Probability and Frequency

(Lecture 3)

Tom Loredo

Dept. of Astronomy, Cornell University

`http://www.astro.cornell.edu/staff/loredo/bayes/`

The Frequentist Outlook

Probabilities for hypotheses are meaningless because hypotheses are not “random variables.”

Data *are* random, so only probabilities for data can appear in calculations.

Strength of inference is cast in terms of long-run behavior of procedures, averaged over data realizations:

- How far is $\hat{\theta}(D)$ from true θ, on average (over D)?
- How often does interval $\Delta(D)$ contain true θ, on average?
- How often am I wrong if I reject a model when $S(D)$ is above S_c?

What is good for the long run
is good for the case at hand.

The Bayesian Outlook

Quantify information about the case at hand as completely and consistently as possible.

No explicit regard is given to long run performance.

But a result that claims to fully use the information in each case should behave well in the long run.

What does the case at hand tell us about what might occur in the long run?

Is what is good for the case at hand also good for the long run?

Lecture 3

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures

Lecture 3

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures

Prediction and Inference w/ Frequencies

Frequencies are relevant when modeling repeated trials, or repeated sampling from a population or ensemble.

Frequencies are *observables:*

- When available, can be used to *infer* probabilities for next trial
- When unavailable, can be *predicted*

Some Frequency Models

Consider binary experiments. Trial t produces result r_t (0 or 1), with probability a that may be known or unknown.

Frequency of 1's in N trials:

$$f = \frac{1}{N} \sum_t r_t = \frac{n}{N}$$

M_1: independent trials, a a known constant; predict f
M_2: a is an unknown constant; f is observed; infer a
M_3: $p(r_1, r_2 \ldots r_N | M_3)$ known (dependence!); predict f
M_4: Parallel experiments on similar systems produce $\{f_i\}$; infer $\{a_i\}$

Independent Trials

M_1: a is a known constant; predict f

Use the binomial dist'n: $f = a \pm \sqrt{a(1-a)/N}$

Special case of (weak) law of large numbers

M_2: a is an unknown constant; f is observed; infer a

Our binary outcome example from Lecture 1—the first use of Bayes's theorem: $a = f \pm \sqrt{n}/N$

Dependent Trials

M_3: $p(r_1, r_2 \ldots r_N|M_3)$ known; predict f

Can show that:

$$\langle f \rangle = \frac{1}{N} \sum_t p(r_t|M_3)$$

where $$p(r_1|M_3) = \sum_{r_2} \cdots \sum_{r_N} p(r_1, r_2 \ldots |M_3)$$

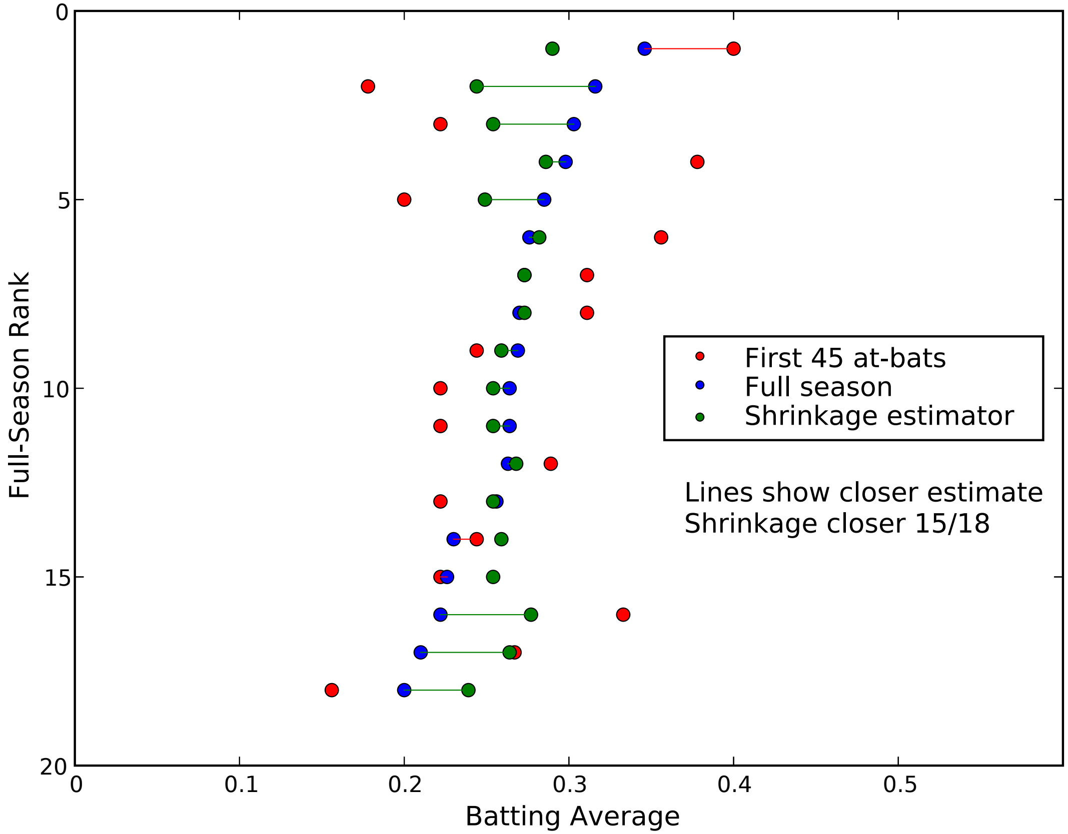
Expected frequency of outcome in many trials $=$ *average* probability for outcome across trials.

But can also show that σ_f needn't converge to 0. The actual frequency may differ significantly from its expectation even after many trials.

Population of Related Systems

M_4: Parallel experiments on similar systems produce $\{f_i\}$; infer $\{a_i\}$

Example: 1977 Batting Averages (Efron & Morris)



Green estimates are deliberately *biased* from observed frequencies—and predict the future better! (“Shrinkage”)

Probability and Frequency

Probabilities and frequencies in repeated experiments are simply related only in the simplest settings (independence, small dimension).

Otherwise, the relationships are subtle. A formalism that distinguishes them from the outset is particularly valuable for exploring this. E.g., shrinkage is explored via hierarchical and empirical Bayes.

Lecture 3

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures

Bayesian Calibration

Credible region $\Delta(D)$ with probability P:

$$P = \int_{\Delta(D)} d\theta \; p(\theta|I) \frac{p(D|\theta, I)}{p(D|I)}$$

What fraction of the time, Q, will the true θ be in $\Delta(D)$?

1. Draw θ from $p(\theta|I)$
2. Simulate data from $p(D|\theta, I)$
3. Calculate $\Delta(D)$ and see if $\theta \in \Delta(D)$

$$Q = \int d\theta \; p(\theta|I) \int dD \; p(D|\theta, I) \, [\theta \in \Delta(D)]$$

$$Q = \int d\theta \; p(\theta|I) \int dD \; p(D|\theta, I) \, [\theta \in \Delta(D)]$$

Note appearance of $p(\theta, D|I) = p(\theta|D, I)p(D|I)$:

$$\begin{aligned}
Q &= \int dD \int d\theta \; p(\theta|D, I) \, p(D|I) \, [\theta \in \Delta(D)] \\
&= \int dD \; p(D|I) \; \int_{\Delta(D)} d\theta \; p(\theta|D, I) \\
&= P \int dD \; p(D|I) \\
&= P
\end{aligned}$$

Bayesian inferences are "calibrated."
Calibration is with respect to choice of prior & $\mathcal{L}$.
This is useful for testing Bayesian computer codes.

Frequentist Coverage and Confidence

Coverage:

Coverage for a rule $\delta(D)$ specifying a parameter interval based on the data:

$$C_\delta(\theta) \quad = \quad \int dD\, p(D|\theta, I)\, [\theta \in \delta(D)]$$

If $C(\theta) = P$, a *constant*, $\delta(D)$ is a *strict confidence region* with confidence level P.

Conservative confidence regions:

It is hard to find $\delta(D)$ giving constant $C(\theta)$; very hard with nuisance parameters, and impossible with discrete data.

Reported confidence level $\equiv \min_\theta C_\delta(\theta)$.

This remains problematic for discrete data. E.g., binomial dist'n: If $a = 0$, then $n = 0$, always. Any $\delta(n)$ will just give one particular interval, $\delta(0)$, for all trials and thus must have $C(0) = 0$ or 1.

Average coverage:

Intuition suggests reporting some kind of average performance: $\int d\theta\ f(\theta) C_\delta(\theta)$

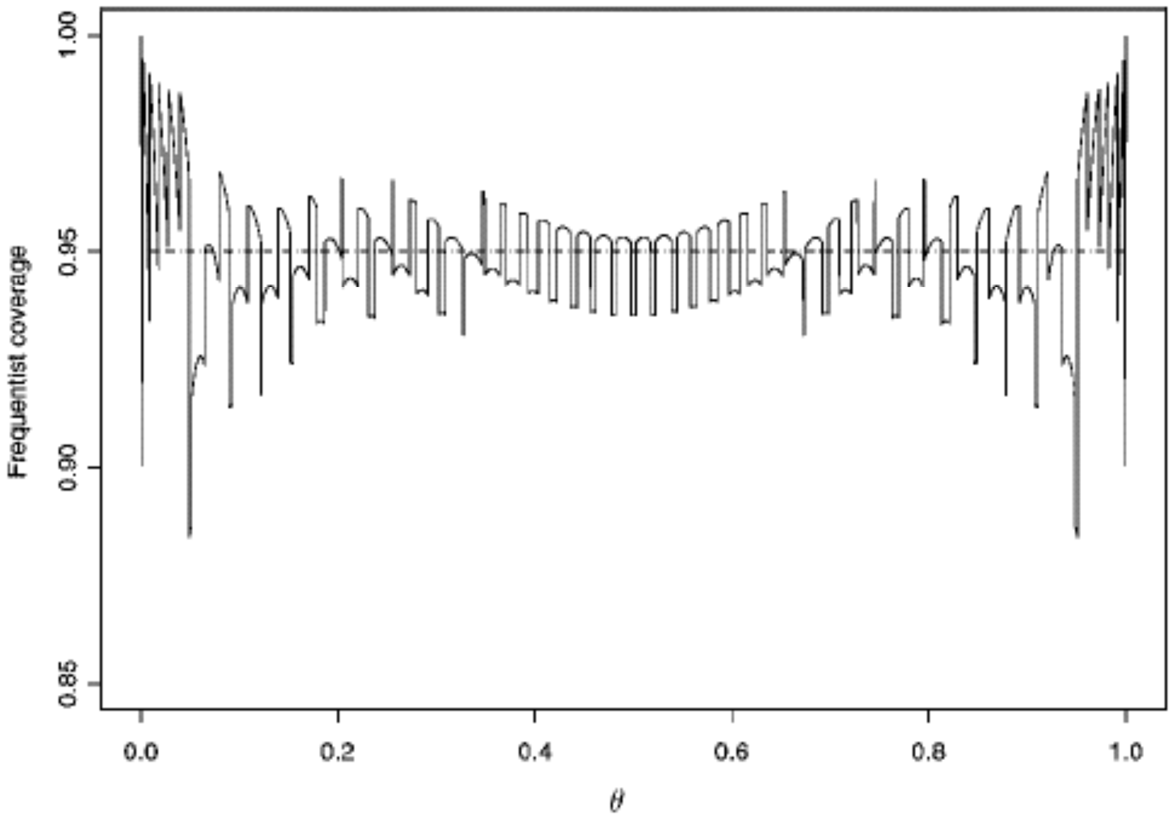
Recall the Bayesian calibration condition:

$$
\begin{aligned}
P &= \int d\theta\, p(\theta|I) \int dD\, p(D|\theta, I)\, [\theta \in \Delta(D)] \\
&= \int d\theta\, p(\theta|I)\, C_\delta(\theta)
\end{aligned}
$$

provided we take $\delta(D) = \Delta(D)$.

- If $C_\Delta(\theta) = P$, the credible region is a strict confidence region.
- Otherwise, the credible region's probability content accounts for a priori uncertainty in θ, via the prior.

Coverage for Binomial Estimation



Binomial CR coverage, $N = 50$

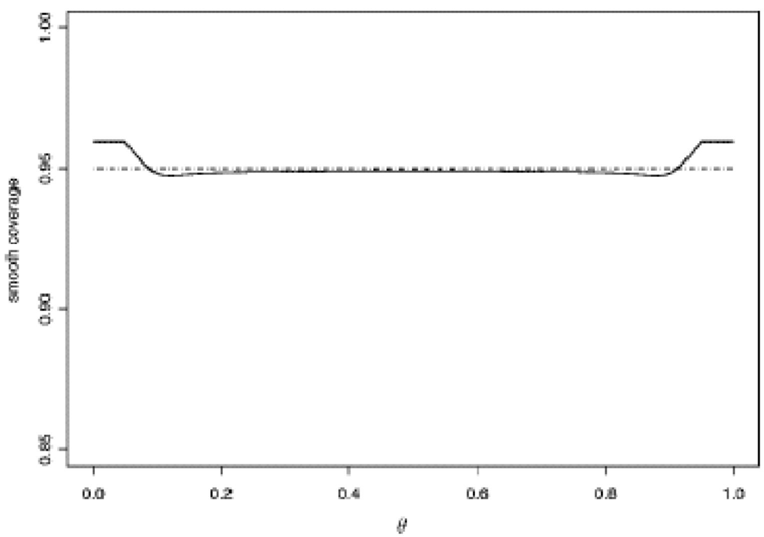
Berger & Bayarri 2004

But the locations and sizes of the “jitter” vary with N.

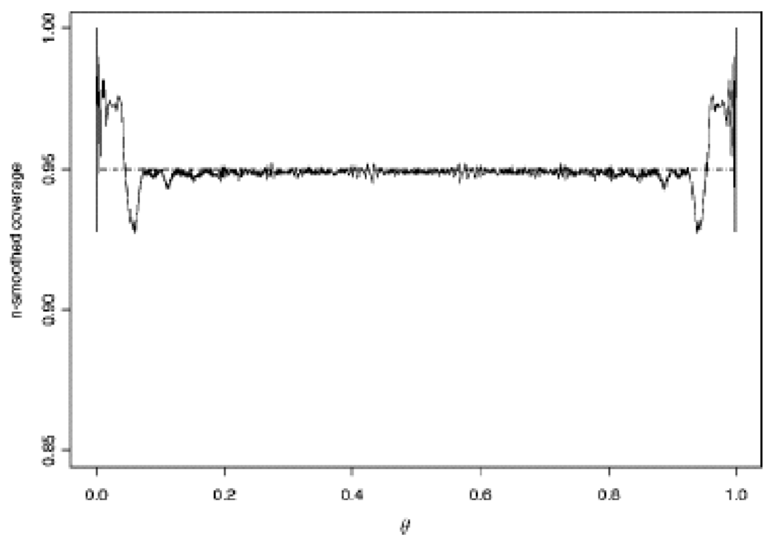
Parameter & Sample Averaged Coverage

It may be more relevant to report coverage for situations “like” the observed one, but not identical to it—nearby parameter values, or similar sample size. $\rightarrow$ average coverage is relevant:

Avg. over nearby θ



Avg. over similar N



Berger & Bayarri

The actual uncertainties in real situations suggest *some* kind of averaging is more relevant, and that conservative coverage is *too* conservative.

Calibration for Bayesian Model Comparison

Assign prior probabilities to N_M different models.

Choose as the true model that with the highest posterior probability, but only if the probability exceeds $P_{\rm crit}$.

Iterate via Monte Carlo:

- 1.Choose a model by sampling from the model prior.
- 2.Choose parameters for that model by sampling from the parameter prior *pdf*.
- 3.Sample data from that model's sampling distribution conditioned on the chosen parameters.
- 4.Calculate the posteriors for all the models; choose the most probable if its $P > P_{\rm crit}$.

$\Rightarrow$ Will be correct $\geq 100 P_{\rm crit}$% of the time that we reach a conclusion in the Monte Carlo experiment.

Robustness to model prior:

What if model frequencies $\neq$ model priors?

Choose between two models based on the Bayes factor, B (assumes equal freq.), but let them occur with *nonequal* frequencies, f_1 and f_2. Let γ be the max prior freq. ratio for a model:

$$\gamma = \max\left[\frac{f_1}{f_2}, \frac{f_2}{f_1}\right]$$

Fraction of time a correct conclusion is made if we require $B > B_{\rm crit}$ or $B < 1/B_{\rm crit}$ is

$$Q > \frac{1}{1 + \frac{\gamma}{B_{\rm crit}}}$$

E.g., if $B_{\rm crit} = 100$:

- Correct $\geq 99\%$ if $\gamma = 1$
- Correct $\geq 91\%$ if $\gamma = 9$

A Worry: Incorrect Models

What if none of the models is “true”?

Comfort from experience: Rarely are statistical models precisely true, yet standard models have proved themselves adequate in applications.

Comfort from probabilists: Studies of consistency in the framework of nonparametric Bayesian inference show “good priors are those that are approximately right for most densities; parametric priors [e.g., histograms] are often good enough” (Lavine 1994). But there remains some controversy about this; if “big” models are required to fit the data, expert care is required.

One should worry somewhat, but there is not yet any theory providing a consistent, quantitative “model failure alert” (Bayesian or frequentist).

Bayesian Consistency & Convergence

Parameter Estimation:

- Estimates are consistent if the prior doesn't exclude the true value.
- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$.
- Using standard nonuniform "reference" priors can improve their performance to $O(n^{-1})$.
- For handling nuisance parameters, regions based on marginal likelihoods have superior long-run performance to regions found with conventional frequentist methods like profile likelihood. Competitive frequentist methods require conditioning on ancillaries and correction factors that mimic marginalization.

Model Comparison:

- Model comparison is asymptotically consistent. Popular frequentist procedures (e.g., χ^2 test, asymptotic likelihood ratio ($\Delta\chi^2$), AIC) are not.
- For separate (not nested) models, the posterior probability for the true model converges to 1 exponentially quickly.
- When selecting between more than 2 models, carrying out multiple frequentist significance tests can give misleading results. Bayes factors continue to function well.

Summary

Parametric Bayesian methods are typically excellent frequentist methods!

Not too surprising—methods that claim to be optimal for each individual case should be good in the long run, too.

Key Ideas

- Connections between probability and frequency can be subtle
- Bayesian results are calibrated (w.r.t. modeling assumptions)
- Parametric Bayesian methods are good frequentist methods