Nuisance parameters and their treatment

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 2, Hilary Term 2008

April 2, 2008

A statistic $A = a(X)$ is said to be *ancillary* if

(i) The distribution of A does not depend on θ;

(ii) there is a statistic $T = t(X)$ so that $S = (T, A)$ taken together are minimal sufficient.

If the MLE $\hat{\theta}$ is not sufficient, it is often possible to find an ancillary statistic A so that $(\hat{\theta}, A)$ is jointly sufficient. Then we also have

$$f(x \mid A = a; \theta) \propto h(x)k\{\hat{\theta}(x), a; \theta\}.$$

Thus $\hat{\theta}$ *is sufficient when considering the conditional distribution given the ancillary* A.

- The *sufficiency principle* (S) says that if $S = s(X)$ is a sufficient statistic, S carries the same evidence for the parameter θ as does X.
- The *conditionality principle* (C) says that if $A = a(X)$ is ancillary, then the conditional distribution given $A = a(x_{\mathrm{obs}})$, carries the same evidence as the unconditional experiment.
- The *likelihood principle* (L) says that all evidence in an experiment is summarized in the likelihood function.

Birnbaum's theorem says that (S) and (C) combined are equivalent to (L)!

Bayesian inference obeys (L) in the strongest form.

A statistic $T = t(X)$ is said to be (boundedly) *complete* w.r.t. θ if for all functions h

$$\mathbf{E}_\theta\{h(T)\} = 0 \text{ for all } \theta \implies h(t) = 0 \text{ a.s.}$$

In a linear exponential family, the canonical statistic $T = t(X)$ is boundedly complete and sufficient.

The Lehmann-Scheffé theorem: *if a sufficient statistic is complete, it is also minimal sufficient.*

Basu's theorem: *If $T = t(X)$ is (boundedly) complete and sufficient for θ and the distribution of A does not depend on θ, then T and A are independent.*

One instrument produces measurements $\mathcal{N}(\theta, 1)$, the other measurements which are $\mathcal{N}(\theta, 100)$.

We wish to check whether a parameter $\theta = 0$, the alternative being that $\theta > 0$.

Toss a coin *with probability λ of landing heads* and let $A = i, i = 1, 2$ denote that the instrument i is chosen. Perform then the measurement to obtain X. The joint distribution of (X, A) is determined as

$$f(x, a; \theta, \lambda) = \phi(x - \theta)1_{\{1\}}(a)\lambda + \phi\{(x - \theta)/10\}1_{\{2\}}(a)(1 - \lambda).$$

Suppose we have chosen the first instrument and observe $X = 4$. Is this consistent with the assumption $\theta = 0$?

The parameter λ is *nuisance parameter* in the sense that we are not interested in its value, but its value modifies the distribution of our observations.

If we now redo the exercise from the case where λ is known, we have the additional problem that the p-value

$$\begin{aligned} p &= P(X > 4; \theta = 0) \\ &= \{1 - \Phi(4)\}\lambda + \{1 - \Phi(.4)\}(1 - \lambda) \\ &= .00003\lambda + .34458(1 - \lambda) \end{aligned}$$

unfortunately *depends on the unknown λ.*

However, *the probability of choosing the instrument seems irrelevant once we know which instrument was in fact used.*

Thus, again we would rather consider $A = a$ fixed and condition on the actual instrument used. That is, also here *calculate the p-value as*

$$\tilde{p} = P(X > 4 \mid A = 1; \theta = 0) = \{1 - \Phi(4)\} = .00003$$

giving very strong evidence against the hypothesis. Note that λ *does not enter in this conditional calculation*.

Motivated by this example, we consider more generally a family of distributions $f(x; \theta), \theta \in \Theta$ where θ is partitioned into $\theta = (\psi, \lambda)$. We also assume that ψ is the *parameter of interest* and λ a *nuisance parameter.*

Suppose that there is a minimal sufficient statistic $T = t(X)$ partitioned as $T = (S, C) = (s(X), c(X))$ where:

C1: the distribution of C depends on λ but not on ψ;

C2: the conditional distribution of S given $C = c$ depends on ψ but not λ, for all c;

C3: the parameters vary independently, i.e. $\Theta = \Psi \times \Lambda$.

Then the likelihood function factorizes as

$$L(\theta \mid x) \propto f(s, c; \theta) = f(s \mid c; \psi) f(c; \lambda)$$

and we say that C is *ancillary for* ψ, S is *conditionally sufficient for* ψ given C, and C is *marginally sufficient for* λ.

We also say that C is a *cut* for λ.

When C is a cut, the likelihood factorizes as

$$L(\theta \mid x) \propto f(s, c; \theta) = f(s \mid c; \psi) f(c; \lambda) = L_1(\psi \mid s, c) L_2(\lambda \mid c).$$

Since ψ and λ vary independently, we may then maximize L by maximizing each of these factors separately. In other words, the maximum likelihood estimator $\hat{\theta}$ of the parameter θ satisfies

$$\hat{\theta} = (\hat{\psi}, \hat{\lambda}), \quad \text{where } \hat{\psi} = \arg\max_{\psi} L_1(\psi \mid s, c), \ \hat{\lambda} = \arg\max_{\lambda} L_2(\lambda \mid c).$$

Hence we get the *same estimate whether we use the joint distribution* $f_{(S,C)}$ *for* θ*, or* $f_{S|C}$ *for* ψ *and* f_C *for* λ*.*

Note that the equation above may indicate a *simple way of maximizing the likelihood function.*

A widely accepted *conditionality principle* says that when C is a cut for a nuisance parameter λ, *inference about ψ should be based on the conditional distribution of S given C.*

In the simple example given, this corresponds to conditioning on the instrument actually used when making inference about θ.

A possibly less well accepted principle says that when C is a cut for λ, *inference about λ should be based on the marginal distribution of C.*

Thus when making inference about the probability λ of choosing the first instrument, we should ignore the fact that the instrument was used, but only consider that it was chosen.

Another example

Consider a sample $X = (X_1, \ldots, X_n)$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Since $(\bar{X}, S^2 = \sum_i (X_i - \bar{X}_i)^2)$ is minimal sufficient, the likelihood function becomes

$$L(\mu, \sigma^2 \mid x) \propto f(\bar{x}; \mu, \sigma^2) f(s^2; \sigma^2),$$

where we have used the independence of $\bar{X}$ and S^2 and the fact that S^2 follows a $\sigma^2\chi^2$-distribution not depending on μ.

Here the situation is less clear cut. It could make sense to think of $\bar{x}$ as being sufficient for μ (which it is if σ^2 is fixed) and S^2 as ancillary for μ and sufficient for σ^2, but *it does not fit into the theory developed* as the distribution of $\bar{X}$ depends on (μ, σ^2).

Since Bayesian inference obeys the likelihood principle only the factorization itself matters:

$$L(\theta \mid x) \propto L_1(\psi \mid s, c) L_2(\lambda \mid c).$$

Still, this fact is not unimportant. Assume that the prior density satisfies

$$\pi(\psi, \lambda) = \eta(\psi)\rho(\lambda),$$

in other words that the parameters ψ and λ are prior independent. Then the posterior density satisfies

$$\pi^*(\psi, \lambda) = \pi(\psi, \lambda \mid x) \propto \eta(\psi)\rho(\lambda) L_1(\psi \mid s, c) L_2(\lambda \mid c) \propto \eta^*(\psi)\rho^*(\lambda).$$

Hence *if C is a cut for λ and ψ and λ are prior independent, they are posterior independent.*

Consider the hypothesis that the parameter of interest ψ has a specific value, i.e. $H_0 : \psi = \psi_0$. This is a *composite* hypothesis and we wish to find a test of size α so the rejection region R satisfies

$$P(X \in R; \psi_0, \lambda) = \alpha \text{ for all values of } \lambda \in \Lambda.$$

A test is said to be *similar* if this condition holds.

One way of constructing a similar test is to find a statistic C which is *sufficient for λ for fixed $\psi = \psi_0$.* This would in particular be the case if C is a cut. Now look for a set $R(c)$ such that

$$P(X \in R(c) \mid C = c; \psi_0, \lambda) = P(X \in R(c) \mid C = c; \psi_0) = \alpha,$$

where we have used the sufficiency of C to remove λ.

If we define R as $x \in R \iff x \in R(c(x))$ we then get

$$\begin{aligned} P(X \in R; \psi_0, \lambda) &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R \mid C; \psi_0)\} \\ &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R(C) \mid C; \psi_0)\} \\ &= \mathbf{E}_{(\psi_0,\lambda)}(\alpha) = \alpha. \end{aligned}$$

We have thus succeeded in constructing a similar test by this conditioning operation.

A test of this kind is said to have *Neyman structure*. An important result is that if C is complete and sufficient for λ for $\psi = \psi_0$, then *any similar rejection region R has Neyman structure.*

This is shown as follows. Assume R is a similar rejection region, i.e.

$$P(X \in R; \psi_0, \lambda) = \alpha \text{ for all } \lambda.$$

Then define $h(C) = P(X \in R \mid C; \psi_0) - \alpha$. We get

$$\begin{aligned}\mathbf{E}_{(\psi_0,\lambda)}\{h(C)\} &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R \mid C; \psi_0) - \alpha\} \\ &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R \mid C; \psi_0, \lambda) - \alpha\} \\ &= P(X \in R; \psi_0, \lambda) - \alpha = 0.\end{aligned}$$

Completeness yields $h(C) = 0$ and $P(X \in R \mid C; \psi_0) - \alpha$.

As a consequence of this result it is common, although not universally accepted, to *condition on the statistic sufficient under the hypothesis when testing composite hypothesis,* i.e. to construct tests with Neyman structure.