CERN, June 2007

D.R.Cox

Nuffield College, Oxford, UK

david.cox@nuffield.ox.ac.uk

Some methodological themes

- Non-probabilistically based ideas
  - clustering algorithms
  - visualization
- design of experiments and sampling procedures
- model construction
  - families of empirical models
  - substantive stochastic models
- model analysis and criticism
- formal theory

Five faces of Bayesian statistics

- empirical Bayes; number of similar parameters with a frequency distribution
- neutral (reference) priors: Laplace, Jeffreys, Jaynes, Berger and Bernardo
- information-inserting priors (evidence-based)
- personalistic priors
- technical device for generating frequentist inference

The killer question

Model formulation

Translates physics question into a statistical question Two broad

approaches

- very general families of models (with appropriate software)
- very specific models, often based on stochastic processes or sometimes deterministic theory with *ad hoc* error terms added

Large numbers of tests

Suppose that a large number n of significance test are done each with its own null hypothesis that may or may not be true. If we report only the most significant p-value clearly we are likely to be misled if in fact all null hypotheses are true. This is a well-understood selection effect.

There are two distinct problems

- some small but almost certainly non-zero number of the hypotheses are false; which are they?
- it is quite possible that all the null hypotheses are true; how strong is the evidence against the hypothesis with $m =$min(p_j)?

Selection of real effects

Two approaches

- false discovery rate aims to control

$$E(F/R \mid R > 0),$$

  where F is the number of falsely rejected hypotheses out of R rejected

- empirical Bayes approach

Stronger formulation in terms of exceedances; controls the probability that FDR exceeds a given bound

Empirical Bayes approach

Simplest formulation

- n test statistics $T_1, \ldots, T_n$
- under relevant null hypothesis each statistic has density $f_0(t)$
- under alternative density is $f_1(t)$
- proportion θ of null hypotheses false

Simplest formulation

- suppose $f_0(t)$ known, wlg N(0, 1)
- suppose $f_1(t)$ is N$(\mu_1, 1)$, one-sided version
- if (θ, μ) known the posterior odds that a value t comes from the non-null versus from the null distribution are

$$\log \frac{P(f_1 \mid t)}{P(f_0 \mid t)} = \log \frac{\theta}{1-\theta} + \mu_1(t - \mu_1/2).$$

- hence once (θ, μ) estimated posterior odds can be found

More selective than accept, reject. If a single threshold is required it can be found to produce an assigned false discovery rate.

More elaborate version

Given extensive data both $f_1(t)$ and even $f_0(t)$ and θ can be estimated nonparametrically.

None of the methods depends critically on the assumed independence of the test statistics. Dependencies that are local in some sense thin out on passing to extremes.

Global null hypothesis

Consider $m = \min(p_j)$ as test statistic. If the individual tests are independent the associated p-value is

$$1 - (1 - m)^n$$

and without the independence assumption

$$mn$$

is an upper bound, often sharp.

If n is large, for example of the order of 10^3 this means that to achieve an interesting level of significance m must be extremely small. This involves sensitivity to the assumptions involved in calculating the p_j to what is typically an unrealistic level.

A possible robust method

Let $y_j = -\log p_j$. Under the null hypotheses these have an exponential distribution of unit mean. Equivalently $2y_j$ is chi-squared with 2 degrees of freedom. Order the values

$$y_{(1)} \geq y_{(2)} \geq \ldots \geq y_{(n)}.$$

It can be shown that under a global null hypothesis these have expected values

$$1 + 1/2 + \ldots + 1/n, 1/2 + \ldots + 1/n, \ldots, 1/n.$$

A plot should show a line of unit slope possibly with the last value (or last few values) above the line. Distortion of distributional form of test statistics would leave smooth curve possibly with outliers.

Illustration provided by David Clayton (Welcome Inst of Genetics, Cambridge)

- approximately 0.5 $\times$ 10^6 p-values
- plot showed good agreement with unit line to some way past 1% point
- thereafter smooth curve up to largest values
- no evidence of sudden jumps upwards

In fact comparisons were of two control groups so that all null hypotheses correct

Bayesian formulation and solution

A few references

Schweder, T. and Spjotvoll, E.J. (1982). *Biometrika* 69, 493-502.

Cox, D.R. and Wong, M.Y. (2004). *J.R. Statist. Soc.* B **66**, 395-400.

Efron, B., Tibshirani, R. and Storey, J. (2001). *J. Amer. Statist. Assoc.* **96**, 1151-1160.

Wong, M.Y. and Cox, D.R. (2007) *J. Appl. Statist.* , to appear

Benjamini, Y. and Hochberg, Y. (1995) *J.R. Statist. Soc.* B, **57**, 289-300.

Genovese, C. R. and Wasserman, L. (2006). *J. Amer. Statist. Assoc.* **101**, 1408-1417.

Storey, J. D. (2002). *J.R. Statist. Soc.* B **64**, 479-498.

Storey, J. D. (2004). *Ann. Statist.* **31**, 2013-2035.