Uncertainty Quantification, Lecture 1

Philip B. Stark

Department of Statistics, University of California, Berkeley

5–8 February 2017, Les Diablerets, Switzerland

Reading list

- Evans, S.N. and P.B. Stark, 2002. Inverse Problems as Statistics, *Inverse Problems*, *18*, R55–R97. http://iopscience.iop.org/0266-5611/18/4/201/pdf/0266-5611_18_4_201.pdf
- Evans, S.N., B. Hansen and P.B. Stark, 2005. Minimax Expected Measure Confidence Sets for Restricted Location Parameters, *Bernoulli, 11*, 571–590.
- Freedman, D.A., 1997. Some Issues in the Foundations of Statistics, *Foundations of Science*, *1*, 19–39. http://link.springer.com/chapter/10.1007%2F978-94-015-8816-4_4#page-1
- Hengartner, N.W., and P.B. Stark, 1992. Finite-Sample Confidence Envelopes for Shape-Restricted Densities, *Ann. Statist.*, *23*, 525–550.

- Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False, *PLoS Medicine*, *2*, e124. http://dx.doi.org/10.1371/journal.pmed.0020124
- Kennedy, M.C., and A. O'Hagan, 2001. Bayesian calibration of computer models, *JRSS B*, *63*, 425–464. DOI:10.1111/1467-9868.00294
- Kuusela, M., and P.B. Stark, 2016. Shape-constrained uncertainty quantification in unfolding steeply falling elementary particle spectra. Submitted to *Annals of Applied Statistics*. http://arxiv.org/abs/1512.00905
- Mulargia, F., P.B. Stark, and R.J. Geller, 2017. Why is Probabilistic Seismic Hazard Analysis (PSHA) still used? *Phys. Earth Planet. Inter.*, http://dx.doi.org/10.1016/j.pepi.2016.12.002
- Regier, J.C. and P.B. Stark, 2015. Uncertainty quantification for emulators. *SIAM/ASA Journal on Uncertainty Quantification*, *3*, 686–708. doi:10.1137/130917909, http://epubs.siam.org/doi/10.1137/130917909

- Russi, T., A. Packard, and M. Frenklach, 2010. Uncertainty quantification: Making predictions of complex reaction systems reliable *Chemical Physics Letters*, *499*, 1–8. http://dx.doi.org/10.1016/j.cplett.2010.09.009
- Sacks, J., W.J. Welch, T.J. Mitchell, and H.P. Wynn, 1989. Design and Analysis of Computer Experiments, *Statistical Science*, *4*, 409–423. http://www.jstor.org/stable/2245858
- Saltelli, A., T.H. Andres, and T. Homma, 1993. Sensitivity analysis of model output: An investigation of new techniques, *Computational Statistics & Data Analysis*, *15*, 211–238. http://dx.doi.org/10.1016/0167-9473(93)90193-W“,
- Saltelli, A., and Funtowicz, S., 2014. When all Models are Wrong, *Issues in Science and Technology*, *30*, http://issues.org/30-2/andrea/

- Saltelli, A., P.B. Stark, W. Becker, and P. Stano, 2015. Climate Models as Economic Guides: Scientific Challenge or Quixotic Quest?, *Issues in Science and Technology*, Spring 2015. http://issues.org/31-3/climate-models-as-economic-guides-scientific-challenge-or-quixotic-quest/
- Schafer, C.M. and P.B. Stark, 2009. Constructing Confidence Sets of Optimal Expected Size, *J. Am. Stat. Assoc.*, *104*, 1080–1089. http://dx.doi.org/10.1198/jasa.2009.tm07420
- Smith, R.C., 2013. *Uncertainty Quantification: Theory, Implementation, and Applications*, SIAM, 383pp.
- Soergel, D.A.W., 2015. Rampant software errors may undermine scientific results. *F1000Research*, *3*:303. doi: 10.12688/f1000research.5930.2

- Stark, P.B., 1992. Inference in infinite-dimensional inverse problems: Discretization and duality, *Journal of Geophysical Research*, *97*, 14,055–14,082. http://onlinelibrary.wiley.com/doi/10.1029/92JB00739/epdf
- Stark, P.B. and D.A. Freedman, 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, v. 32, Kluwer, Dordrecht, The Netherlands, 201–213. Preprint: http://www.stat.berkeley.edu/~stark/Preprints/611.pdf
- Stark, P.B., 2008. Generalizing resolution, *Inverse Problems*, *24*, 034014. Reprint: http://www.stat.berkeley.edu/~stark/Preprints/resolution07.pdf

- Stark, P.B. and L. Tenorio, 2010. 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 and Sons, NY. Preprint: http://www.stat.berkeley.edu/~stark/Preprints/freqBayes09.pdf
- Stark, P.B., 2015. Constraints versus priors. *SIAM/ASA Journal of Uncertainty Quantification*, *3*(1), 586–598. http://epubs.siam.org/doi/10.1137/130920721
- Stark, P.B., 2016. Pay no attention to the model behind the curtain. http://www.stat.berkeley.edu/~stark/Preprints/eucCurtain15.pd
- Urban, M.C., 2015. Accelerating extinction risk from climate change, *Science*, *348*, 571–573. DOI: 10.1126/science.aaa4984

Topics (out of order)

- Why UQ?
- quality of evidence
  - "it's the data, stupid" (and the code, too)
  - known unknowns and unknown unknowns
  - experimental design
  - error models v model errors
  - bugs, reproducibility, replicability
  - epistemic and aleatory uncertainty
- constraints and priors
- models all the way down
  - Freedman's Rabbit-Hat theorem
  - hierarchical priors
  - the ludic fallacy
- rates versus probabilities
- theories of probability
- probability as metaphor
- responsible quantification v. *quantifauxcation*

- how big is your model space?
  - bias/variance tradeoffs
  - is the bias bounded?
  - regularization for inference vs estimation
  - assuming the problem away
- lampposting and hypocognition
- how bad can it be?
  - lower bounds on minimax uncertainties for emulators
  - uncertainty quantification for the HEP unfolding problem
  - curse of dimensionality
  - bugs, bugs, bugs
  - reproducibility and replicability; verification and validation

Abstract

UQ tries to appraise and quantify the uncertainty of models of physical systems calibrated to noisy data (and of predictions from those models), including contributions from ignorance; systematic and stochastic measurement error; limitations of theoretical models; limitations of numerical representations of those models; limitations of the accuracy and reliability of computations, approximations, and algorithms; and human error (including software bugs).
Much UQ research focuses on developing efficient numerical approximations (emulators) of computationally expensive numerical models.
In some circles, UQ is nearly synonymous with the study of emulators and Bayesian models.

What sources of uncertainty does a UQ analysis take into account? What does it ignore? How ignorable are the ignored sources? What assumptions were made? What evidence supports those assumptions? Are the assumptions testable? What happens if the assumptions are false?

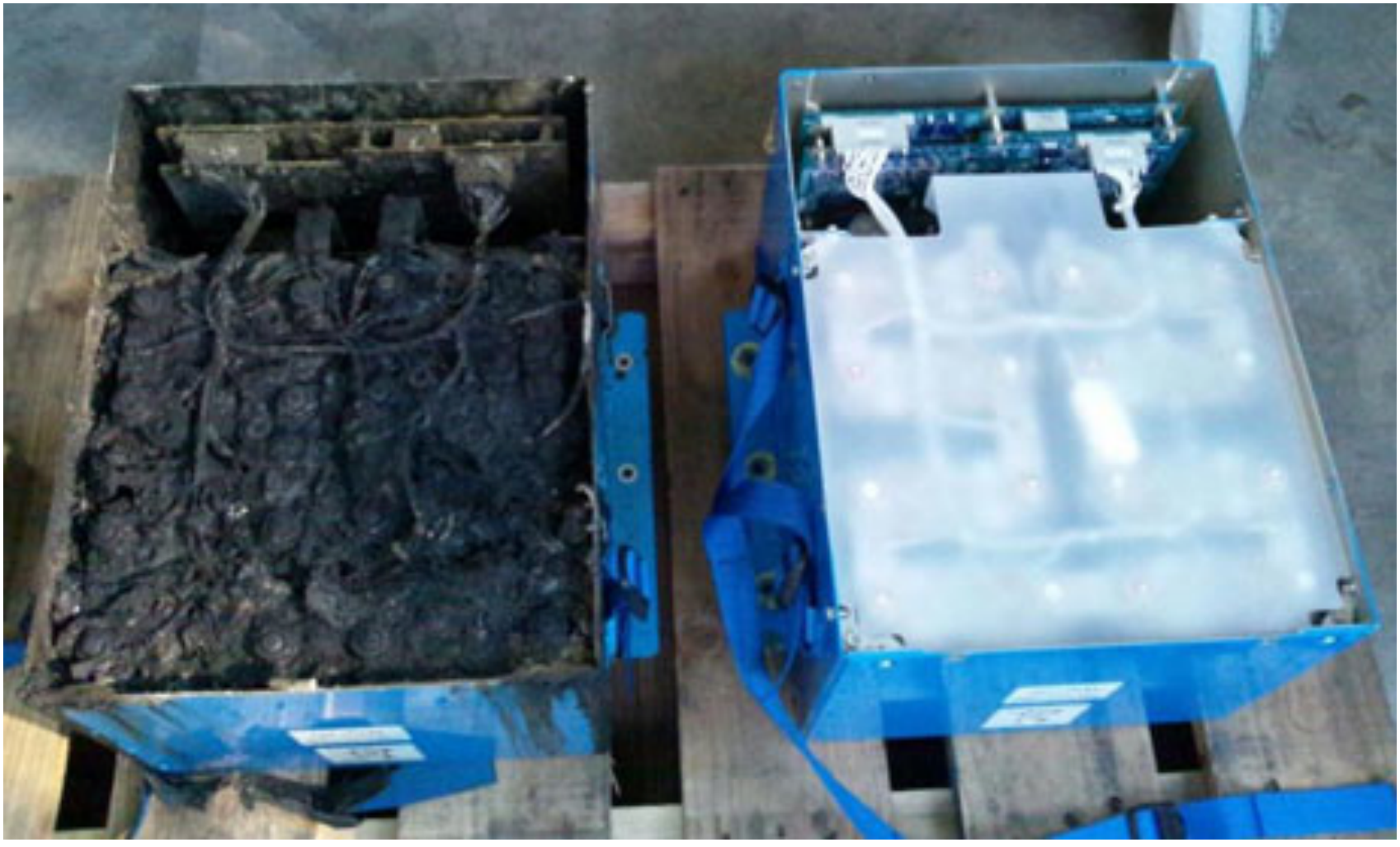
I will sketch an embedding of UQ within the theory of statistical estimation and inverse problems.

I will point to a few examples of work that quantifies uncertainty from systematic measurement error and discretization error.

Bad examples will be drawn from the 2009 NAS report, "Evaluation of Quantication of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile," the Climate Prospectus, and probabilistic seismic hazard analysis (PSHA).

Why Uncertainty Quantification Matters

James Bashford / AP



Reuters / Japan TSB

Figure: L'Aquila

What is UQ?

UQ is typically more specific than just quantifying uncertainty:

$$\text{UQ} = \text{inverse problems} + \text{approximate forward model.}$$

What are inverse problems?

statistics

What is the effect of an approximate forward model, discretization, etc.?

additional systematic measurement error

How can we deal with systematic measurement error?

statistics

So,

$$\text{UQ} = \text{statistics}$$

What makes UQ special?

- the particular sources of systematic error
- poorly understood/characterized measurement error
- poorly understood/characterized properties of the underlying "model"
- heavy computational burden (in some applications)
- numerical approximations
- reliance on simulation
- big data (in some applications)
- heterogeneous and legacy data (in some applications)
- need for speed (in some applications)
- societal consequences (in some applications)

Abstract mumbo-jumbo

How can we embed UQ in the framework of statistics?[1]

Statistical decision theory.

Ingredients:

- *The the state of the world* θ. Math object that represents the physical system.
- Set of possible states of the world Θ. Know *a priori* that $\theta \in \Theta$.
- Observations Y. Sample space of possible observations $\mathcal{Y}$.
- *measurement model* that relates the probability distribution of Y to θ. If θ is state of the world, then $Y \sim \Pr_\theta$. Incorporates the forward model.
- one or more *parameters* of interest, $\lambda = \lambda[\theta]$
- an *estimator* $\hat{\lambda}(Y)$ of the parameter (might be set-valued)
- a risk function that measures the expected loss from estimating $\lambda[\theta]$ by $\hat{\lambda}(Y)$

[1] Moreover, does it help?

How does UQ fit into this framework?

- What's $\Pr_\theta$?
- Systematic errors are additional unknown parameters.
  - need constraints on them or can't say much
- Augment θ, Θ to include the systematic errors as parameters.
- Systematic errors are *nuisance parameters*: the distribution of the data depends on them, but they are not of interest.

What's missing?

- Given θ, do we actually know (or can we simulate from) Pr_θ?
  Do we know the mapping $\theta \to \mathrm{Pr}_\theta$?
  If not, more unknowns to take into account.
- Usefully constrained sets Θ of possible models.
- Ways of quantifying/bounding the systematic error.
- Ways of assessing the stochastic errors.
- Estimators $\hat{\lambda}$ for $\lambda[\theta]$ in light of the stochastic and systematic errors, Θ, $\theta \to \mathrm{Pr}_\theta$.

What can we do with the framework?

- Bayes or frequentist analysis?
- Nature of the assumptions.
- Where does the prior come from?

Back to basics: Data quality

Tendency to gloss over data uncertainties:

- ignore systematic error
- treat all error bars as if they were SDs (or a multiple)
- treat all measurement error as Normal (or Poisson, for counts)
- treat measurement errors as independent
- ignore data reduction steps, normalization, calibration background fits, etc.
- treat inverse of final Hessian of nonlinear LS as if it characterizes the uncertainty.

Where do the data come from?

Design & processing matter

- random sampling?
- random assignment to treatment or control?
- understood instrumental errors?
- Many steps of reduction and processing from raw instrumental/experimental/observational data to produce the numbers that statisticians work with.
  - can take place in the instrument or the pipeline
  - poorly understood effect on uncertainties/errors
  - often based on heuristics
  - raw data often not recorded or not retained

Data quality: It ain't what we pretend it is

- "n =all": Boston bump, predictive policing
- Helioseismology. Nominal "statistical" uncertainties didn't even account for numerical instability in the data reduction.
- Post-Enumeration Survey data from the U.S. Census
- online behavior monitoring
- historical nuclear test data used to calibrate numerical models for "Reliable Replacement Warhead."
  - instruments gone
  - people who recorded the data retired
  - transformations & data reduction mysterious
  - lots of $\pm 10\%$: What does $\pm 10\%$ mean?

Can't get off the ground

How can you know how well the model should fit the data, if you don't understand the nature and probable / possible / plausible size of systematic and stochastic errors in the data?

Theory and Practice

- In theory, there's no difference between theory and practice. But in practice, there is.
  –*Jan L.A. van de Snepscheut*
- The difference between theory and practice is smaller in theory than it is in practice.
  –*unknown UQ master*

Bad incentives:

Grappling with Data Quality ain't Sexy

- Academic statisticians rewarded for proving hard theorems, doing heroic numerical work (speed or size), making splashy images that get on the cover of *Nature*, being "first."
- We fall in love with technology, models, technique, tools.
- Digging into data quality, systematic errors, etc., is crucial, unglamorous, and unrewarded—but crucial.
- Can't Q U without understanding limitations of the data.

The society which scorns excellence in plumbing as a humble activity and tolerates shoddiness in philosophy because it is an exalted activity will have neither good plumbing nor good philosophy: neither its pipes nor its theories will hold water.
–*John W. Gardner*

What does the analysis tell us?

- If UQ gives neither an upper bound nor a lower bound on a sensibly defined measure of uncertainty, what have we learned?
- At the very least, should list what we have and have not taken into account.

Examples with uncertain forward models and discretization error

- Stark (1992) treats a problem in helioseismology in which the forward model is known only approximately; bounds the systematic error that introduces and takes it into account to find confidence sets for a fully infinite-dimensional model; also gives a general framework.
- Evans & Stark (2002) give a more general framework.
- Stark (2008) discusses generalizing "resolution" to nonlinear problems and problems with systematic errors.
- Gagnon-Bartsch & Stark (2012) treat a problem in gravimetry with discretized domain; bound systematic error from discretization and take it into account to find confidence sets for a fully infinite-dimensional model.

Generic approach: Strict Bounds

- Find sup and inf of parameter $\lambda[\theta]$ of interest over a confidence set for the model θ, including stochastic and systematic error included in $\Pr_\theta$.
- Leads to infinite-dimensional optimization problems
  - can be exactly reduced to finite-dimensional problems in some cases
  - prior constraints usually essential
  - functionals that can be estimated w finite uncertainty are limited
  - convexity and other properties help
  - often solvable using Fenchel duality

The optimization problem

$$Y = \mathcal{K}[\theta] + \epsilon$$

$$\mu \equiv \mathcal{K}[\theta]$$

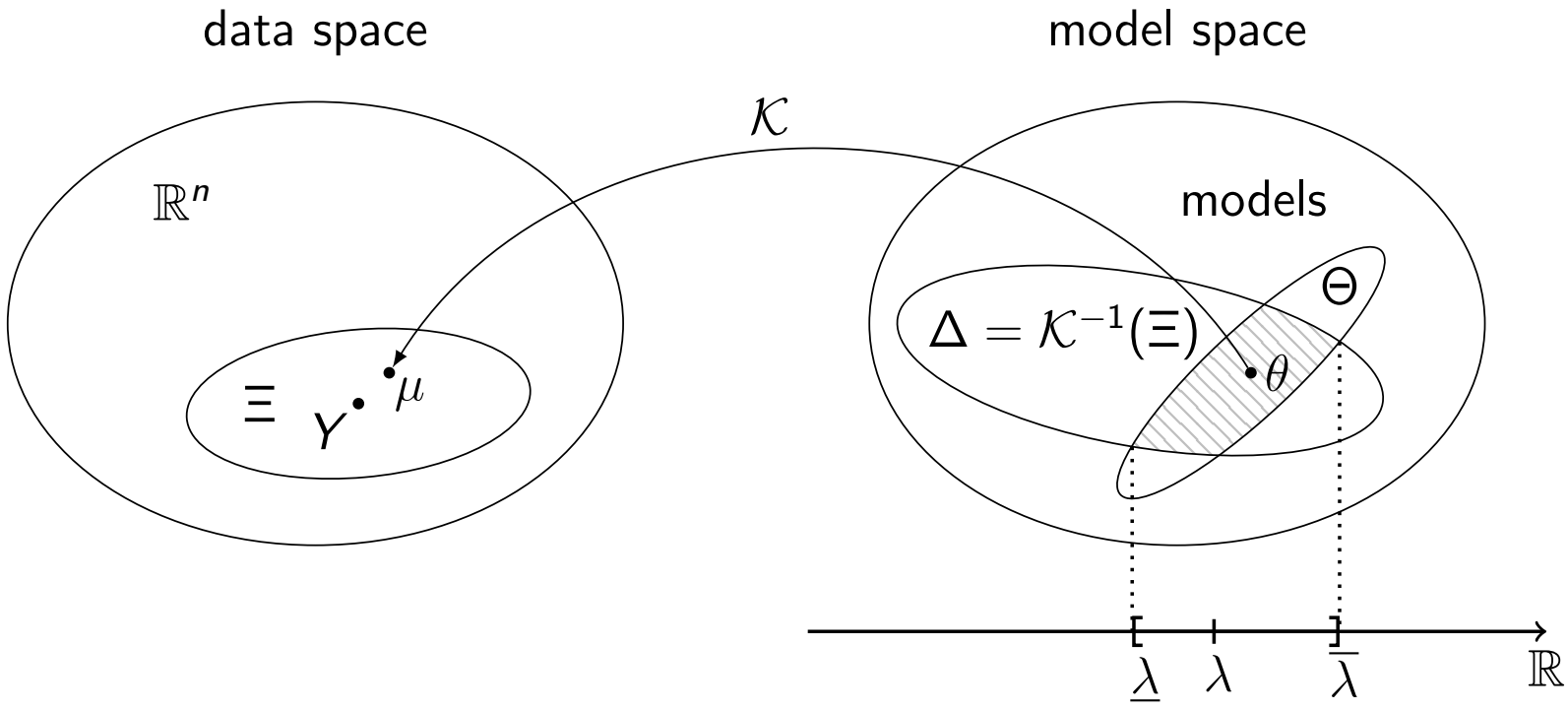
$$\Xi = \Xi(Y) \text{ satisfies } \Pr\{\Xi \ni \mu\} \geq 1 - \alpha$$

$$\Delta \equiv \{\nu : \mathcal{K}[\nu] \in \Xi\}$$

Then

$$[\inf_{\nu \in \Delta \cap \Theta} \lambda[\nu], \sup_{\nu \in \Delta \cap \Theta} \lambda[\nu]]$$

is a $1 - \alpha$ confidence interval for $\lambda[\theta]$.



(Thanks to Mikael Kuusela!)

Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile (EQMU)

Committee on the Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile, 2009.
http://www.nap.edu/openbook.php?record_id=12531&page=R1

Fundamental Theorem of Physics

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

Lemma: Nobody knows more about physics than physicists.[2]

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

[2] Follows from the axiom: Nobody knows more about *anything* than physicists.

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.

NAS panel included physicists, nuclear engineer, senior manager, probabilistic risk assessor, and one statistician

Cream of EQMU (p25)

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.

EQMU example, 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 $\hat{V}_T = 100V$, with uncertainty $U_T = 5V$.
V_A estimated as $\hat{V}_A = 150V$, 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, p46)

"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, p14)

Scratching the veneer

- Are V_A and/or V_T random? Or simply unknown?
- Are $\hat{V}_A$ and $\hat{V}_T$ 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, p13)
- "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, p15)

Vision 1: Us are error bars

Suppose V_A and V_T are independent random variables[3] with known means $\hat{V}_A$ and $\hat{V}_T$, respectively. Suppose $\Pr\{\hat{V}_A - V_A \leq U_A\} = 90\%$ and $\Pr\{V_T - \hat{V}_T \leq U_T\} = 90\%$.

- What's $\Pr\{V_A - V_T \geq 0\}$?
  Can't say, but ...
- Bonferroni's inequality:

$$\Pr\{\hat{V}_A - V_A \leq U_A \text{ and } V_T - \hat{V}_T \leq U_T\} \geq 80\%.$$

- Conservative bound. What's the right answer?

[3] 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, p27)

- 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 $\hat{V}_A$ and $\hat{V}_T$ 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 they are correlated, SD could be $\sqrt{U_A^2 + U_T^2 + 2U_A U_T}$

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

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

Chebychev's inequality:

$$\Pr\left\{|V_A - V_B - (\hat{V}_A - \hat{V}_B)| \leq kU\right\} \geq 1 - \frac{1}{k^2}.$$

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

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

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

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

Vision 3: one of each

From description, makes sense that:

- V_T is an unknown parameter
- $\hat{V}_T$ is an already-computed estimate of V_T from data
- $\hat{V}_A$ is a design parameter
- 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 $V_T - \hat{V}_T \leq U_T$ or not: no probability left, only ignorance.

Whether $\hat{V}_A - V_A \leq U_A$ 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?
- 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 can retire, 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 nonsense.

Bad quotes (EQMU, p21)

"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, p23)

"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.
  - Maxwell-Boltzman. State vector gives the quantum state of each particle: r^n possible values.
  - Bose-Einstein. State vector gives # particles in each quantum state: $\binom{n+r-1}{n}$ possible values.
  - Fermi-Dirac. State vector gives the number of particles in each quantum state, but no two particles can be in the same state: $\binom{r}{n}$ possible values.

- Maxwell-Boltzman common in probability theory, but but describes 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.

Constraints versus prior probabilities

Bayesian machinery 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: $\Pr(B|A) = \Pr(A|B)\Pr(B)/\Pr(A)$. "Just math."
  - To have posterior $\Pr(B|A)$, need prior $\Pr(B)$.
  - The prior matters. Where does it come from?

Conservation of Rabbits

The Rabbit Axioms

1. For the number of rabbits in a closed system to increase, the system must contain at least two rabbits.
2. No negative rabbits.

Freedman's Rabbit-Hat Theorem

You cannot pull a rabbit from a hat unless at least one rabbit has previously been placed in the hat.

- The prior puts the rabbit in the hat
- PRA puts many rabbits in the hat
- Hierarchical priors put many rabbits in the hat
- Bayes/minimax duality: minimax uncertainty is Bayes uncertainty for least favorable prior.[4]

[4] 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 approach: capture constraint using prior, e.g., $\theta \sim U[-\tau, \tau]$.
  - Credible region: 95% posterior probability.
- Frequentist approach: use constraint directly.
  - Confidence interval: 95% coverage probability whatever be θ.

95% Confidence sets vs. credible regions

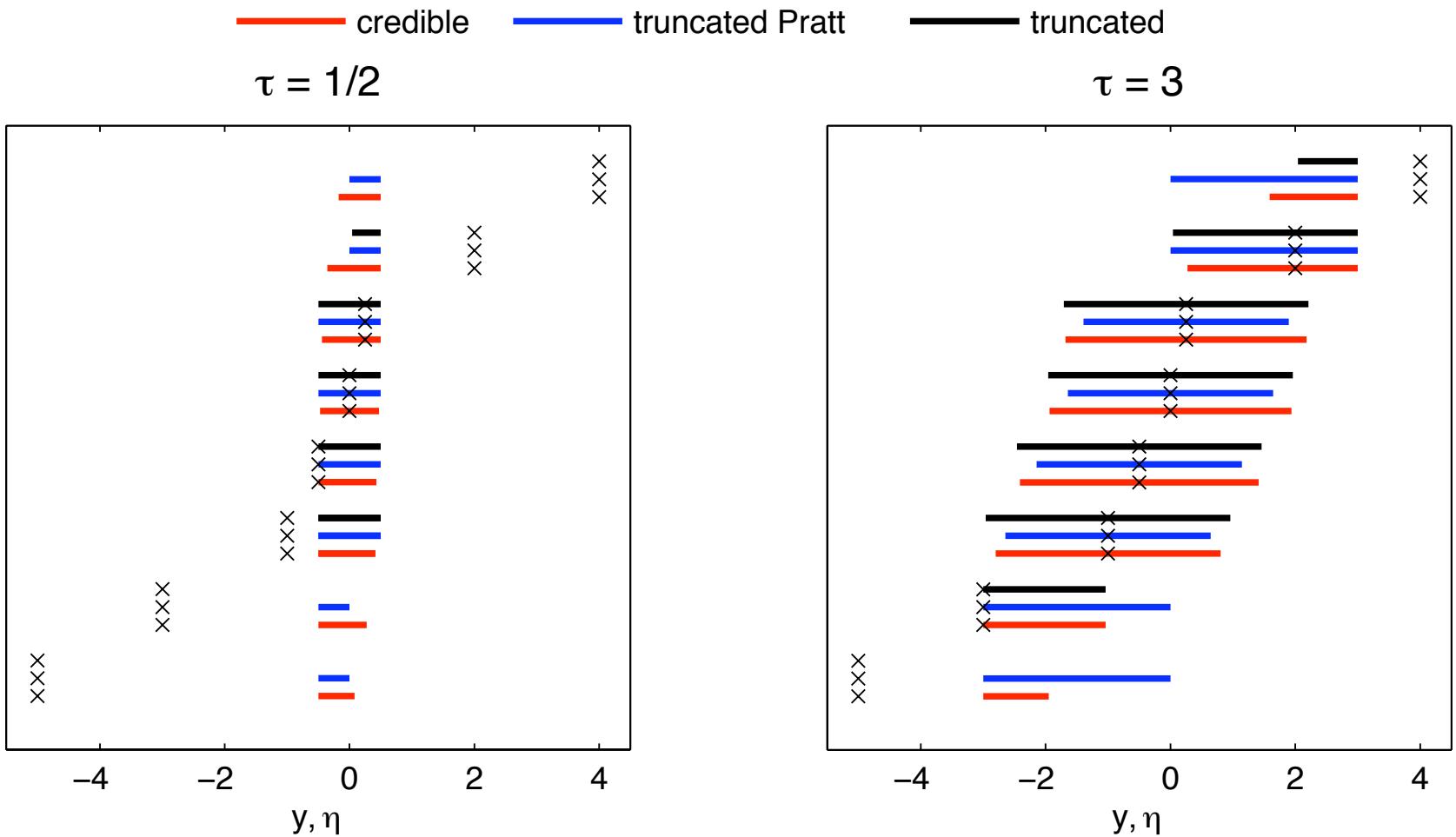


Figure: ../../Ms/Sandia09/intervals

Coverage of 95% credible regions

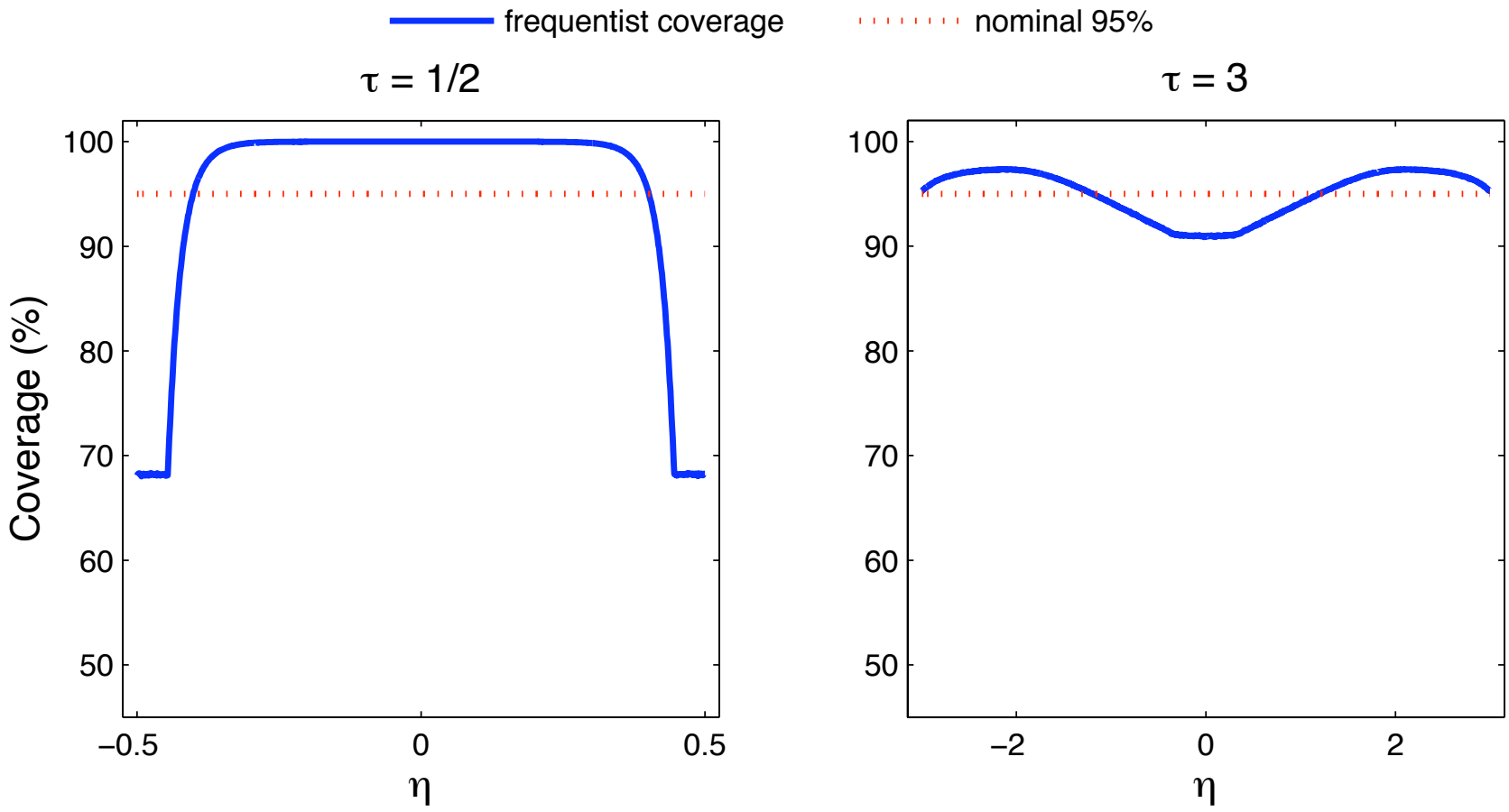


Figure: ../../Ms/Sandia09/freqcvge

Expected size of credible regions and confidence intervals

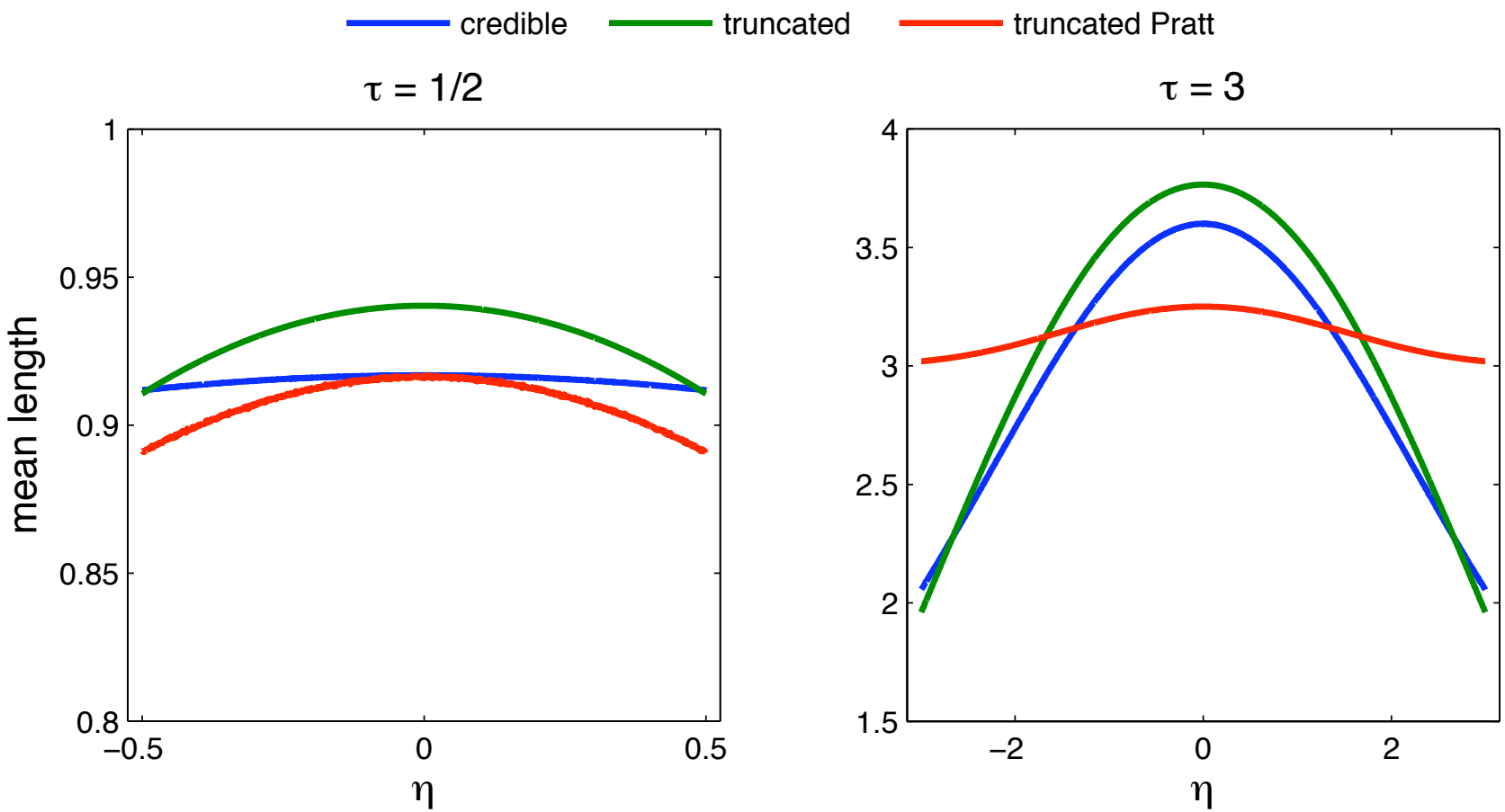


Figure: ../../Ms/Sandia09/lengths

Interpreting earthquake predictions (with D.A. Freedman)

Globally, on the order of 1 magnitude 8 earthquake per year. Locally, recurrence times for big events $O(100\text{ y})$. 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

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

- Identify (by expert opinion) ways segments of each fault can rupture separately and together. Each combination of segments is a 'seismic source.'
- Determine (by expert opinion) extent that long-term fault slip is accommodated by rupture of each combination of segments for each fault.
- 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.
- Using the chosen relationship and the assumed probability distribution for its parameters, determine a mean event magnitude for each seismic source by Monte Carlo.

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

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

Stage 2:

- Fit 3 stochastic models for earthquake recurrence—Poisson, Brownian passage time and "time-predictable"—to long-term seismicity rates estimated in stage 1.
- Combine stochastic models to estimate chance of a large earthquake:
  - Use Poisson and Brownian passage time models 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.

- contd.
  - 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.

- contd.
  - 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.
  - 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?

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

More uncertainty: failures of reproducibility

- attempts to replicate experiments or data analyses often fail to support the original claims.[5]
- *P*-hacking, ignoring multiplicity, small effects, file-drawer effect, bugs, etc.
- Failures contribute to uncertainty: hard to quantify
- Journals generally uninterested in publishing negative results or replications of positive results: emphasis is on "discoveries."
- Thermo ML found ~20% of papers that otherwise would have been accepted had serious errors, discovered b/c required sharing data

[5] E.g., http://science.sciencemag.org/content/349/6251/aac4716.full http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off

- Selecting data, hypotheses, data analyses, and results, to produce (apparently) positive results inflates the apparent signal-to-noise ratio and overstates statistical significance.
- Automation of data analysis, including feature selection and model selection, combined with the large number of variables measured in many modern studies and experiments, including "omics," high-energy physics, and sensor networks: inevitable that many "discoveries" will be wrong.

Most software has bugs

- 2014 study by Coverity, based on code-scanning algorithms, found 0.61 errors per 1,000 lines of source code in open-source projects and 0.76 errors per 1,000 lines of source code in commercial software[6]
- Few scientists use sound software engineering practices, such as rigorous testing—or even version control.[7]

---

[6] http://go.coverity.com/rs/157-LQW-289/images/2014-Coverity-Scan-Report.pdf

[7] See, e.g.: Merali, Z., 2010. Computational science: ... Error ... why scientific programming does not compute. *Nature*, *467*, 775–777 doi:10.1038/467775a http://www.nature.com/news/2010/101013/full/467775a.html; Soergel, D.A.W., 2015. Rampant software errors may undermine scientific results. *F1000Research*, *3*, 303. doi:10.12688/f1000research.5930.2 http://f1000research.com/articles/3-303/v2

“Rampant software errors undermine scientific results”

Soergel, 2015

> *Errors in scientific results due to software bugs are not limited to a few high-profile cases that lead to retractions and are widely reported. Here I estimate that in fact most scientific results are probably wrong if data have passed through a computer, and that these errors may remain largely undetected. The opportunities for both subtle and profound errors in software and data management are boundless, yet they remain surprisingly underappreciated.*

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation
- Coding standards/conventions
- Pair programming
- Issue trackers
- Code reviews (and in teaching, grade *code*, not just *output*)
- Code tests: unit, integration, coverage, regression

Integration tests

Figure: Integration testing

http://imgur.com/qSN5SFR by Datsun280zxt

Spreadsheets especially bad

- Easier to commit errors and harder to find them.
- 2010 work of Reinhart and Rogoff[8] used to justify austerity measures in southern Europe. Errors in their Excel spreadsheet lead to the wrong conclusion[9]
- Spreadsheets might be OK for data entry, but not for calculations.
- Conflate input, output, code, presentation; facilitate & obscure error
- According to KPMG and PWC, over 90% of corporate spreadsheets have errors
- Not just errors: bugs in Excel too: +, *, random numbers, statistical routines
- "Stress tests" of international banking system use Excel simulations!

[8] Reinhart, C. and K. Rogoff, 2010. Growth in a Time of Debt, Working Paper no. 15639,
National Bureau of Economic Research,
http://www.nber.org/papers/w15639; Reinhart, C. and K. Rogoff, 2010

Questions for reproducibility

- materials (organisms), instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?
- code & data available to re-generate figures and tables?
- code readable and checkable?
- software build environment specified adequately?
- what is the evidence that the result is correct?
- how generally do the results hold? how stable are the results to perturbations of the experiment?

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?
- Was it done correctly?
- Were the results reported correctly?
- Were there ad hoc aspects? What if different choices had been made?
- What other analyses were tried?
- How was multiplicity treated?
- Can someone else use the procedures and tools?

Abridged catalog of sources of uncertainty

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

- faulty assumptions
- error in the calibration data, including noise and systematic error, and assumptions about these
- approximations in the model, including physics and parametrization
- finite-precision arithmetic
- numerical approximations to the approximate physics embodied in the simulator
- algorithmic errors in the numerical approximation, tuning parameters in the simulations
- sampling variability in stochastic algorithms and simulations
- limitations of PRNGs and other algorithms; numerical approximations
- choices of the training points for the interpolator/emulator
- choices of the interpolator: functional form, tuning parameters, fitting algorithm
- choice of the measure of agreement between observation

Conclusions

- UQ is hard to do well.
- Most attempts ignore sources of uncertainty that could contribute more than the sources they include: lampposting.
- Some of those sources can be appraised.
- Errors and error bars for the original measurements are poorly understood: insurmountable?
- Bayesian methods make very strong assumptions about the probability distribution of data errors, models and output; reduce apparent but not real uncertainty.
- Extrapolating complex simulations requires refusing to contemplate violations of assumptions that cannot be tested using calibration data.
- Numerical experiments are not an adequate substitute for real experiments.