Calibrated Bayes

Roderick Little

UNIVERSITY OF MICHIGAN

Outline of talk

1. Strengths and weaknesses of the frequentist paradigm
2. Strengths and weaknesses of the Bayesian paradigm
3. Towards a resolution of the Bayes / frequentist schism: calibrated Bayes

See: Little, R.J.A. (2006). Calibrated Bayes: A Bayes/frequentist Roadmap. *The American Statistician*, 60, 3, 213-223.

Bayes (B) vs Frequentist (F): Why it matters

- "Inferential schizophrenia" should be avoided – mixing B and F yields inconsistencies
  - E.g. model vs design-based survey inference: current B/F compromise can give wider intervals for more data.
- Philosophical differences sow confusion and division in many areas of statistical application
- Our credibility as statisticians is undermined when we can't agree on the fundamentals of our subject
- Bayesians (B) and frequentists (F) can get different answers, on basic and complex problems

B & F get different answers: a CI is not a PI

Example: Single sample inference with bound on precision

An iid normal sample with $n = 7$, with $\bar{y} = 1$, $s = 1$ yields

$$I^{BRP}_{.05}(s) = I^{F}_{.05}(s) = \bar{y} \pm 2.447\left(s/\sqrt{n}\right) = 1 \pm 0.92 \qquad (1)$$

Experimenter E tells us that sd $\sigma = 1.5$

$$I^{BRP}_{.05}(\sigma = 1.5) = I^{F}_{.05}(\sigma = 1.5) = 1 \pm 1.96\left(1.5/\sqrt{7}\right) = 1 \pm 1.11 \qquad (2)$$

E: oops there's more variance! In fact $\sigma > 1.5$!

$$I^{BRP}_{.05}(\sigma > 1.5) = 1 \pm 1.45 \qquad (3)$$

What does a frequentist do? Pick your poison:

(1) is an exact 95% CI but is clearly the wrong inference!

(2) is an anti-conservative 95% CI (though it contains (1)!)

(3) is correctly wider than (2), but it's Bayes, not a 95% CI, and depends on the choice of prior

Models or methods?

- In general, we seem divided about whether the goal of statistics is to model the data or develop an estimation procedure
  - Model versus estimating equation?
  - Likelihood versus method of moments?
  - Methods approach seeks to avoid assumptions – but assumptions are sometimes buried
  - Models make assumptions explicit, but do modelers pay enough attention to model checks?
  - See e.g. Breiman (2001).

Strengths of frequentist inference

- Focus on repeated sampling properties tends to ensure that inferences have good frequentist properties (are well calibrated)
  - E.g. in survey sampling setting, automatically takes into account complex survey design features (unlike early model approaches)
- No need to specify prior distributions
- Flexible range of procedures
  - Come up with a method (even Bayes), and we can assess it's frequentist properties

Weaknesses of the frequentist paradigm

- Incomplete, ambiguous, incoherent
- Incomplete: "not enough answers" – exact finite-sample frequentist solutions are limited to a narrow class of problems.
  - E.g. Behrens-Fisher problem: comparison of means in two independent samples with different means and variances. Lots of approximate answers, rather than approximations of an exact answer
- Ambiguous: e.g. about conditioning (the choice of "reference sets" for frequentist inference
- Incoherent: violates the likelihood principle

F is ambiguous, B gets different answers

- Example 1: Independence in 2x2 Contingency Table

Outcome

Treatment	S	F
A	170	2
B	162	9

$H_0 : \pi_A = \pi_B;\ H_a : \pi_A > \pi_B$

Alternative tests

Pearson chi-squared (C)	P=0.016
Yates continuity corrected (Y)	P=0.032
Fisher exact test (F)	P=0.030
Bayes $\Pr(\pi_A < \pi_B \mid data)$	Pr=0.013

Independence in 2x2 tables

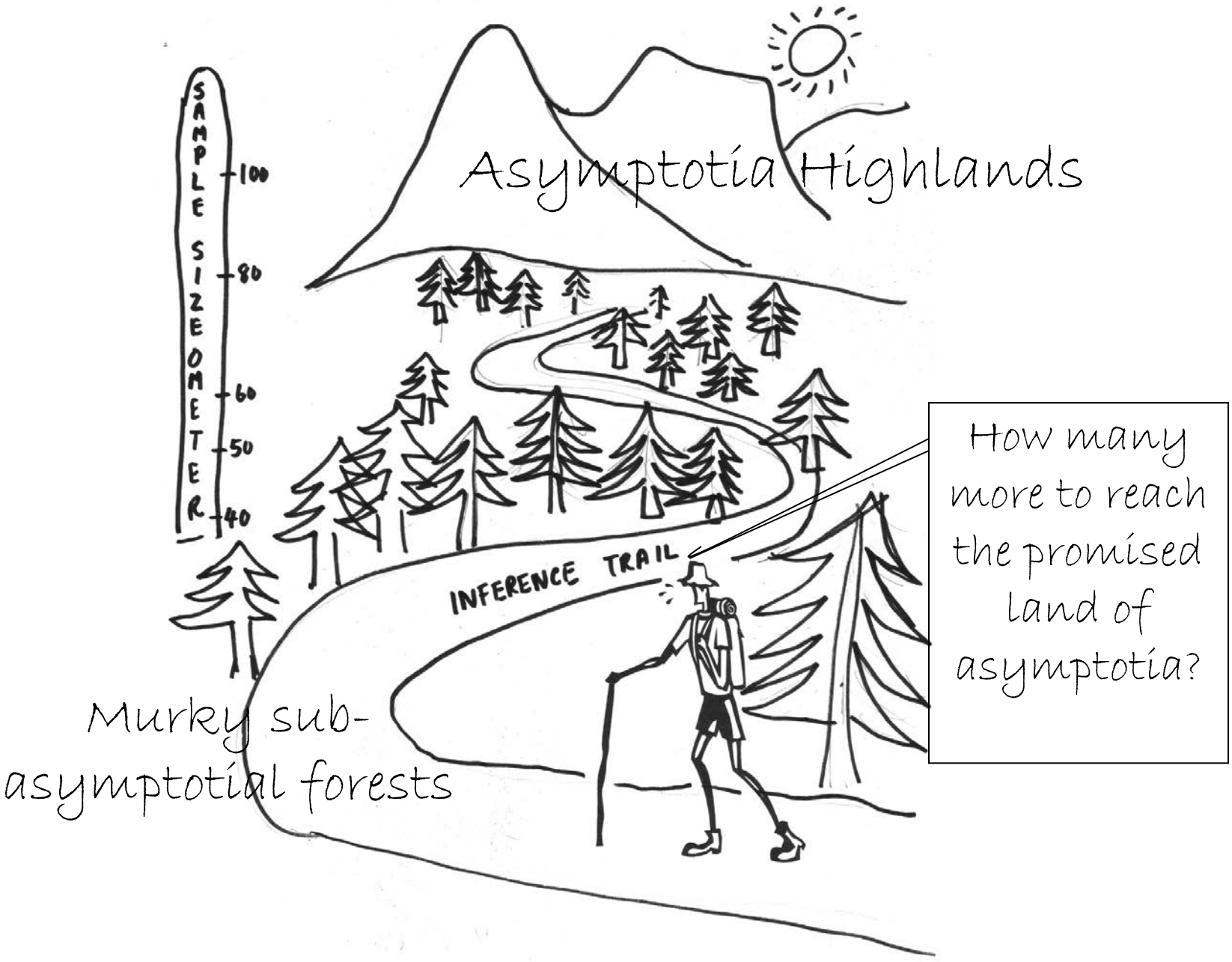
- Choice of test doesn't matter in large samples, but it does in small/moderate samples
- Fisher test is conservative when one margin is fixed (as is common in many practical designs), but exact if both margins are fixed
- Should we condition on second margin or not? It's approximately (but not exactly) ancillary for odds ratio (Yates 1984, Little 1989)
- Frequentist theory is ambiguous, Bayes yields different results

Weaknesses of the frequentist paradigm

- Not prescriptive: a set of principles for assessing properties of inference procedures rather than an inferential system
  - Distinguish “the inference” from “properties of the inference”
- No unified theory for how to generate these procedures, e.g.
  - Least squares? Too limited!
  - ML/GEE? OK, but how to choose the equations, and theory is basically asymptotic
  - Unbiasedness? Doesn’t work! (e.g. Basu’s elephant example)

(Over?)emphasis on asymptotic properties

- Lack of a satisfactory exact small-sample theory has led much frequentist theory to be asymptotic
- Current enthusiasm for semi-parametric efficiency, asymptotic results is driven by a search for robustness without modeling assumptions (not to mention elegant mathematics)
- Much useful work here, but unclear how relevant asymptotic theory is for finite sample sizes …


Asymptotia Highlands
SAMPLE SIZEOMETER
100
80
60
50
40
INFERENCE TRAIL
How many more to reach the promised land of asymptotia?
Murky sub-asymptotial forests

Strengths of Bayesian paradigm

- Complete, coherent, prescriptive for inference
- Complete
  - Solutions where no exact frequentist answer exists
  - Results with reference priors mimic many results from frequentist inference
  - Allows prior information to be incorporated when available
  - Satisfying treatment of nuisance parameters
  - Credibility intervals, not confidence intervals, are what people really want.
- Coherent: no ancillarity issues; satisfies likelihood principle
- Prescriptive: the prescription is Bayes Theorem

Invalid weaknesses of Bayes

- Bayes too subjective for scientific inference
  - Bayesian approach can encompass a full gamut of subjectivity, depending on strength of data and prior
  - Frequentist methods (under explicit or implicit models) often involve major subjective elements
  - For example, regression coefficients of excluded covariates are implicitly assigned prior distributions with all prior probability at zero.
- Bayes denies the role of randomization for design
  - Randomization is vital for credible Bayesian inferences if selection / allocation mechanism is included in the model (e.g. chapter 7 of Gelman, Carlin, Stern, and Rubin 2003)

Valid weaknesses of Bayes

- Requires and relies on full specification of a model (likelihood and prior)
  - Where does the model come from?
  - "Too many answers", corresponding to all the possible choices of model/prior
  - Models are always wrong, and bad models lead to bad answers; no built in "calibration"
  - Unclear how to incorporate uncertainty from misspecification of models – tends to be informal, at best. (To some extent this applies to frequentist methods as well.)

Bayesian model formulation

- B is less effective for model formulation and assessment than for inference under a model.
- For example, Bayesian hypothesis testing for comparing models of different dimension is tricky
  - sensitive to choice of priors; can't just slap down a reference prior
  - Strict subjective Bayesians claim they can make pure Bayesian model selection work, but this approach is a hard sell for scientific inference
  - Most use the data for model selection, in some form
  - Model formulation and assessment will never achieve the degree of clarity of Bayesian inference under an agreed model

Summary

Activity	Bayes	Frequentist
Inference under assumed model	Strong	Weak
Model formulation / assessment	Weak	Strong

Conclusion: Calibrated Bayes

Activity	Bayes	Frequentist
Inference under assumed model	Strong	
Model formulation / assessment		Strong

Bayesian for inference

Frequentist for model assessment (enriched by Bayesian ideas)

Capitalizes on strengths of both paradigms!

Calibrated Bayes

- Calibrated Bayes ideas go back to the 1960's (see e.g. Peers (1965); Welch (1965); Dawid (1982))
- Two landmark papers by George Box (1980) and Don Rubin (1984) discuss the approach in some generality

Bayes/frequentist compromises

"I believe that … sampling theory is needed for exploration and ultimate *criticism* of the entertained model in the light of the current data, while Bayes' theory is needed for *estimation* of parameters conditional on adequacy of the model."

George Box (1980)

Bayes/frequentist compromises

- Box's calibrated Bayes factorization:

$$p(Y,\theta \mid M) = p(Y \mid M)\, p(\theta \mid Y, M)$$

Model criticism Parameter inference

Prior predictive checks: compare $d(Y_{\text{obs}})$ with distribution of d given M. d = discrepancy (Gelman, Meng and Stern 1996)

Bayes/frequentist compromises

"The applied statistician should be Bayesian in principle and calibrated to the real world in practice – appropriate frequency calculations help to define such a tie."

"… frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

Don Rubin (1984)

Bayes/frequentist compromises

- Rubin's calibrated Bayes factorization:

$$p(Y^*, \theta^*, \theta \mid Y, M) = p(Y^*, \theta^* \mid Y, \theta, M)\, p(\theta \mid Y, M)$$

Model criticism (integrating out θ)

Parameter inference

Posterior predictive checks: compare $d(Y_{\text{obs}}, \theta)$ with distribution of $d(Y^*, \theta^*)$ given Y and M. d = discrepancy (Gelman, Meng and Stern 1996)

Posterior predictive checks seem preferable to prior predictive checks, since they focus on validity of predictions and avoid sensitivity to choice of the prior

Advantages of calibrated Bayes

- Retains optimal properties of Bayes inference under well-specified models
- Focus on frequentist calibration creates useful resistance to “bad models”
  - E.g. in surveys, “design-consistency” forces models that account for survey design (Hansen, Madow and Tepping 1983)
- Limits ambiguities of frequentist assessments (e.g degree of conditioning) to evaluation of model, rather than model inference itself

Advantages of calibrated Bayes

- Assists in selection of “reference priors”
- Fisherian significance tests still have role for model checking
  - Global tests of null that a model is consistent with data, avoiding the need for specifying an alternative model
  - E.g. testing “no linkage” in genetics
  - Posterior predictive checks greatly expand range of model assessments over frequentist approaches
- Neyman-Pearson hypothesis testing does not have a role for inference about model parameters
  - No great loss, in my view …
  - See Christensen (2005)

Problems with Calibrated Bayes

- What is it?? The inference under model is prescriptive, but not the model formulation
- Ambiguities at the frontier between model inference and model checking
  - How much peeking at the data is allowed in developing the model without corrupting the inference?
  - Model selection vs. model averaging (Draper 1995)
- Choice of checks is often unclear
  - No prescription here
  - Should posterior predictive P-Values be uniform under the posited model? (Bayarri and Berger 2000, Robins, van der Vaart and Ventura 2000)
- How to assess uncertainty in model misspecification

Consequences for teaching

- Bayesian statistical inference needs to be taught!
  - Bayesian statistics is “optional” or even absent in many programs for training MS and even Ph.D. statisticians
  - unsupportable given prominence of Bayes in science (e.g. 2002 Science Watch citations of mathematicians)
  - consumers of statistics learn nothing at all about Bayes
- Barriers to teaching Bayes to non-mathematicians are overrated!
  - Basic idea of Bayes Theorem does not require calculus
  - Focus on interpretation of answers rather than details of Bayesian calculations
  - Frequentist theory is no picnic to teach to consumers!

Consequences continued

- More emphasis on statistical modeling over methods
  - All models are wrong – how do we pick a good one?
  - Formulating statistical models for real data is not simple, e.g.
  - Models with better fits can yield worse answers (e.g Heckman selection models for missing data)
  - All model assumptions are not equal, e.g. how do the assumptions rank in importance?
  - Difficulties of picking priors in high-dimensional complex models, objective or subjective
  - Students need more instruction on how to fit models to complicated data sets

Consequences ctd.

- More attention is needed to assessments of model fit
  - Models are wrong, need careful checking
  - Frequentist methods have a role here
  - Includes hypothesis tests with sharp nulls
  - Diagnostics well known for regression, less developed and taught for other models

Summary

- Bayes and frequentist ideas are both important for good statistical inference
- Both sets of ideas should be taught
- The calibrated Bayes compromise capitalizes on strengths of Bayes and frequentist paradigms
  - A good roadmap for the 21st century

References

Basu, D (1971) p203-242, *Foundations of Statistical Inference*, Holt, Rinehart and Winston.
Bayarri, M.J. and Berger, J.O. (2000), *JASA* 95, 1127-1172.
Berger, J.M. (2000), *JASA* 95, 1269-1276.
Bernado, JM (1979) *JRSSB* 41, 113-147.
Birnbaum, A (1962) *JASA* 57, 269-326.
Box, GEP (1980) *JRSSA* 143, 383-430.
Breiman, L. (2001) Statist. Sci. 16, 199–231.
Dawid, A.P. (1982), *JASA* 77, 605-610.
Dawid, AP., Stone, M. & Zidek, JV (1973) *JRSSB* 35, 189-233.
Draper D (1995) *JRSSB*, 57, 45--97
Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003), *Bayesian Data Analysis*, 2nd. Ed. CRC Press
Gelman, A., Meng, X.-L. and Stern, H. (1996) *Statist. Sinica* 6, 733-807.
Hansen, MH, Madow, WG & Tepping, BJ (1983) *JASA* 78, 776-793.
Little, RJA (1989) *The American Statistician,* 43, 283-288.
Little, RJA (2004) *JASA* 99, 546-556.
Peers, H.W. (1965 *JRSSB* 27, 9-16.
Robins, J.M., van der Vaart, A., and Ventura, V. (2000) *JASA* 95, 1143-1172.
Rubin, DB (1984) *Annals of Statistics* 12, 1151-1172.
Welch, B.L. (1965) *JRSSB* 27, 1-8.
Yates, F (1984) *JRSSA* 147, 426-463.