

# PHYSTAT DARK MATTER 2019

## Nuisance parameters

H. S. Battey

*Department of Mathematics, Imperial College London*

The talk is based on joint work with David R. Cox (Oxford)

## TERMINOLOGY

Accounts of statistical theory start from several premises:

- Data are realizations of random variables.
- There is a given family of possible probability distributions for these random variables to which the “true” distribution belongs<sup>a</sup>.
- This family is called a statistical model.

---

<sup>a</sup>To an adequate order of approximation.

## STATISTICAL MODELS

*The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models, especially substantive ones, do not seem essentially different from other kinds of model.*

D. R. Cox, 1995.

## NUISANCE PARAMETERS

- Typically arise when a relatively complicated model is required.
- Needed to complete the specification (often inexplicitly) but are not of immediate subject matter concern.

## THE DANGERS OF NUISANCE PARAMETERS

- When the number of nuisance parameters is appreciable (relative to the number of independent observations), profile maximum likelihood typically produces severely biased estimators of interest parameters (Bartlett, 1937)<sup>a</sup>.
- There are comparable difficulties with Bayesian inference based on high-dimensional flat priors.

---

<sup>a</sup>In Bartlett's example, direct interpretation of the likelihood tells you that you are almost certain that the interest parameter is very close to half its true value.

## TWO BROAD APPROACHES

- 1 Treat the nuisance parameters as fixed arbitrary constants and eliminate them by appropriate manoeuvres.
- 2 Treat the nuisance parameters as realizations of independent and identically distributed random variables with a parametric distribution.

The second approach entails stronger modelling assumptions.  
The “appropriate manoeuvres” in the first take take different forms.

## EXAMPLE: MATCHED COMPARISONS

- Individuals (more generally experimental units) matched on their intrinsic features and randomized to “treatment” and “control”.
- Idealized case: monozygotic twins; two eyes of the same patient.
- Record an outcome  $T_i$  and  $C_i$  for pairs indexed by  $i = 1, \dots, n$ .

## EXAMPLE (CONTINUED): TWO MODELS

Suppose that  $T_i$  and  $C_i$  are exponentially distributed with rates

- 1  $\gamma_i\psi$  and  $\gamma_i/\psi$  respectively (multiplicative model); or
- 2  $\rho_i + \Delta$  and  $\rho_i - \Delta$  respectively (additive model).

Interest is in the “treatment effects”  $\psi$  and  $\Delta$ .

Pair effects  $\gamma_i$  and  $\rho_i$  are of no direct concern.

There are  $2n$  observations and  $n + 1$  parameters.

## TREATING THE $(\gamma_i)_{i=1}^n$ AS FIXED CONSTANTS

$T_i \sim \exp(\gamma_i \psi)$  and  $C_i \sim \exp(\gamma_i / \psi)$ .

The ratio  $Z_i \triangleq T_i / C_i$  has density function at  $z$

$$\frac{\psi^2}{(1 + \psi^2 z)^2}.$$

Free of nuisance parameters.

Fit  $\psi$  by maximum likelihood based on the ratios.

Fisher information per observation is  $(4/3)\psi^{-2}$ .

## TREATING THE $(\gamma_i)_{i=1}^n$ AS RANDOM VARIABLES

$T_i \sim \exp(\gamma_i \psi)$  and  $C_i \sim \exp(\gamma_i / \psi)$ .

Assume that the  $(\gamma_i)_{i=1}^n$  are i.i.d. gamma distributed (shape parameter  $\alpha$ , rate  $\beta$ ).

The induced joint density function of  $T_i$  and  $C_i$  at  $(t, c)$  is

$$\frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\psi t + c/\psi + \beta)^{\alpha+2}}.$$

Only 2 nuisance parameters instead of  $n$ .

Fisher information per observation is  $2(\alpha + 2)(\alpha + 3)^{-1}\psi^{-2}$ .

The two limits are  $2\psi^{-2}$  and  $(4/3)\psi^{-2}$  as  $\alpha \rightarrow \infty$  and  $\alpha \rightarrow 0$ .

## ERRONEOUSLY TREATING THE $(\gamma_i)_{i=1}^n$ AS RANDOM VARIABLES

Suppose that the model is as in the previous slide, but the assumption of i.i.d. gamma distributed random effects is erroneous.

- The resulting maximum likelihood estimator is consistent.
- It has potentially much larger variance than the estimator obtained by assuming the nuisance parameters are arbitrary constants.
- The difference depends in a rather complicated way on the apparent relative dispersion of the nuisance parameters.

## TREATING THE $(\rho_i)_{i=1}^n$ AS FIXED CONSTANTS

$T_i \sim \exp(\rho_i + \Delta)$  and  $C_i \sim \exp(\rho_i - \psi)$ .

The density function of  $T_i$  at  $t$ , conditional on the realization  $s_i$  of  $S_i \triangleq T_i + C_i$  is

$$\frac{2\Delta e^{-2\Delta t}}{1 - e^{-2\Delta s_i}}.$$

Free of nuisance parameters.

Fit  $\Delta$  by maximum likelihood based on the the conditional density.

## GENERAL PRINCIPLES

Let  $\psi$  be an interest parameter and  $\lambda$  be a nuisance parameter.  
From an arbitrary partition  $(T, C)$ , make a bijective transformation  $(T, C) \rightarrow (S, R)$  such that one of the following factorizations holds.

- (i)  $f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \lambda) f_S(s; \psi),$
- (ii)  $f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \psi) f_S(s; \lambda),$
- (iii)  $f_{S,R}(s, r; \psi, \lambda) = f_R(r; \lambda) f_S(s; \psi),$
- (vi)  $f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \lambda, \psi) f_S(s; \psi),$
- (v)  $f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \psi) f_S(s; \psi, \lambda).$

## FACTORIZATION (i)

$$f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \lambda)f_S(s; \psi).$$

A case for marginal likelihood<sup>a</sup> based on  $f_S(s; \psi)$   
with  $S$  sufficient for  $\psi$ .

---

<sup>a</sup> Statisticians' terminology differs from physicists' terminology here.

## FACTORIZATION (ii)

$$f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \psi) f_S(s; \lambda).$$

A case for conditional likelihood based on  
 $f_{R|S}(r|s; \psi)$ .

## FACTORIZATION (iii)

$$f_{S,R}(s, r; \psi, \lambda) = f_R(r; \lambda) f_S(s; \psi).$$

A case for marginal likelihood based on  $f_S(s; \psi)$ .  
The jointly sufficient statistic is two independent  
sufficient statistics.

## FACTORIZATIONS (iv) and (v)

$$f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \lambda, \psi)f_S(s; \psi),$$

$$f_{S,R}(s, r; \psi, \lambda) = f_{R|S}(r|s; \psi)f_S(s; \psi, \lambda).$$

Marginal likelihood is applicable for (iv) and conditional likelihood for (v), but information on  $\psi$  is lost in either case.

## CONNECTIONS

- Parameter orthogonalization, i.e., interest-respecting reparameterization (Cox and Reid, 1987).
- Modified profile likelihood (Barndorff-Nielsen, 1983; Cox and Reid, 1987; Cox and Reid, 1992).

## ASSESSING MODEL ADEQUACY: GENERAL STRATEGIES

- ① Parameterize model space.
- ② Sufficiency arguments.

## PARAMETERIZATION OF MODEL SPACE

A single parameterization embracing the separate models:

$$\{1 + \lambda(\alpha_i - \theta)\}^{1/\lambda}, \quad \{1 + \lambda(\alpha_i + \theta)\}^{1/\lambda}.$$

Here  $\lambda$  specifies the form of relation to be fitted.

- $\lambda \rightarrow 0$  recovers the multiplicative rates model ( $\alpha_i = \log \gamma_i$ ,  $\theta = \log \psi$ );
- $\lambda = 1$  recovers the additive rates model ( $\alpha_i = \rho_i - 1$ ,  $\theta = \Delta$ );
- $\lambda = -1$  captures another model in which the treatment has an additive effect on the means ( $\alpha_i = \xi_i + 1$ ,  $\theta = \phi$ , say).

## SUFFICIENCY ARGUMENTS

- The data  $z$  are treated as realizations of  $Z = (Z_1, \dots, Z_n)$ .
- For any reasonable model  $m \in \mathcal{M}$  with parameter vector  $\theta_m$ , identify the sufficient statistic  $S_m$  for  $\theta_m$ .
- A model is compatible with the data if  $z$  is not extreme when calibrated against the conditional distribution of  $Z$  given  $S_m = s_m$ .

## SUFFICIENCY ARGUMENT: EXAMPLE

- Suppose (for a potential contradiction) that the multiplicative treatment effect model is true. The likelihood contribution from the  $i$ th pair is

$$\gamma_i^2 \exp(-\gamma_i c_i / \psi) \exp(-\gamma_i \psi t_i).$$

- For any given  $\psi$ ,  $S_i(\psi) \triangleq C_i / \psi + T_i \psi$  is sufficient for  $\gamma_i$  and has density function

$$f_{S_i(\psi)}(s) = \gamma_i^2 s \exp(-\gamma_i s).$$

## SUFFICIENCY ARGUMENT: EXAMPLE (CONTINUED)

- The conditional density of  $T_i$  at  $t_i$ , given  $S_i(\psi) = s_i(\psi)$ , is

$$\frac{\gamma_i^2 \exp\{-\gamma_i s_i(\psi)\}}{\gamma_i^2 s_i(\psi) \exp\{-\gamma_i s_i(\psi)\}} = \frac{1}{s_i(\psi)},$$

showing that  $T_i | S_i(\psi) = s_i(\psi)$  is uniformly distributed between 0 and  $s_i(\psi)$ .

- For any hypothesized value  $\psi_0$  of  $\psi$ , compatibility of the proportional treatment effects model and  $\psi_0$  with the data corresponds to compatibility of the realizations of  $U_i(\psi_0) \triangleq T_i/s_i(\psi_0)$  with a uniform distribution on  $(0,1)$  for all  $i = 1, \dots, n$ .

## SEVERAL POTENTIAL MODELS

- If the data are compatible with several reasonable models with different subject-matter implications, one should aim to specify them all. A “confidence set of models”.
- Any choice between models in the confidence set would require additional data or subject-matter expertise.

# Thank you for your attention

## References

The lecture was based on:

Battey, H. S. and Cox, D. R. (2019), High-dimensional nuisance parameters: an example from parametric survival analysis. *Under review*.

and (for the last slide)

Cox, D. R. and Battey, H. S. (2017), Large numbers of explanatory variables, a semi-descriptive analysis. *Proc. Nat. Acad. Sci.*, 114, 8592–8595.

Battey, H. S. and Cox, D. R. (2018), Large numbers of explanatory variables: a probabilistic assessment. *Proc. R. Soc. Lond. A*, 474, 20170631.

Other papers referenced:

- Barndorff-Nielsen, O. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*, 160, 268–282.
- Barndorff-Nielsen, O. (1983). On a formula for the distribution of a maximum likelihood estimator. *Biometrika*, 70, 343–365.
- Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc. B*, 49, 1–39.
- Cox, D. R. and Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika*, 79, 408–4011.