\theta}|S)] \geq \mathrm{var}[\mathbb{E}(\tilde{\theta}|S)] = \mathrm{var}(\tilde{\theta}^*)$$

In addition, note that

$$\mathrm{var}(\tilde{\theta}|S) := \mathbb{E}[(\tilde{\theta} - \mathbb{E}[\tilde{\theta}|S])^2|S] = \mathbb{E}[(\tilde{\theta} - \tilde{\theta}^*)^2|S]$$

so that $\mathbb{E}[\mathrm{var}(\tilde{\theta}|S)] = \mathbb{E}(\tilde{\theta} - \tilde{\theta}^*)^2 > 0$ unless if $\mathbb{P}(\tilde{\theta}^* = \tilde{\theta}) = 1$. □

Suppose that $\tilde{\theta}$ is an unbiased estimator of $g(\theta)$ and T, S are θ-sufficient.

- What is the relationship between $\mathrm{var}(\underbrace{\mathbb{E}[\tilde{\theta}|\,T]}_{\tilde{\theta}^*_T}) \overset{?}{\gtrless} \mathrm{var}(\underbrace{\mathbb{E}[\tilde{\theta}|S]}_{\tilde{\theta}^*_S})$
- Intuition suggests that whichever of T, S carries the least irrelevant information (in addition to the relevant information) should "win"
  - ↪ More formally, if $T = h(S)$ then we should expect that $\tilde{\theta}^*_T$ dominate $\tilde{\theta}^*_S$.

Proposition

For $\tilde{\theta}$ an unbiased estimator of θ and T, S two θ-sufficient statistics, define

$$\tilde{\theta}^*_T := \mathbb{E}[\tilde{\theta}|\,T] \quad \& \quad \tilde{\theta}^*_S := \mathbb{E}[\tilde{\theta}|S].$$

Then, the following implication holds

$$T = h(S) \implies \mathrm{var}(\tilde{\theta}^*_T) \leq \mathrm{var}(\tilde{\theta}^*_S)$$

- Essentially this means that the best possible "Rao-Blackwellisation" is achieved by conditioning on a minimally sufficient statistic.

Proof.

Recall the *tower property* of conditional expectation: if $Y = f(X)$, then

$$\mathbb{E}[Z|Y] = \mathbb{E}\{\mathbb{E}(Z|X)|Y\}.$$

Since $T = f(S)$ we have

$$\begin{aligned} \tilde{\theta}^*_T &= \mathbb{E}[\tilde{\theta}|T] \\ &= \mathbb{E}[\mathbb{E}(\tilde{\theta}|S)|T] \\ &= \mathbb{E}[\tilde{\theta}^*_S|T] \end{aligned}$$

The conclusion now follows from the Rao-Blackwell theorem. □

- 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.

Theorem

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.

Example

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

- 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 (by Lehman-Scheffé).

Theorem (Lehman-Scheffé)

For $\{X_i\}_{i=1}^n \sim f_\theta$ for some $\theta \in \Theta$. Let T be a complete sufficient statistic for θ. If $\psi(T)$ is the unbiased statistic based on T, then it is the unique minimum variance unbiased estimator (MVUE) of θ.

Proof.

By Rao-Blackwell, if S is unbiased, we know that $g(T) = \mathrm{E}[S|T]$ is an unbiased estimator of θ with variance no larger than that of S. What remains to be shown is the uniqueness of this estimator. Let W be another possible MVUE of θ. Then, $h(T) = \mathrm{E}[W|T]$ is another unbiased estimator with variance no larger than that of W. Then,

$$\mathrm{E}[g(T) - h(T)] = 0 \ \forall \theta \in \Theta.$$

Since T is complete it must be true that $g(T) - h(T) = 0$ for all $\theta \in \Theta$. Thus, $\psi(T)$ is unique with a variance no large than that of any other unbiased estimator. □

- 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 data whose distribution does not depend on the parameter. Then

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

  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 (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 (i.e., fix $N = n$).

Example (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 $f(y \mid x; \theta) f(x)$, so that the parameters θ 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.

Theorem (Basu)

Let $(P_\theta; \theta \in \Theta)$ be a family of distributions on a measurable space $X, \mathcal{A}$. Let T be a boundedly complete sufficient statistic for θ and suppose A is ancillary to θ. Then, conditional on θ, $T \perp\!\!\!\perp A \mid \theta$.

Proof.

Let F_T and F_A be the marginal distributions of T and A respectively and define, for some set B,

$$p_A = \Pr(A \in B)$$
$$q_A(T(X)) = \Pr(A \in B | T(X)).$$

Then, we see that

$$\mathrm{E}_\theta[q_A(T) - p_A] = p_A - p_A = 0,$$

by the fact that p_A being independent of θ by definition of an ancillary statistic and q_A being independent of θ by sufficiency of T. Furthermore, since T is complete, this implies that $q_A = p_A$ almost surely.

proof continued.

Thus, for any measurable set C,

$$\begin{aligned}\Pr_\theta(A \in B, T \in C) &= \int q_A(t)\mathbf{1}(t \in C)\mathrm{d}F_T(t) \\ &= \int p_A(t)\mathbf{1}(t \in C)\mathrm{d}F_T(t) \\ &= \Pr_\theta(A \in B)\Pr_\theta(T \in C).\end{aligned}$$

□

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

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

Lemma

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

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.

- In theoretical discussion we usually write something like

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

  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.

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

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

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

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

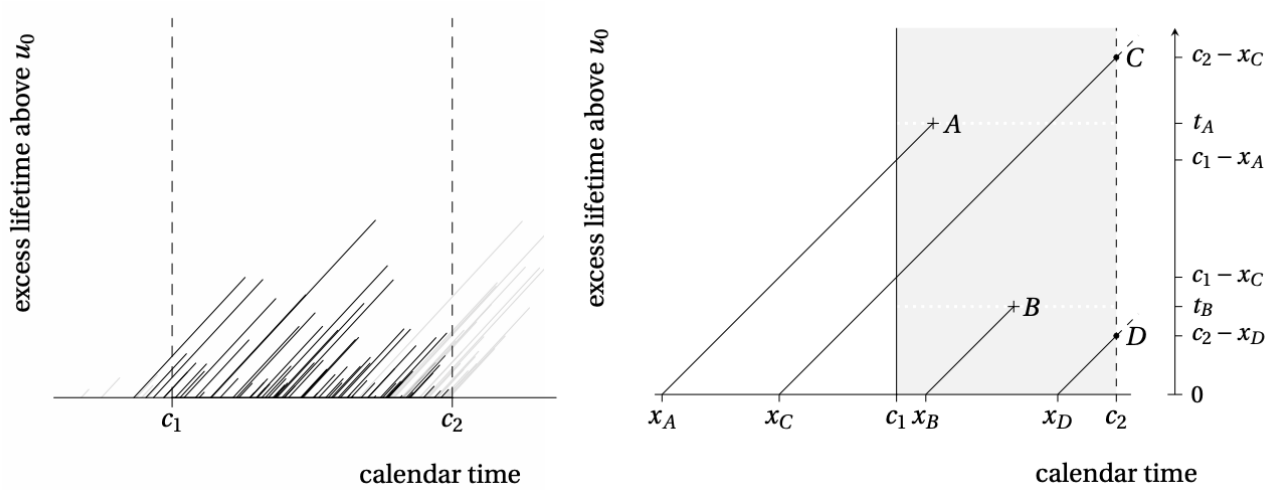
Example

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.

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.

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

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

- 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.

- 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.

- 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 (Poisson birth process)

Find the likelihood when $Y_0 \sim \mathrm{Poiss}(\theta)$ and $Y_0, \ldots, Y_n$ are such that $Y_{j+1} \mid Y_0 = y_0, \ldots, Y_j = y_j \sim \mathrm{Poiss}(\theta y_j)$.

- 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

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

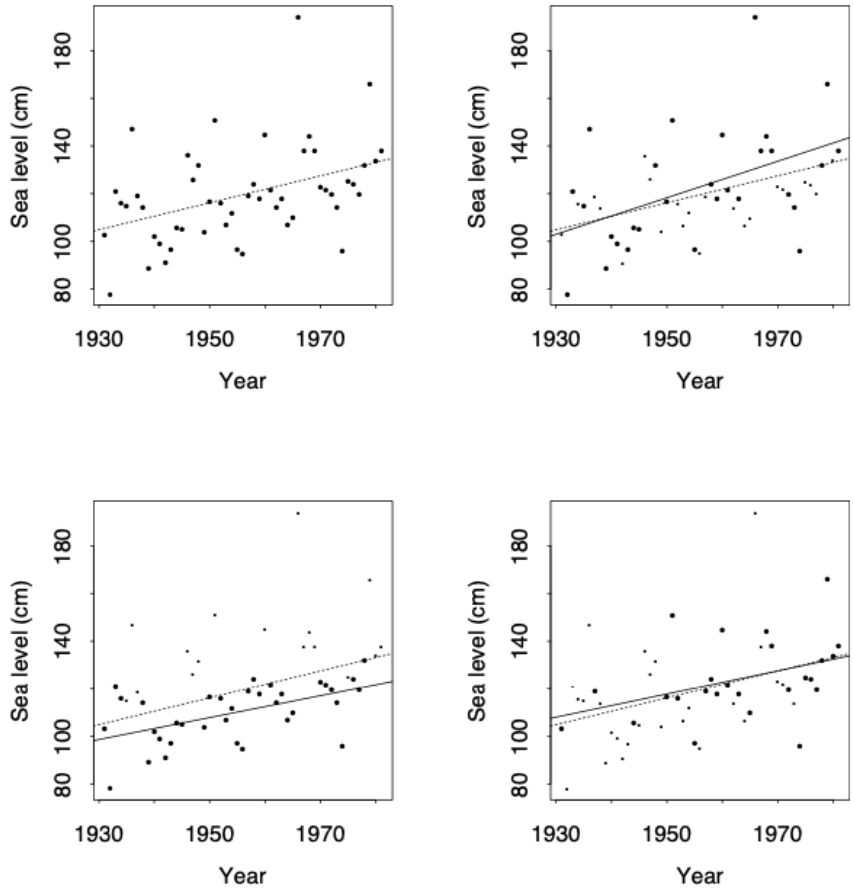
  and there are three possibilities:

  - data are **missing completely at random**, $\Pr(I = 0 \mid x, y) = \Pr(I = 0)$;
  - data are **missing at random**, $\Pr(I = 0 \mid x, y) = \Pr(I = 0 \mid x)$; and
  - **non-ignorable non-response**, $\Pr(I = 0 \mid x, y)$ 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 missing data.

	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) mechanisms of the form

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

  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.

- 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.

- 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 $F(t^{\circ}; \psi_\alpha) = \alpha$ for appropriate values of α.

Why is $F(T;\psi_0)$ uniform?

- Write $F_0(t) = \Pr(T \leq t;\psi_0)$, and note if $T \sim F_0$, then

$$\Pr\{F_0(T) \leq u\} = \Pr\{T \leq F_0^{-1}(u)\} = F_0\{F_0^{-1}(u)\} = u, \quad 0 < u < 1,$$

  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. In such cases $F_0(T)$ can only take a finite or countable number of values that give the **achievable confidence levels**.

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

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

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

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

  for $p(\psi_0)$ as the P-value for a test of $H_0 : \psi = \psi_0$.
- 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.

Example (Normal sample)

Provide a confidence interval for the mean of a normal random sample with known variance.

- 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

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

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

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

- 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.

Example (Uniform sample)

Provide inference for the upper limit of a uniform sample.

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

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

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

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

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

  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.