Uncertainty Quantification Qualification

Lawrence Livermore National Laboratory
26 March 2009

Philip B. Stark
Department of Statistics
University of California, Berkeley
statistics.berkeley.edu/~stark

Parts joint with D.A. Freedman, L. Tenorio

Abstract

Uncertainty quantification (UQ) estimates the uncertainty in models of physical systems calibrated to noisy data. Uncertainty quantification *qualification* (UQQ) is needed. What sources of uncertainty does a UQ analysis take into account? What does it ignore? How ignorable are the ignored sources of uncertainty? What assumptions does the UQ estimate depend on? What evidence supports those assumptions? Are the assumptions testable? What happens if the assumptions are false?

The 2009 NAS report, *Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile* (EQMU), has some good advice. But M/U has little connection to "confidence." EQMU is inconsistent and confuses absence of evidence with evidence of absence. Probabilistic Risk Assessment as proposed is not a serious solution. "UQ by committee" has difficulties evident in the USGS earthquake forecast for the Bay Area.

Outline

- 2009 NAS Report (EQMU): digs, cream, margins, uncertainties, confidence, howlers, constraints versus priors.

- Earthquake probabilities: The USGS forecast for the SF Bay Area.

- History lessons: EOS/helioseismology; solar free oscillations.

- A (partial) catalog of sources of uncertainty.

Present company excluded, of course!

Axiom 1: Anything that comes up in a physics problem is physics.

Axiom 2: Nobody knows more about physics than physicists.*

Theorem: There's no reason for physicists to talk to anybody else to solve physics problems.

Practical consequence: Physicists often re-invent the wheel. It is not always as good as the wheel a mechanic would build.

Some "unsolved" problems–according to EQMU—are solved. But not by physicists.

Who was on the NAS panel?

*Special case of the general axiom: Nobody knows more about *anything* than physicists, and physicists are smarter than everybody else.

What is UQ?

UQ = inverse problems + approximate forward model.

Lots of work on this, including approximate forward model.

Cream of EQMU (p. 25)

Assessment of the accuracy of a computational prediction depends on assessment of model error, which is the difference between the laws of nature and the mathematical equations that are used to model them. Comparison against experiment is the only way to quantify model error and is the only connection between a simulation and reality. . . .

Even if model error can be quantified for a given set of experimental measurements, it is difficult to draw justifiable broad conclusions from the comparison of a finite set of simulations and measurements. . . . it is not clear how to estimate the accuracy of a simulated quantity of interest for an experiment that has not yet been done. . . . In the end there are inherent limits [which] might arise from the paucity of underground nuclear data and the circularity of doing sensitivity studies using the same codes that are to be improved in ways guided by the sensitivity studies.

Example from EQMU (pp. 9–11, 25–6; notation changed)

Device needs voltage V_T to detonate. Detonator applies V_A. "Boom" if $V_A \geq V_T$.

V_T estimated as $\widehat{V}_T = 100V$, with uncertainty $U_T = 5V$.

V_A estimated as $\widehat{V}_A = 150V$, with uncertainty $U_A = 10V$.

Margin $M = 150V - 100V = 50V$.

Total uncertainty $U = U_A + U_T = 10V + 5V = 15V$.

"Confidence ratio" $M/U = 50/15 = 3\frac{1}{3}$.
Magic ratio $M/U = 3$. (EQMU, p. 46)

"If $M/U >> 1$, the degree of confidence that the system will perform as expected should be high. If M/U is not significantly greater than 1, the system needs careful examination." (EQMU, p. 14)

Scratching the veneer.

Are V_A and/or V_T random? Or simply unknown?

Are $\widehat{V}_A$ and $\widehat{V}_T$ design parameters? Estimates from data?

Why should U_A and U_T add to give total uncertainty U?

How well are U_A and U_T known?

If U is a bound on the possible error, then have complete confidence if $M > U$: ratio doesn't matter.

If U isn't a bound, what does U mean?

EQMU says:

"Generally [uncertainties] are described by probability distribution functions, not by a simple band of values."
(EQMU, p. 13)

"An important aspect of [UQ] is to calculate the (output) probability distribution of a given metric and from that distribution to estimate the uncertainty of that metric. The meaning of the confidence ratio (M/U) depends significantly on this definition . . ."
(EQMU, p. 15)

Vision 1: Us are error bars

Suppose V_A and V_T are independent random variables* with known means $\hat{V}_A$ and $\hat{V}_T$, respectively.

Suppose $\mathbb{P}\{\hat{V}_A - V_A \le U_A\} = 90\%$ and $\mathbb{P}\{V_T - \hat{V}_T \le U_T\} = 90\%$.

What's $\mathbb{P}\{V_A - V_T \ge 0\}$? Can't say, but . . .

Bonferroni's inequality:

$$\mathbb{P}\{\hat{V}_A - V_A \le U_A \text{ and } V_T - \hat{V}_T \le U_T\} \ge 80\%.$$

That's a conservative bound. What's the right answer?

*Are they random variables? If so, why not dependent?

Vision 2: Us are (multiples of) SDs

"...if one knows the type of distribution, it could be very helpful to quantify uncertainties in terms of standard deviations. This approach facilitates meaningful quantitative statements about the likelihood of successful functioning." (EQMU, p. 27)

Does one ever know the type of distribution? Is the SD known to be finite? Can very long tails be ruled out?

Even if so, that's not enough: what's the joint distribution of V_A and V_T?

If V_A and V_T were independent with means $\widehat{V}_A$ and $\widehat{V}_T$ and SDs U_A and U_T, the SD of $V_A - V_T$ would be $\sqrt{U_A^2 + U_T^2}$, not $U_A + U_T$.

If Us are multiples of SDs, what's the confidence?

Suppose $U = SD(V_A - V_T)$.
What does $M/U = k$ imply about $\mathbb{P}\{V_A > V_T\}$?

Chebychev's inequality:

$$\mathbb{P}\left\{|V_A - V_B - (\hat{V}_A - \hat{V}_B)| \le kU\right\} \ge 1 - \frac{1}{k^2}.$$

E.g., $k = 3$ gives "confidence" $1 - 1/9 = 88.9\%$.

C.f. typical Gaussian assumption: $k = 3$ gives "confidence"

$$\mathbb{P}\left\{\frac{V_A - V_B - (\hat{V}_A - \hat{V}_B)}{\sigma(V_A - V_T)} \ge 3\right\} \approx 99.9\%.$$

$$88.9\% < 99.9\% < 100\%.$$

Vision 3: one of each

From the description, makes sense that V_T is an unknown parameter, $\widehat{V}_T$ is an already-computed estimate of V_T from data, $\widehat{V}_A$ is a design parameter, and V_A is a random variable that will be "realized" when the button is pushed.

If so, makes sense that U_T is an "error bar" computed from data.

Either $V_T - \widehat{V}_T \leq U_T$ or not: no probability left, only ignorance.

Whether $\widehat{V}_A - V_A \leq U_A$ is still a random event; depends on what happens when the button is pushed.

EQMU is careless about what is known, what is estimated, what is uncertain, what is random, etc.
The "toy" lead example is problematic.

Historical error bars

How to make sense of error bars on historical data? Crucial!

Seldom know how the bars were constructed or what they were intended to represent.

Variability in repeated experiments?

Spatial variability (e.g., across-channel variation) within a single experiment?

Instrumental limitation or measurement error?

Hunch? Wish? Prayer? Knee-jerk "it's 10%?"

Measuring apparatus retired along with institutional memory. Can't repeat experiments.

Good quote (EQMU, p. 27, fn 5)

"To the extent (which is considerable) that input uncertainties are epistemic and that probability distribution functions (PDFs) cannot be applied to them, uncertainties in output/integral parameters cannot be described by PDFs."

And then gibberish ensues.

Bad quotes (EQMU, p. 21)

"Given sufficient computational resources, the labs can sample from input-parameter distributions to create output-quantity distributions that quantify code sensitivity to input variations."

"Sampling from the actual high-dimensional input space is not a solved problem."

" . . . the machinery does not exist to propagate [discretization errors] and estimate the uncertainties that they generate in output quantities."

Fallacy (EQMU, p. 23)

"Analysis shows that 90 percent of the realistic input space (describing possible values of nature's constants) maps to acceptable performance, while 10 percent maps to failure. This 90 percent is a confidence number . . . we have a 90 percent confidence that all devices will meet requirements and a 10 percent confidence that all will fail to meet requirements."

Laplace's principle of insufficient reason: if there's no reason to think possibilities have different probabilities, assume that the probabilities are equal.

No evidence of difference $\neq$ evidence of no difference.

Example: Gas thermodynamics

Gas of of n non-interacting particles. Each can be in any of r quantum states; possible values of "state vector" equally likely.

1. Maxwell-Boltzman. State vector gives the quantum state of each particle: r^n possible values.

2. Bose-Einstein. State vector gives # particles in each quantum state: $\binom{n+r-1}{n}$ possible values.

3. Fermi-Dirac. State vector gives the # particles in each quantum state, but no two particles can be in the same state: $\binom{r}{n}$ possible values.

Gas thermodynamics, contd.

Maxwell-Boltzman common in probability theory (e.g., "coin gas"), but but describe no known gas.

Bose-Einstein describes bosons, e.g., photons and He^4 atoms.

Fermi-Dirac describes fermions, e.g., electrons and He^3 atoms.

Outcomes can be defined or parametrized in many ways. Not clear which–if any–give equal probabilities.

Principle of Insufficient Reason is insufficient for physics.

Constraints versus prior probabilities

Bayesian machinery (LANL approach) is appealing but can be misleading.

Capturing constraints using priors adds "information" not present in the constraints.

- Why a particular form?
- Why particular values of the parameters?
- What's the relation between the "error bars" the prior represents and specific choices?

Distributions on states of nature

Bayes' Rule: $\mathbb{P}(B|A) = \mathbb{P}(A|B)\mathbb{P}(B)/\mathbb{P}(A)$.

"Just math."

To have posterior $\mathbb{P}(B|A)$, need prior $\mathbb{P}(B)$.

The prior matters. Where does it come from?

Misinterpretation of LLNL "ensemble of models" approach to UQ: no prior.

Conservation of Rabbits

> Freedman's Principle of Conservation of Rabbits:
> To pull a rabbit from a hat, a rabbit must first be placed in the hat.

The prior puts the rabbit in the hat.

PRA puts many rabbits in the hat.

Bayes/minimax duality: minimax uncertainty is Bayes uncertainty for least favorable prior.*

*Least favorable $\neq$ "uninformative."

Bounded normal mean

Know that $\theta \in [-\tau, \tau]$.

Observe $Y = \theta + Z$.

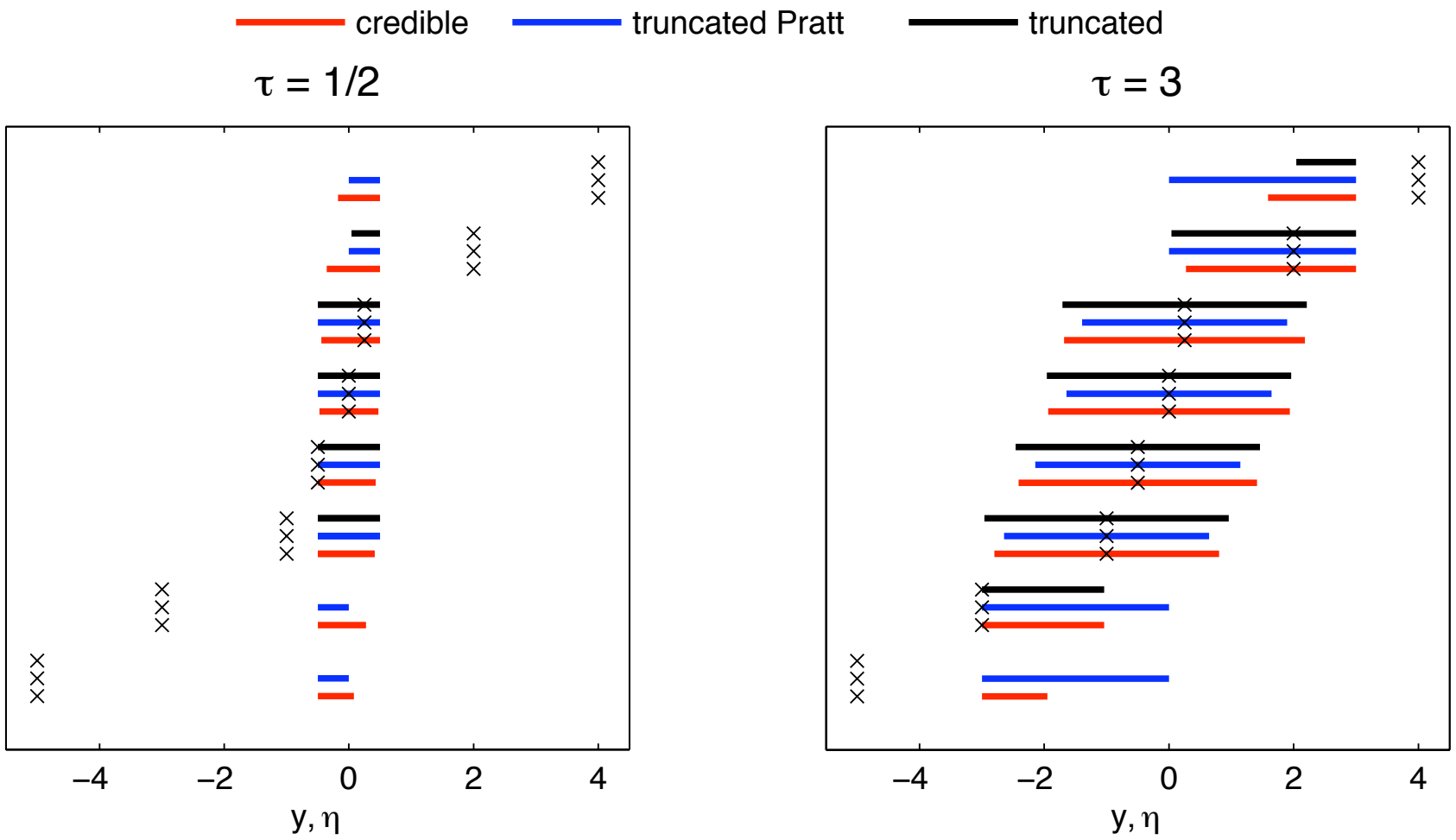
$Z \sim N(0, 1)$.

Want to estimate θ.

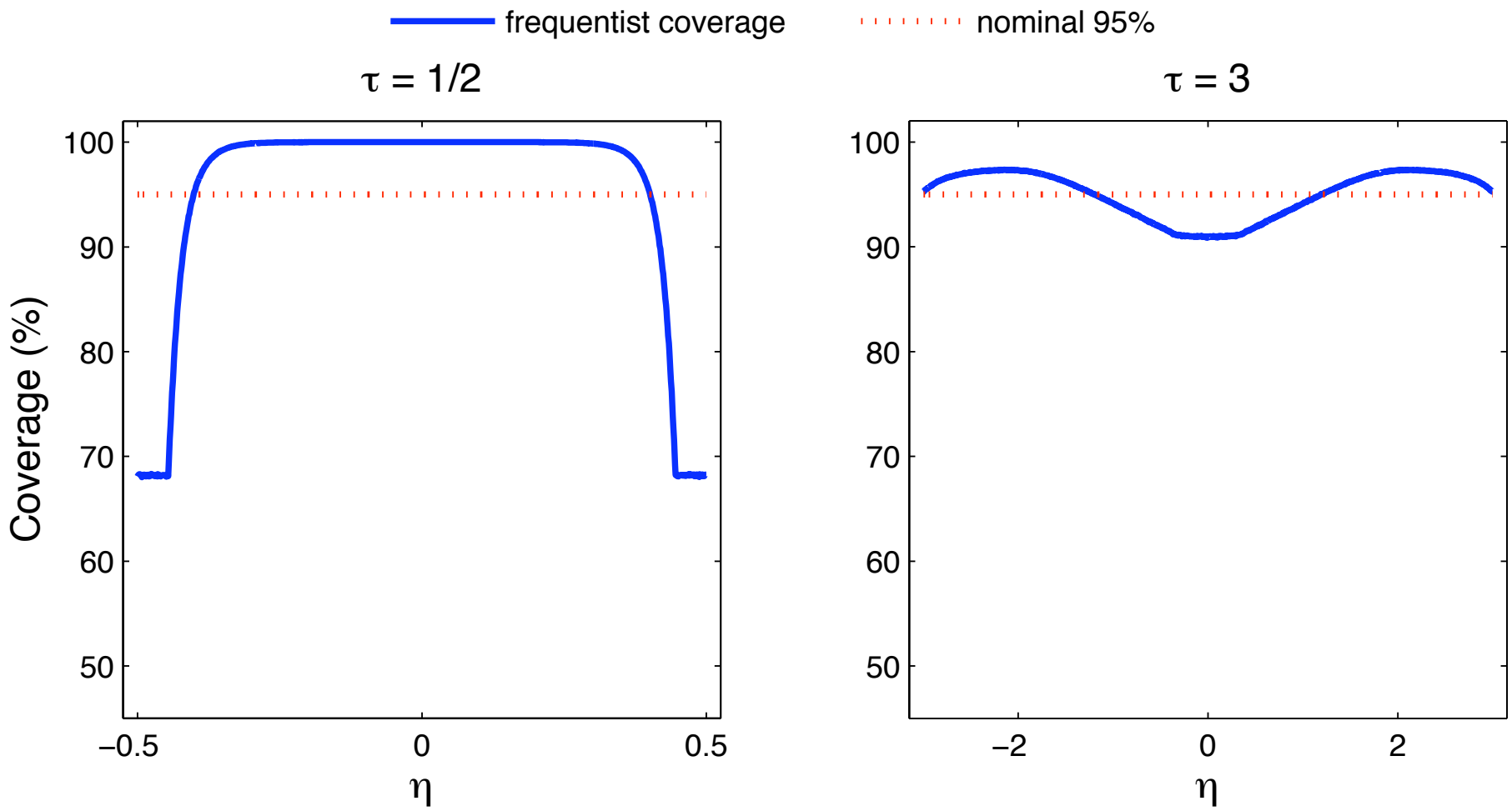
Bayes: capture constraint using prior, e.g., $\theta \sim U[-\tau, \tau]$.

Credible region: 95% posterior probability.
Confidence interval: 95% chance before data are collected.

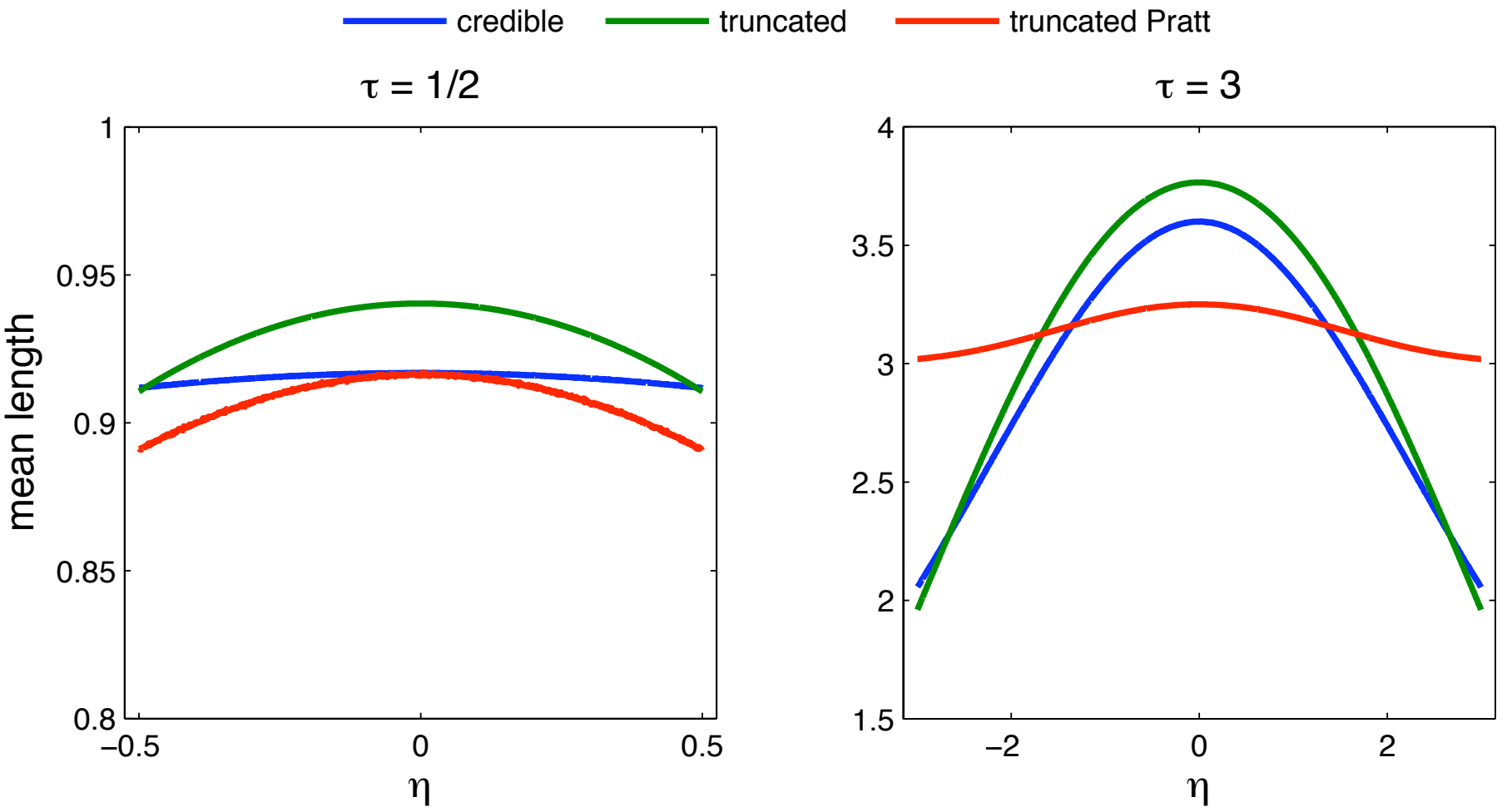
95% Confidence sets vs. credible regions



Coverage of 95% credible regions



Expected size of credible regions and confidence intervals



Interpreting earthquake predictions (joint with D.A. Freedman)

Globally, on the order of 1 magnitude 8 earthquake per year.

Locally, recurrence times for big events $O(100$ y$)$.

Big quakes deadly and expensive.

Much funding and glory in promise of prediction.

Would be nice if prediction worked.

Some stochastic models for seismicity:

- Poisson (spatially heterogeneous; temporally homogeneous; marked?)
- Gamma renewal processes
- Weibull, lognormal, normal, double exponential, ...
- ETAS
- Brownian passage time

Coin Tosses. What does $P(\text{heads}) = 1/2$ mean?

- Equally likely outcomes: Nature indifferent; principle of insufficient reason
- Frequency theory: long-term limiting relative frequency
- Subjective theory: strength of belief
- Probability models: property of math model; testable predictions

Math coins $\neq$ real coins.

Weather predictions: look at sets of assignments. Scoring rules.

USGS 1999 Forecast

$$P(\text{M}\geq 6.7 \text{ event by } 2030) = 0.7 \pm 0.1$$

What does this mean?

Where does the number come from?

Two big stages.

Stage 1

1. Determine regional constraints on aggregate fault motions from geodetic measurements.

2. Map faults and fault segments; identify segments with slip ≥ 1 mm/y. Estimate the slip on each fault segment principally from paleoseismic data, occasionally augmented by geodetic and other data. Determine (by expert opinion) for each segment a 'slip factor,' the extent to which long-term slip on the segment is accommodated aseismically. Represent uncertainty in fault segment lengths, widths, and slip factors as independent Gaussian random variables with mean 0. Draw a set of fault segment dimensions and slip factors at random from that probability distribution.

3. Identify (by expert opinion) ways segments of each fault can rupture separately and together. Each combination of segments is a 'seismic source.'

4. Determine (by expert opinion) extent that long-term fault slip is accommodated by rupture of each combination of segments for each fault.

5. Choose at random (with probabilities of 0.2, 0.2, and 0.6) 1 of 3 generic relationships between fault area and moment release to characterize magnitudes of events that each combination of fault segments supports. Represent the uncertainty in generic relationship as Gaussian with zero mean and standard deviation 0.12, independent of fault area.

6. Using the chosen relationship and the assumed probability distribution for its parameters, determine a mean event magnitude for each seismic source by Monte Carlo.

7. Combine seismic sources along each fault 'to honor their relative likelihood as specified by the expert groups;' adjust relative frequencies of events on each source so every fault segment matches its estimated slip rate. Discard combinations of sources that violate a regional slip constraint.

8. Repeat until 2,000 regional models meet the slip constraint. Treat the 2,000 models as equally likely for estimating magnitudes, rates, and uncertainties.

9. Estimate background rate of seismicity: Use an (unspecified) Bayesian procedure to categorize historical events from three catalogs either as associated or not associated with the seven fault systems. Fit generic Gutenberg-Richter magnitude-frequency relation $N(M) = 10^{a-bM}$ to events deemed not to be associated with the seven fault systems. Model background seismicity as a marked Poisson process. Extrapolate Poisson model to $M \geq 6.7$, which gives a probability of 0.09 of at least one event.

Stage 1: Generate 2,000 models; estimate long-term seismicity rates as a function of magnitude for each seismic source.

Stage 2:

1. Fit 3 stochastic models for earthquake recurrence—Poisson, Brownian passage time and 'time-predictable'—to long-term seismicity rates estimated in stage 1.

2. Combine stochastic models to estimate chance of a large earthquake.

Poisson and Brownian passage time models used to estimate the probability an earthquake will rupture each fault segment.

Some parameters fitted to data; some set more arbitrarily. Aperiodicity (standard deviation of recurrence time, divided by expected recurrence time) set to three different values, 0.3, 0.5, and 0.7. Method needs estimated date of last rupture of each segment.

Model redistribution of stress by earthquakes; predictions made w/ & w/o adjustments for stress redistribution.

Predictions for segments combined into predictions for each fault using expert opinion about the relative likelihoods of different rupture sources.

'Time-predictable model' (stress from tectonic loading needs to reach the level at which the segment ruptured in the previous event for the segment to initiate a new event) used to estimate the probability that an earthquake will originate on each fault segment. Estimating the state of stress before the last event requires date of the last event and slip during the last event. Those data are available only for the 1906 earthquake on the San Andreas Fault and the 1868 earthquake on the southern segment of the Hayward Fault. Time-predictable model could not be used for many Bay Area fault segments.

Need to know loading of the fault over time; relies on viscoelastic models of regional geological structure. Stress drops and loading rates modeled probabilistically; the form of the probability models not given. Loading of San Andreas fault by the 1989 Loma Prieta earthquake and the loading of Hayward fault by the 1906 earthquake were modeled.

Probabilities estimated using time-predictable model were converted into forecasts using expert opinion for relative likelihoods that an event initiating on one segment will stop or will propagate to other segments.

The outputs of the 3 types of stochastic models for each segment weighted using opinions of a panel of 15 experts. When results from the time-predictable model were not available, the weights on its output were 0.

So, what does it mean?

Dunno. No standard interpretation of probability applies.

Aspects of Fisher's fiducial inference, frequency theory, probability models, subjective probability.

Frequencies equated to probabilities; outcomes assumed to be equally likely; subjective probabilities used in ways that violate Bayes' Rule.

Calibrated using incommensurable data–global, extrapolated across magnitude ranges using 'empirical' scaling laws.

PRA is very similar—made-up models for various risks, hand enumeration of possibilities. Lots of "expert judgment" turned into the appearance of precise quantification.

UQ for RRW similar to EQ prediction: can't do relevant experiments to calibrate the models, lots of judgment needed.

History lessons

Helioseismology discovered errors in the EOS for hydrogenic approximation for the bound state of Fe.

Hessian-based error bars for normal mode frequencies: "statistical" versus algorithmic uncertainties—what uncertainties are included in the error bars?

Abridged catalog of sources of uncertainty

Broad categories: issues with the calibration data, issues with the theoretical approximation to the physical system, issues with the numerical approximation of the theoretical approximation in the simulator, issues with the interpolation of the simulated results, and issues with the sampling and testing of candidate models, coding errors

1. error in the calibration data, including noise and systematic error
2. approximations in the physics, including the choice of parametrization
3. numerical approximations to the approximate physics embodied in the simulator
4. algorithmic errors in the numerical approximation, tuning parameters in the simulations
5. sampling variability in stochastic algorithms and simulations
6. choices of the training points for the interpolator
7. choices of the interpolator: functional form, tuning parameters, fitting algorithm
8. choice of the measure of agreement between observation and prediction—the tester
9. choices in the sampler, including the probability distribution used and the number of samples drawn.
10. bugs

Conclusions

UQ is hard. Most UQ analyses ignore sources of uncertainty that could contribute more to the overall uncertainty than the sources they include. Some of those sources can be appraised. Errors and error bars for the original measurements are poorly understood; that might be insurmountable.

Bayesian methods make very strong assumptions about the probability distribution of data errors, models and output. Those assumptions are implausible and produce the illusion that the uncertainty is smaller than it really is.

Extrapolating complex simulations requires refusing to contemplate violations of assumptions that cannot be tested using the calibration data alone. If there is no way to design and carry out real-world experiments (not just numerical experiments), a potentially large source of uncertainty remains unquantified.

References

Evans, S.N., B. Hansen and P.B. Stark, 2005. Minimax Expected Measure Confidence Sets for Restricted Location Parameters, *Bernoulli, 11*, 571–590.

Evans, S.N. and P.B. Stark, 2002. Inverse problems as Statistics, *Inverse Problems, 18*, R1–R43.

Freedman, D.A. and P.B. Stark, 2003. What is the Chance of an Earthquake?, in *Earthquake Science and Seismic Risk Reduction*, F. Mulargia and R.J. Geller, eds. *NATO Science Series IV: Earth and Environmental Sciences, 32*, Kluwer, Dordrecht, The Netherlands, 201–213.

References, contd.

Schafer, C.M. and P.B. Stark, 2004. Using what we know: inference with physical constraints, in *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology PHYSTAT2003*, Lyons, L., R. Mount and R. Reitmeyer, eds., Stanford Linear Accelerator Center, Menlo Park, CA, 25–34.

Schafer, C.M. and P.B. Stark, 2009. Constructing Confidence Sets of Optimal Expected Size, *J. Am. Stat. Assoc.*, in press.

Stark, P.B., 1992. Inference in Infinite-Dimensional Inverse Problems: Discretization and Duality, *J. Geophys. Res., 97*, 14,055–14,082.

Stark, P.B. and L. Tenorio, 2009. A Primer of Frequentist and Bayesian Inference in Inverse Problems, in *Large Scale Inverse Problems and Quantification of Uncertainty*, Biegler, L., G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders and K. Willcox, eds. John Wiley & Sons, NY.