Marginalise or Profile? A Statisticians' View

Anthony Davison

Desiderata

- We aim for (frequentist) inferences that are:
  - **relevant** — we compare the data actually observed to an appropriate **reference set** $\mathcal{S}$ of datasets that we might have observed;
  - **calibrated** — probability statements made with respect to $\mathcal{S}$ are accurate, in R. A. Fisher's sense:

    > *We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.*

  - **secure** — the conclusion does not depend too much on secondary details of the problem formulation, is not strongly perturbed by bad data, ...

Trivial example: Relevance

- $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{U}(\theta - 1/2, \theta + 1/2)$, with θ unknown.
- Minimal sufficient statistic consists of the smallest and largest order statistics, $S = (Y_{(1)}, Y_{(n)})$, or equivalently $A = Y_{(n)} - Y_{(1)}$ and $T = (Y_{(1)} + Y_{(n)})/2$.
- A tells us nothing about θ itself, so the joint density of the data Y

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

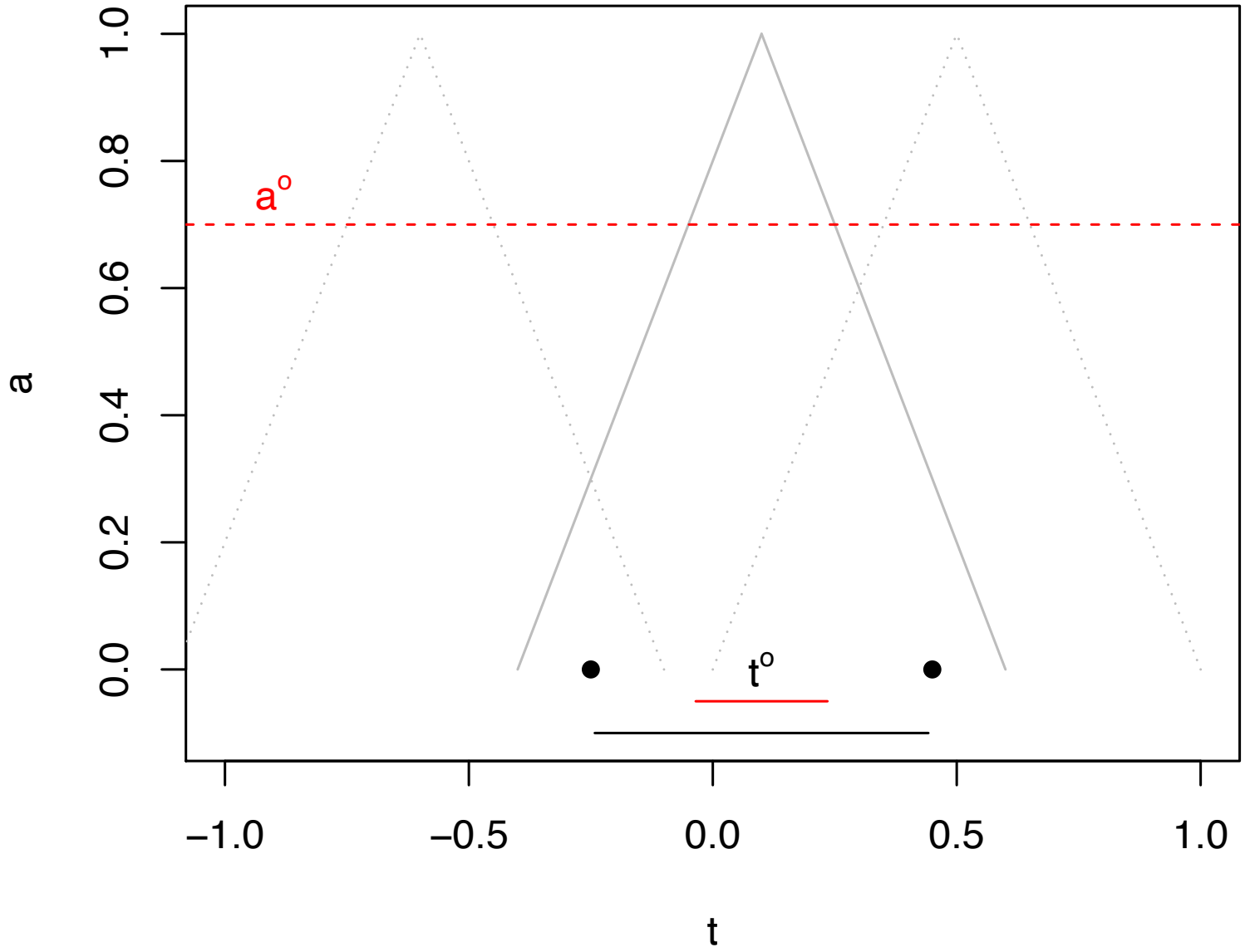
where

$$f_{T|A}(t \mid a;\theta) = \frac{1}{1-a}, \quad \theta - \tfrac{1-a}{2} \leq t \leq \theta + \tfrac{1-a}{2},$$

and we see that A specifies the precision of inference on θ: $a \lesssim 1$ gives high precision, $0 \lesssim a$ gives low precision.

- Here the reference set is $\mathcal{S} = \{y : a(y) = a^{\circ}\}$, where a° is the observed value of A.

Trivial example



Calibration

- We base inference (tests, confidence sets) for θ on the **significance probability/P-value function**

$$\mathrm{P}(T \leq t^{\mathrm{o}} \mid A = a^{\mathrm{o}}; \theta) = \int_{-\infty}^{t^{\mathrm{o}}} f_{T|A}(t \mid a^{\mathrm{o}}; \theta)\, \mathrm{d}t,$$

  for example, solving

$$\mathrm{P}(T \leq t^{\mathrm{o}} \mid A = a^{\mathrm{o}}; \theta) = \alpha/2, 1 - \alpha/2,$$

  to get $1 - \alpha$ **conditional confidence interval**

$$\mathcal{I}^{\mathrm{o}}_{1-\alpha} = (\theta_{\alpha/2}, \theta_{1-\alpha/2}) \equiv t^{\mathrm{o}} \pm (1 - \alpha)(1 - a^{\mathrm{o}})/2.$$

- $\mathcal{I}^{\mathrm{o}}_{1-\alpha}$ is **relevant** because it takes into account the precision of the inference, and **calibrated** because $\mathrm{P}_{\mathcal{S}}(\theta \in \mathcal{I}^{\mathrm{o}}_{1-\alpha} \mid A = a^{\mathrm{o}}) = 1 - \alpha$ for any α.
- The **unconditional confidence interval** based on f_T is not relevant, and hence can give logical inconsistencies — e.g., with $n = 2$, $y_{(1)} = -0.25$, $y_{(2)} = 0.45$, $(t, a) = (0.1, 0.7)$ and
  - conditional 0.9 CI $(-0.035, 0.235)$ of length 0.27,
  - unconditional 0.9 CI $(-0.242, 0.442)$ of length 0.68, which contains the logically impossible value $\theta = 0.3$.

Any statistical models we set up will have

- **primary** aspects, which relate to the question of interest, usually
  - the basic model structure, and
  - with the key issues summarised in an **interest parameter** ψ;

  and
- **secondary** aspects, needed to complete the model but not a main focus, such as
  - the error structure (sometimes), and
  - **nuisance parameters** λ summarising unknowns of secondary interest.
- Ideally inferences will not depend heavily on the nuisance parametrisation (can be useful for numerical work), so aim for **invariance** to interest-preserving reparametrisations

$$\psi \mapsto \eta = g(\psi), \quad \lambda \mapsto \zeta = h(\lambda, \psi).$$

  Not all methods achieve this.
- We also hope for **robustness** to bad data etc.

Complications

- **Nuisance parameters** and **discreteness** degrade calibration.
- Nuisance parameters: ideally we would have factorisation

$$f(y;\psi,\lambda) = f(t,a;\psi,\lambda)f(y \mid t,a) = f(t \mid a;\psi)f(a;\lambda)f(y \mid t,a),$$

  and would then base inferences on (no loss of information on ψ)

$$\mathrm{P}(T \leq t^{\mathrm{o}} \mid A = a^{o};\psi).$$

- More usually we have one of the factorisations (some loss of information on ψ)

$$f(t,a;\psi,\lambda) = f(t \mid a;\psi)f(a;\psi,\lambda), \quad f(t;\psi,\lambda) = f(t;\psi)f(a \mid t;\psi,\lambda),$$

  leading to a conditional or a marginal likelihood and corresponding inference.
- In the worst case there is no explicit sufficient reduction from Y to $S = (T,A)$, and we base inferences on

$$\mathrm{P}(Z \leq z^{\mathrm{o}};\psi,\lambda)$$

  with Z some (approximate) pivot involving ψ (e.g., the MLE $\widehat{\psi}$, the LR statistic, ...).
- We have to replace the unknown λ by some estimate(s), and this introduces error.

Exponential families

- Linear exponential family in natural parametrisation has density

$$f(y;\psi,\lambda) = m(y)\exp\left\{t(y)^{\mathrm{T}}\psi + a(y)^{\mathrm{T}}\lambda - k(\psi,\lambda)\right\},$$

and it turns out that

$$f(t \mid a;\psi) = m(t,a)\exp\left\{t(y)^{\mathrm{T}}\psi - k_a(\psi)\right\},$$

so dependence on λ is removed by conditioning on $A = a^o$.

- If ψ is scalar then inference is based on the conditional significance function of $T = t(Y)$ given $A = a(Y^{\mathrm{o}}) = a^o$, i.e.,

$$\mathrm{P}(T \leq t^{\mathrm{o}} \mid A = a^o;\psi),$$

though with some (often rather little) loss of information on ψ.

- In the Poisson model a **cut** allows the density to be rewritten as

$$f(y;\psi,\lambda) \propto f(t \mid a;\psi)f(a;\zeta), \quad \text{where} \quad \zeta = h(\psi,\lambda),$$

i.e., no information on ψ can be recovered from a.

- **Tangent exponential approximation** gives very accurate inferences for general densities.

Calibration error

- If the random variable $\psi_\alpha \equiv \psi_\alpha(Y)$ is an upper α-level confidence limit for ψ, having

$$\mathrm{P}(\psi \leq \psi_\alpha) = \alpha, \quad 0 < \alpha < 1,$$

would give perfectly calibrated inferences.

- In a **classical** asymptotic setup with p parameters (fixed) and 'sample size' $n \to \infty$,

$$\mathrm{P}(\psi \leq \psi_\alpha) = \alpha + \underbrace{\underbrace{\underbrace{A_1 n^{-1/2}}_{\text{first order error}} + \underbrace{A_2 n^{-1}}_{\text{second order error}}}_{} + \underbrace{A_3 n^{-3/2} + \cdots}_{\text{third order error}}}_{}, \quad 0 < \alpha < 1,$$

where $A_1, A_2, \ldots$ depend on ψ, λ and on α.

- In a '**modern**' asymptotic setup with $p \to \infty$ as $n \to \infty$, (very broadly) calibration results for the classical case hold if $p = o(n^{1/3})$ (Sartori, 2003, *Biometrika*; Tang and Reid, 2020, *JRSSB*). This is as good as we could hope for in general setups, because the bias for ML estimation of ψ is $O(p^3/n)$.

- These are **asymptotic** results, numerical work suggests surprising accuracy for certain situations even when the asymptotics 'fail'.

Profiling

- Under classical asymptotics and when ψ equals its true value ...
- Log likelihood function $\ell(\psi, \lambda)$ and Wilks' theorem lead to basing inference on **likelihood ratio statistic**

$$W(\psi) = 2\left\{\ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi, \widehat{\lambda}_\psi)\right\} \dot{\sim} \chi^2_{\dim \psi},$$

  where $(\widehat{\psi}, \widehat{\lambda})$ are overall MLE, and $(\psi, \widehat{\lambda}_\psi)$ is MLE for fixed ψ.
- When ψ is scalar we can base inferences on the **likelihood root**

$$R(\psi) = \text{sign}(\widehat{\psi} - \psi)\sqrt{W(\psi)} \dot{\sim} \mathcal{N}(0, 1).$$

- W and R are invariant to interest-preserving reparametrisations.
- It turns out that for W and R,
  - one-sided inference has first-order error,
  - two-sided inference has second-order error (A_1 terms cancel).
- The issue is that profiling introduces bias in finite samples, so the log likelihood is shifted to the left or right.
- This bias is reduced by the **modified likelihood root** $R^*(\psi)$, which gives inferences that are accurate to third order for continuous y and to second order for discrete y.

Simulation

- The standard normal approximation to $R(\psi)$ uses the approximate significance function

$$\mathrm{P}\left\{R(\psi) \leq r^{\mathrm{o}}(\psi)\right\} \doteq \Phi\left\{r^{\mathrm{o}}(\psi)\right\},$$

  which has first-order error.

- One obvious improvement is to generate S datasets $Y^{\dagger}$ from the model with parameters $(\psi, \widehat{\lambda}_{\psi})$ and to replace the RHS of the above with

$$\mathrm{P}^{\dagger}\left\{R^{\dagger}(\psi) \leq r^{\mathrm{o}}(\psi)\right\} \doteq \frac{\#\{r^{\dagger}(\psi) \leq r^{\mathrm{o}}(\psi)\}}{S},$$

  for very large S and a range of ψ.

- Lee and Young (2005, *Statistics and Probability Letters*) and DiCiccio and Young (2008, 2010, *Biometrika*) show that this parametric bootstrap procedure has third-order relative accuracy in linear exponential families, both conditionally and unconditionally, and is closely linked to objective Bayes approaches.

- I think it is unknown whether the third-order accuracy is preserved in more general settings (even curved exponential families) . . . but more general bootstrap considerations suggest that at best second-order accuracy is possible.

Marginalisation

- Marginalisation involves a prior for λ and then basing inference on the marginal density

$$f(y;\psi) = \int f(y;\psi,\lambda)\pi(\lambda)\,\mathrm{d}\lambda,$$

  so even first-order inference for ψ will depend on the chosen $\pi(\cdot)$ (neither secure nor calibrated).

- Cox and Wang (2010, *Biometrika*) compare this with conditioning to remove λ in a Poisson model, and show that little information is gained by marginalisation, and that mis-specifying π may lead to appreciable bias. Ongoing research aims to find what aspect of the mis-specification is crucial.

- We could regard this as **pauper's Bayes**, with joint prior $\pi(\psi,\lambda) \propto \pi(\lambda)$, leading to posterior CDF

$$\mathrm{P}(\psi \leq \psi' \mid y^{\mathrm{o}}) = \frac{\int_{-\infty}^{\psi'} \int f(y^{\mathrm{o}};\psi,\lambda)\pi(\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}{\int\int f(y^{\mathrm{o}};\psi,\lambda)\pi(\lambda)\,\mathrm{d}\lambda\mathrm{d}\psi}.$$

- Laplace approximation to both integrals then leads to a Bayesian version of the modified likelihood root r^*, which is a second-order approximation to the CDF for continuous y.

TABLE 3

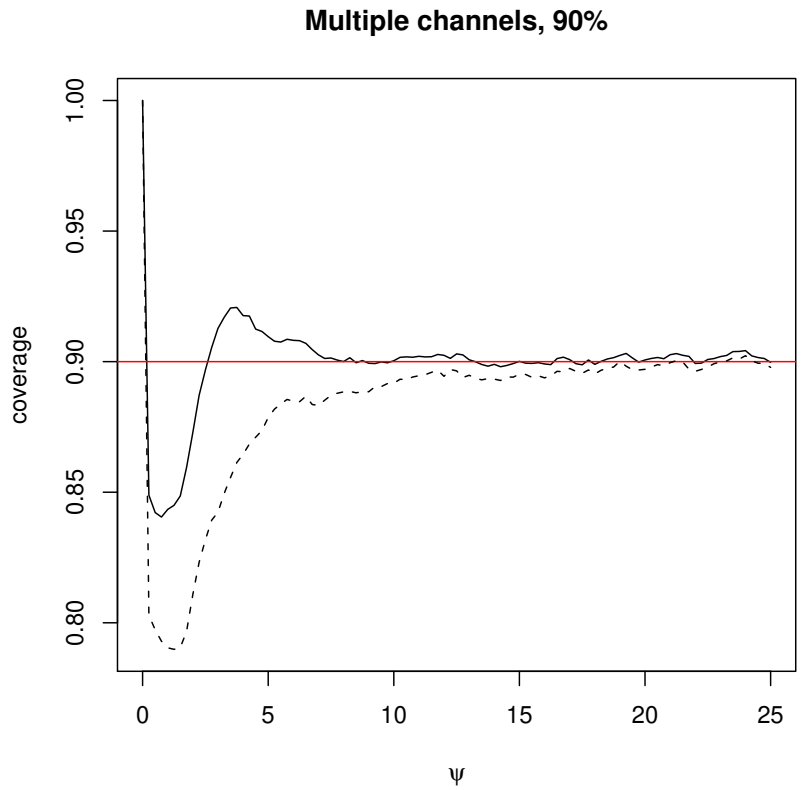
Empirical coverage probabilities in a multiple-channel simulation with 10,000 *replications*, $\psi = 2$, $\beta = (0.20, 0.30, 0.40, \ldots, 1.10)$, $\gamma = (0.20, 0.25, 0.30, \ldots, 0.65)$, $t = (15, 17, 19, \ldots, 33)$ *and* $u = (50, 55, 60, \ldots, 95)$

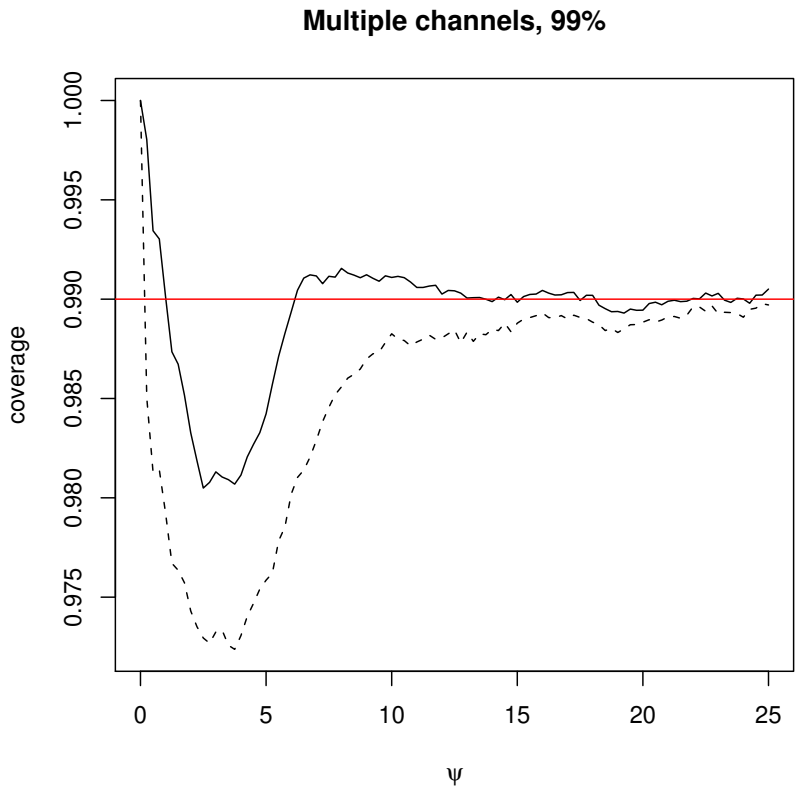
Probability	r	r^*	r^*_B
0.0100	0.0099	0.0101	0.0109
0.0250	0.0244	0.0255	0.0273
0.0500	0.0493	0.0519	0.0542
0.1000	0.0967	0.1012	0.1035
0.5000	**0.4869**	0.5043	0.5027
0.9000	**0.8900**	0.9013	0.8942
0.9500	**0.9421**	0.9499	**0.9427**
0.9750	**0.9687**	0.9759	**0.9689**
0.9900	**0.9875**	0.9913	**0.9864**

Figures in bold differ from the nominal level by more than simulation error.

(Davison and Sartori, 2008, *Statistical Science*)

Multichannel simulation, $c = 10$



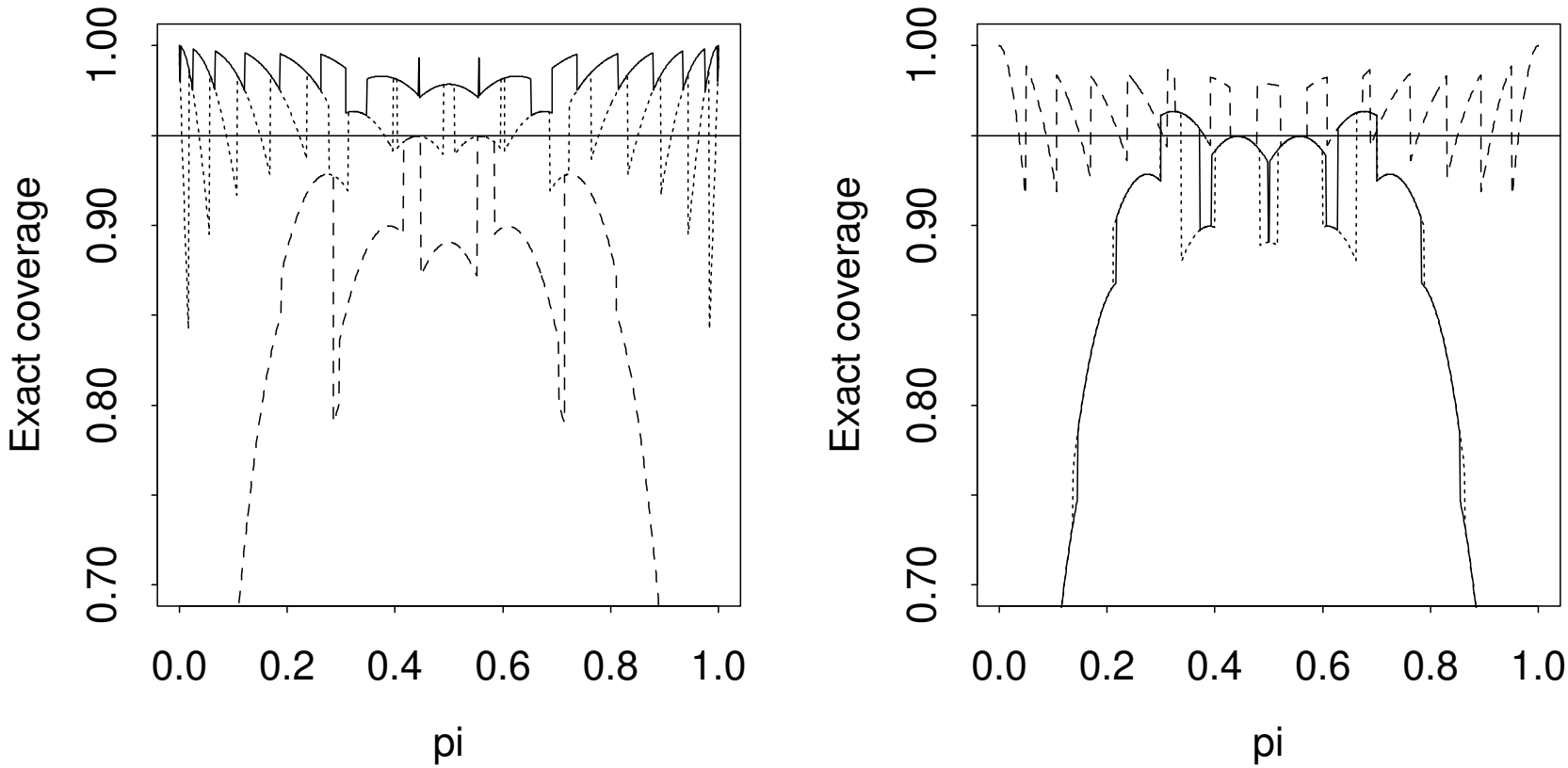


Target coverage (red), coverage of r^* (black), coverage of r_B^* (dashes), as a function of interest parameter ψ.

(Davison and Sartori, 2008, *Statistical Science*)

Discreteness

Exact coverages of 0.95 binomial confidence intervals, with $m = 10$. The horizontal line shows the target coverage. Left: exact (solid), score (dots) and maximum likelihood estimator (dashes). Right: signed likelihood ratio statistic (solid), modified signed likelihood ratio statistic (dots) and modified maximum likelihood (or Bayes) estimator obtained by replacing m and r by $m + 2$ and $r + 1$ (dashes).



Discreteness

The binomial example works better with higher m, but anyway suggests

- conservative exact intervals/loss of power for exact tests;
- approximate intervals (especially Bayes ones) may be preferable.

Conclusion

- Frequentist inferences should be be relevant, calibrated and secure.
- Relevance involves comparison of actual data with suitable reference set, and often conditioning (explicit or implicit).
- Many models of interest are (curved) exponential families, so some conditioning is needed for relevance and to eliminate nuisance parameters.
- To achieve calibration we aim beyond first-order approximation.
- Likelihood ratio statistic (profiling) is first-order accurate in general, but modified versions are (conditionally) second- or third-order accurate.
- Simulation reduces the error from profiling in special cases, but (probably) not in general. Still, it will improve numerical behaviour.
- Marginalisation seems unwise in general unless the prior used is known to be well-specified.
- Discreteness tends to give conservative inferences.

Thanks!