Ancillarity and Conditional Inference

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 1, Hilary Term 2009

January 18, 2009

Consider an experiment with two instruments available:

One instrument is very precise and produces measurements $\mathcal{N}(\theta, 1)$. The other instrument is older and less accurate; it produces measurements which are $\mathcal{N}(\theta, 100)$.

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

Toss a fair coin 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

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

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

The p-value is

$$p = P(X > 4; \theta = 0) = \{1 - \Phi(4)\}/2 + \{1 - \Phi(.4)\}/2 = .1723,$$

so there is nothing to worry about?

However, we did in fact use the precise instrument. So, with a standard deviation of 1, a value of $X = 4$ should be very unlikely. Why should it matter that we could have used the other instrument, but didn't?

Should we not rather have considered $A = a$ fixed and condition on the actual instrument used? That is, 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.

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.

Intuitively A is then *uninformative* about the unknown parameter.

In the example just given, A is such an ancillary statistic since $\hat{\theta} = X$ can play the role of T as (X, A) clearly is jointly (minimal) sufficient.

The word ‘ancillary’ both means secondary and auxiliary, each meaning referring to each of the two conditions.

Notion of ancillarity seems fundamental in statistics and is due to Fisher, but its role is less clear than that of sufficiency.

Various forms of the *conditionality principle* say that the distribution used for inference should be conditional on any ancillary, such as the instrument actually used.

Note this is a frequentist concept and plays little role in a Bayesian paradigm.

In the Fisherian paradigm, we should not compare the measurement obtained to anything we could have seen, but did not. Rather we should define a relevant *reference set* of values, for example by conditioning with an ancillary statistic, and use this set for inference calculations.

The relevant reference set may not simply be the original sample space!

In a Bayesian paradigm we only consider the value observed trough the likelihood function, which modifies the prior distribution into the posterior.

The likelihood function when observing $X = 4, A = 1$ would be

$$L(\theta \mid X = 4, a = 1) \propto \phi(4 - \theta)$$

which in itself appears to give very strong evidence against $\theta = 0$. In fact, the likelihood ratio is

$$L(0 \mid X = 4, a = 1)/L(4 \mid X = 4, a = 1) = \phi(4)/\phi(0) = 0.0003355.$$

but in the Bayesian paradigm we must combine with a prior distribution over to quantify exactly how much this modifies our beliefs about θ.

In general, 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 since

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

we also have

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

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

This is yet another argument for considering using the conditional distribution as a reference distribution.

It has several time been attempted to give statistical inference a firm foundation through so-called inference principles, for example:

- 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.

It has several time been attempted to give statistical inference a firm foundation through so-called inference principles, for example:

- 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 $A = a(x_{\text{obs}})$, carries the same evidence as the unconditional experiment.

It has several time been attempted to give statistical inference a firm foundation through so-called inference principles, for example:

- 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

Whereas some variant of (S) and (C) are commonly accepted among statisticians, (L) is not.

Birnbaum showed in 1972 that (S) and (C) combined are equivalent to (L)!

Reactions on this result have been different. The theorem depends heavily on the precise formulation of the principles (weak and strong forms) and is therefore not generally accepted as a fact.

Bayesian inference obeys (L) in the strongest form.

Attitudes towards this fact are varied...

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

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

It is *boundedly complete* if the same holds when only bounded functions h are considered.

It would be more precise to say the family of densities of T

$$\mathcal{F}_T = \{f_T(t;\theta), \theta \in \Theta\}$$

is complete, but the shorter usage has become common.

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

Consider an exponential family, with densities

$$f(x;\theta) = b(x)e^{a(\theta)^\top t(x) - c(\theta)}, \quad x \in \mathcal{X}.$$

If the family is linear, then $T = t(X)$ is boundedly complete and sufficient.

This is a non-trivial result. The proof uses analytic function theory and is outside the scope of this course.

The case of a linear exponential family is essentially the only case where a complete sufficient statistic exists, or at least where this can be proved.

For curved exponential families there is typically no complete sufficient statistic.

Sometimes it does not matter, whether we condition on A or not:

If $T = t(X)$ is complete and sufficient for θ and the distribution of A does not depend on θ, then T and A are independent.

Here is a nice application of this:

If $(X_1, \ldots, X_n)$ is a sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with *known* variance $\sigma^2 = \sigma_0^2$, it holds that $\hat{\mu} = \bar{X}$ complete and sufficient. Since the distribution of $\sum(X_i - \bar{X})^2$ cannot depend on μ, it follows that $\bar{X}$ and $\sum(X_i - \bar{X})^2$ are independent.

The proof is surprisingly simple: Let g be an arbitrary bounded function of a and let $m = \mathbf{E}_\theta\{g(A)\}$. Note m does not depend on θ as the distribution of A did not. Now let

$$h\{t(x)\} = \mathbf{E}_\theta[\{g(A) - m\} \mid T = t(x)]$$

which also does not depend on θ because T was sufficient.

Iterating expectations and using the definition of m yields

$$\begin{aligned}\mathbf{E}_\theta\{h(T)\} &= \mathbf{E}_\theta \mathbf{E}_\theta[g\{A\} - m \mid T] \\ &= \mathbf{E}_\theta\{g(A) - m\} = 0\end{aligned}$$

for all θ. Completeness then implies

$$\mathbf{E}_\theta\{g(A) \mid T = t(x)\} = \mathbf{E}\{g(A)\},$$

thus that A and T are independent.