Bayesian data analysis: what it is and what it is not

Prof. Andrew Gelman
Dept. of Statistics
Columbia University

Talk for Columbia University Department of Computer Science, 15 Dec 2003

Themes

- Popular view of Bayesian statistics
  - Subjective probability
  - Elicited prior distributions
- Bayesian data analysis as we do it
  - Hierarchical modeling
  - Many applications
- Conceptual framework
  - Fit a probability model to data
  - Check fit, ride the model as far as it will take you
  - Improve/expand/extend model

Overview

- Decision analysis for home radon
- Where did the prior dist come from?
- Quotes illustrating misconceptions of Bayesian inference
- State-level opinions from national polls (hierarchical modeling and validation)
- Serial dilution assay

  (handling uncertainty in a nonlinear model)
- Simulation-based model checking
- Some open problems in BDA

Decision analysis for home radon

- Radon gas
  - Causes 15,000 lung cancers per year in U.S.
  - Comes from underground; home exposure
- Radon webpage
  - Click for recommended decision
- Bayesian inference
  - Prior + data = posterior
  - Where did the prior distribution come from?

Prior distribution for your home's radon level

- Example of Bayesian data analysis
- Radon model
  - theta_i = log of radon level in house i in county j(i)
  - linear regression model:
    theta_i = a_j(i) + b_1*base_i + b_2*air_i + … + e_i
  - linear regression model for the county levels a_j,
    given geology and uranium level in the county, with county-level errors
- Data model
  - y_i = log of radon *measurement* in house I
  - y_i = theta_i + Bias + error_i
  - Bias depends on the measurement protocol
  - error_i is not the same as e_i in radon model above

Radon data sources

- National radon survey
  - Accurate unbiased data—but sparse
  - 5000 homes in 125 counties
- State radon surveys
  - Noisy biased data, but dense
  - 80,000 homes in 3000 counties
- Other info
  - House level (basement status, etc.)
  - County level (geologic type, uranium level)

Bayesian inference for home radon

- Set up and compute model
  - 3000 + 19 + 50 parameters
  - Inference using iterative simulation (Gibbs sampler)
- Inference for quantities of interest
  - Uncertainty dist for any particular house (use as prior dist in the webpage)
  - County-level estimates and national averages
  - Potential $7 billion savings
- Model checking
  - Do inferences make sense?
  - Compare replicated to actual data, cross-validation
  - Dispersed model validation (“beta-testing”)

Bayesian inference for home radon

- Allows estimation of over 3000 parameters
- Summarizes uncertainties using probability
- Combines data sources
- Model is testable (falsifiable)

Pro-Bayesian quotes

- Hox (2002):

  "In classical statistics, the population parameter has only one specific value, only we happen not to know it. In Bayesian statistics, we consider a probability distribution of possible values for the unknown population distribution."

- Somebody's webpage:

  "To a true subjective Bayesian statistician, the prior distribution represents the *degree of belief* that the statistician or client has in the value of the unknown parameter . . . it is the responsibility of the statistician to *elicit the true beliefs* of the client."

Why these views of Bayesian statistics are wrong!

- Hox quote (distribution of parameter values)
  - Our response: parameter values are “fixed but unknown” in Bayesian inference also!
  - Confusion between *quantities of interest* and *inferential summaries*
- Anonymous quote
  - Our response: the statistical model is an assumption to be held if useful
  - Confusion between *statistical analysis* and *decision making*

Anti-Bayesian quotes

- Efron (1986):

  "Bayesian theory requires a great deal of thought about the given situation to apply sensibly."
- Ehrenberg (1986):

  "Bayesianism assumes: (a) *Either* a weak or uniform prior, in which case why bother?, (b) *Or* a strong prior, in which case why collect new data?, (c) *Or* more realistically, something in between, in which case Bayesianism always seems to duck the issue."

Why these views of Bayesian statistics are wrong!

- Efron quote (difficulty of Bayes)
  - Our response: demonstration that Bayes solves many problems more easily that other methods
  - Mistaken focus on the simplest problems
- Ehrenberg quote (arbitrariness of prior dist)
  - Our response: the "prior dist" represents the information provided by a group-level analysis
  - One "prior dist" serves many analyses

State-level opinions from national polls

- Goal: state-level opinions
  - State polls: infrequent and low quality
  - National polls: N is too small for individual states
  - Also must adjust for survey nonresponse
- Try solving a harder problem
  - Estimate opinion in each of 3264 categories:
    - 51 states
    - 64 demographic categories (sex, ethnicity, age, education)
- Logistic regression model
- Sum over 64 categories within each state
- Validate using pre-election polls

State-level opinions from national polls

- Logistic regression model
  - y_i = 1 if person i supports Bush, 0 otherwise
  - logit (Pr(y_i=1)) = linear predictor given demographics and state of person i
  - Hierarchical model for 64 demographic predictors and 51 state predictors (including region and previous Republican vote share in the state)
  - State polls could be included also if we want
- Sum over 64 categories within each state
  - "Post-stratification"
  - In state s, the estimated proportion of Bush supporters is sum(j=1 to 64) N_j Pr(y=1 in category j) / sum(j=1 to 64) N_j
  - Also simple to adjust for turnout of likely voters

Compare to “no pooling” and “complete pooling”

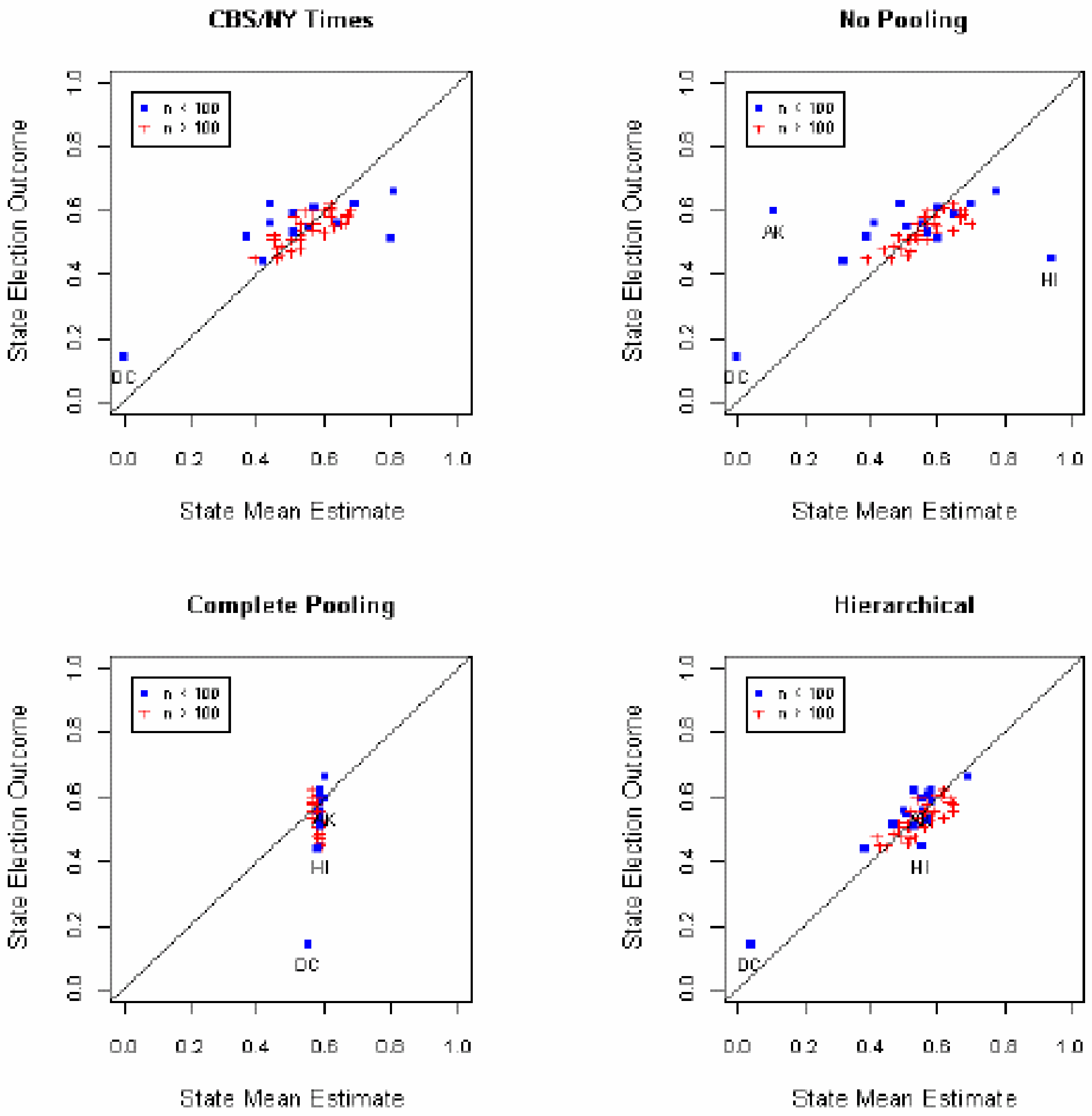
- “No pooling”
  - Separate estimate within each state
  - Treat the survey as 49 state polls
  - Expect “overfitting”: too many parameters
- “Complete pooling”
  - Include demographics only
  - Give up on estimating 51 state parameters
- Competition
  - Use pre-election polls and compare to election outcome
  - Estimated Bush support in U.S.
  - Estimates in individual states

Election poll analysis and Bayesian inference

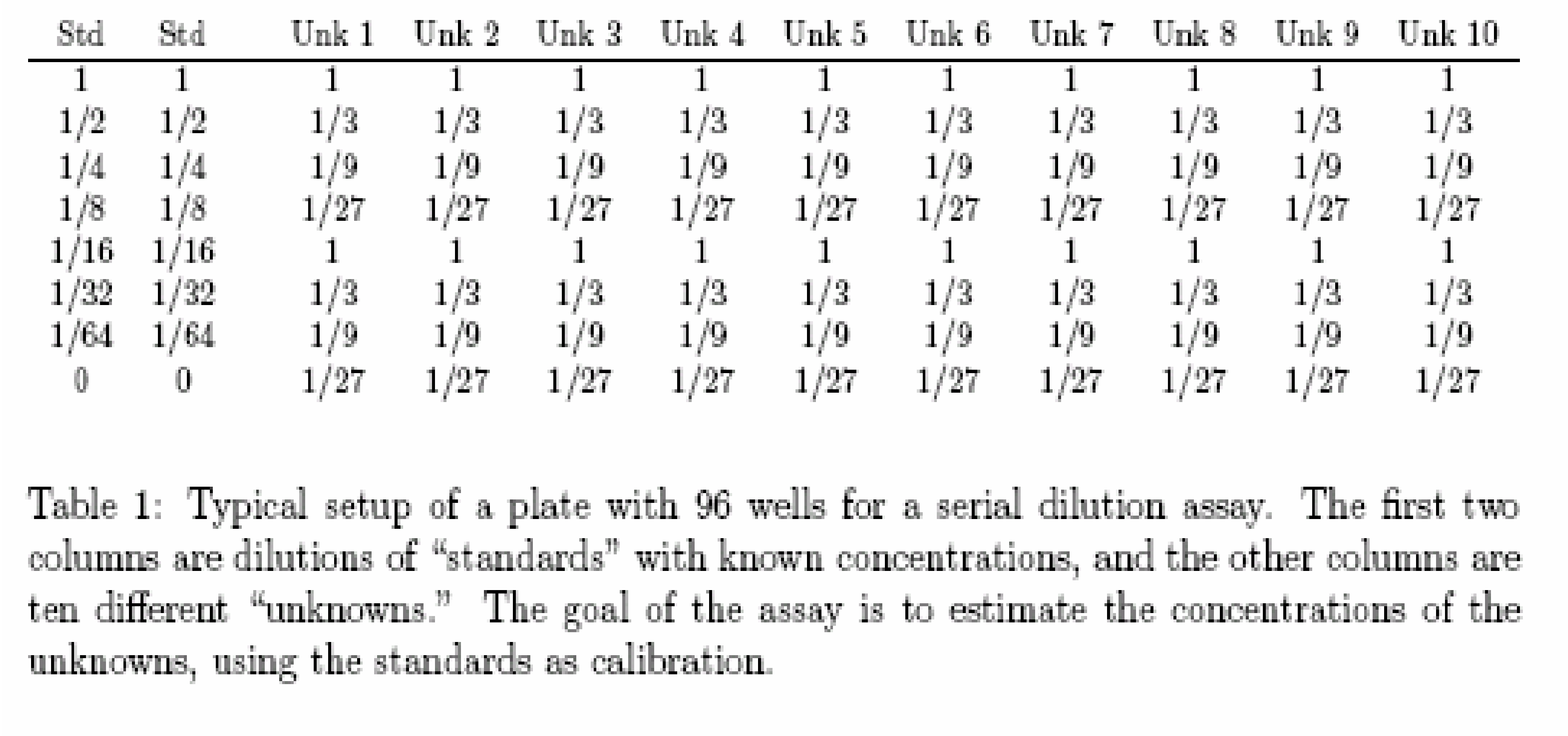
- Where was the “prior distribution”?
  - In logistic regression model, 51 state effects, a_k
  - a_k = b*presvote_k + c_region(k) + e_k
  - Errors e_k have Normal (0, sigma^2) distribution; sigma estimated from data
- Where was the “subjectivity”?

Election poll analysis: validation

	No pooling	Complete pooling	Bayes
Avg std errors of state estimates	5.1%	0.9%	3.1%
Avg of actual absolute state errors	5.1%	5.9%	3.2%


CBS/NY Times
No Pooling
Complete Pooling
Hierarchical
n < 100
n > 100
State Election Outcome
State Mean Estimate
0.0
0.2
0.4
0.6
0.8
1.0
DC
AK
HI

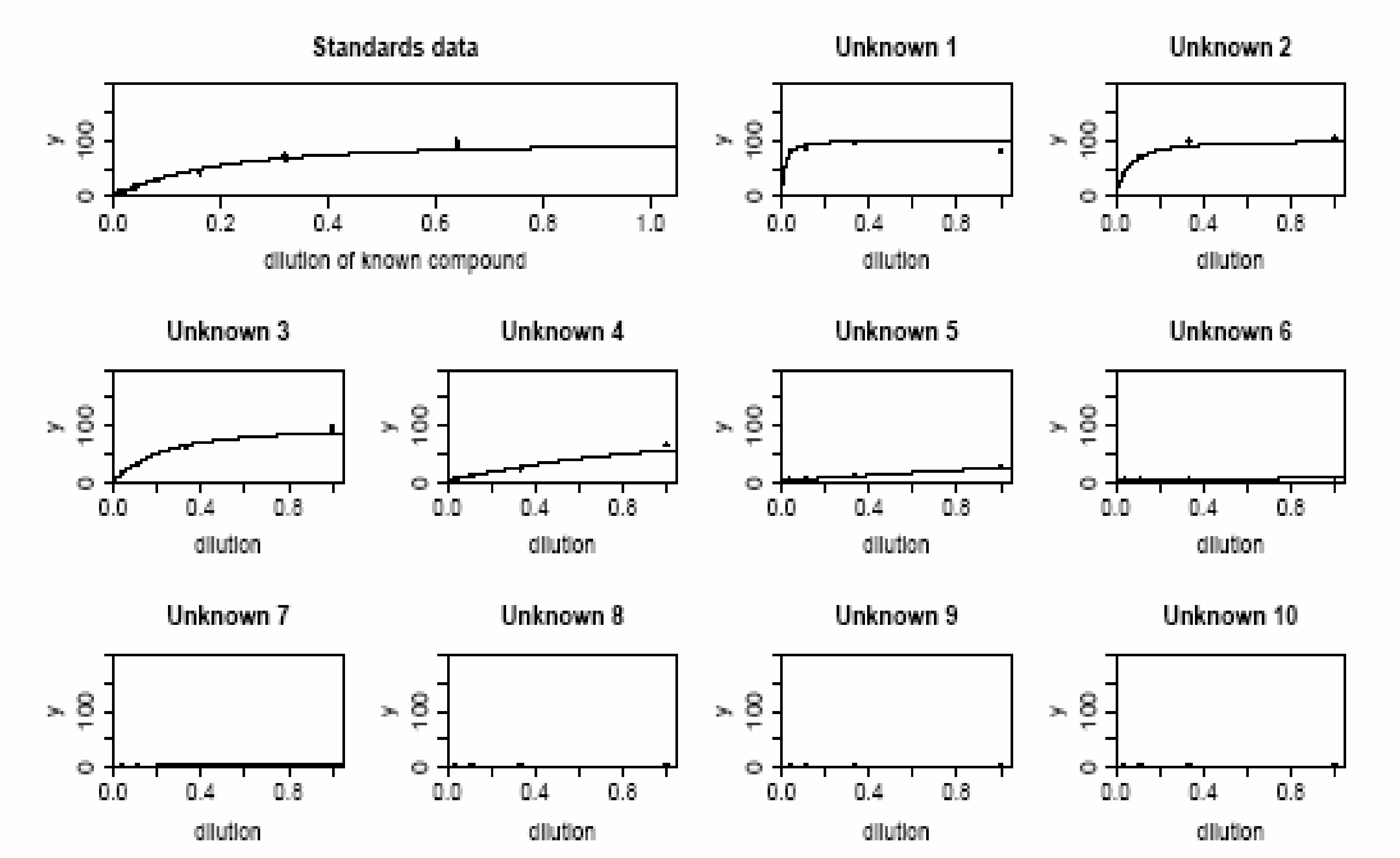
Serial dilution assays



Std	Std	Unk 1	Unk 2	Unk 3	Unk 4	Unk 5	Unk 6	Unk 7	Unk 8	Unk 9	Unk 10
1	1	1	1	1	1	1	1	1	1	1	1
1/2	1/2	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
1/4	1/4	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
1/8	1/8	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27
1/16	1/16	1	1	1	1	1	1	1	1	1	1
1/32	1/32	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
1/64	1/64	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
0	0	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27	1/27

Table 1: Typical setup of a plate with 96 wells for a serial dilution assay. The first two columns are dilutions of "standards" with known concentrations, and the other columns are ten different "unknowns." The goal of the assay is to estimate the concentrations of the unknowns, using the standards as calibration.

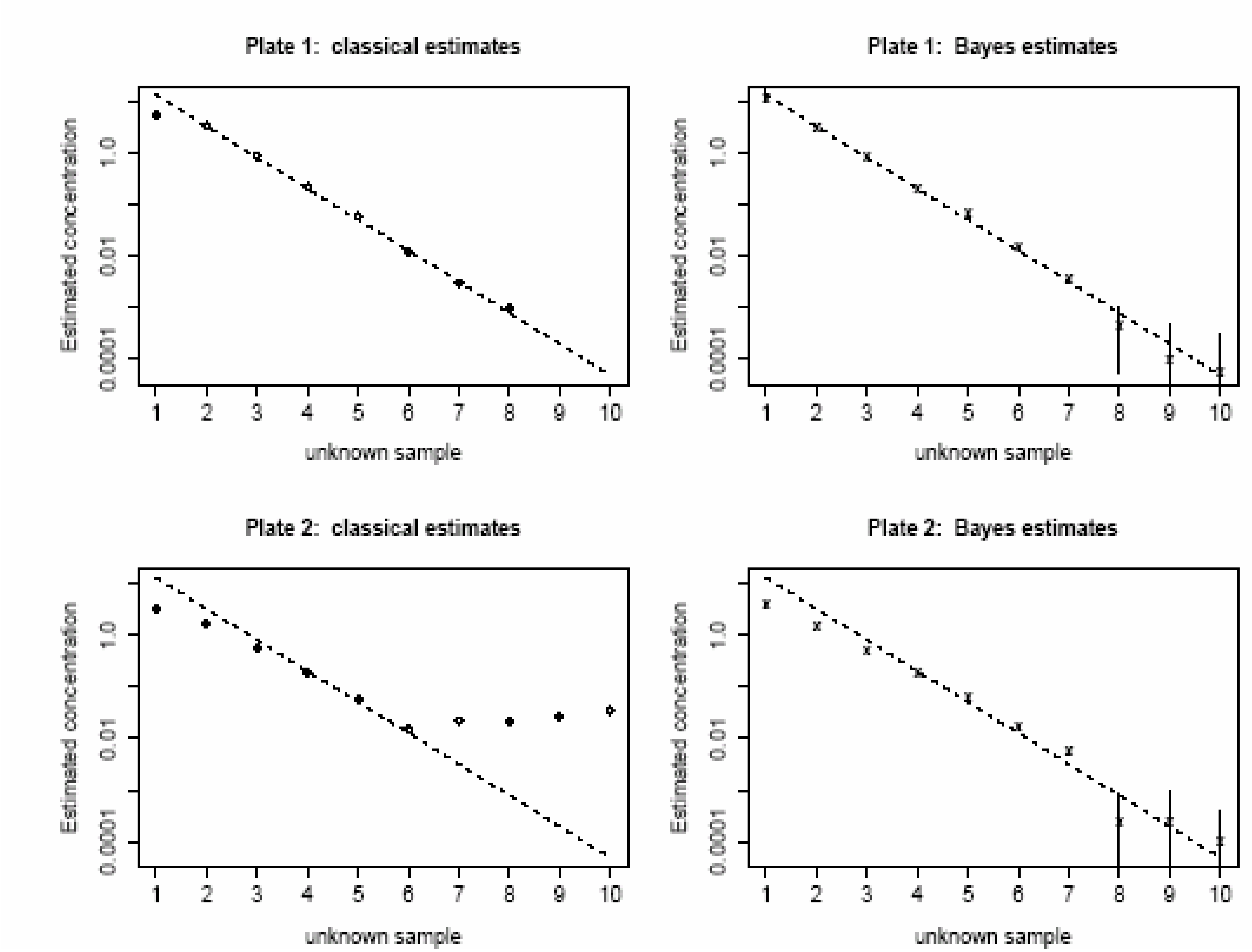
Serial dilution assays



Serial dilution assays: motivation for Bayesian inference

- Classical approach: read the estimate off the calibration curve
- Difficulties
  - "above detection limit": curve is too flat
  - "below detection limit": signal/noise ratio is too low
  - For some samples, **all the data** are above or below detection limits!
- Goal: downweight—but don't discard—weak data
- Maximum likelihood (weighted least squares)
- Bayes handles uncertainty in the parameters of the calibration curve

Serial dilution: validation

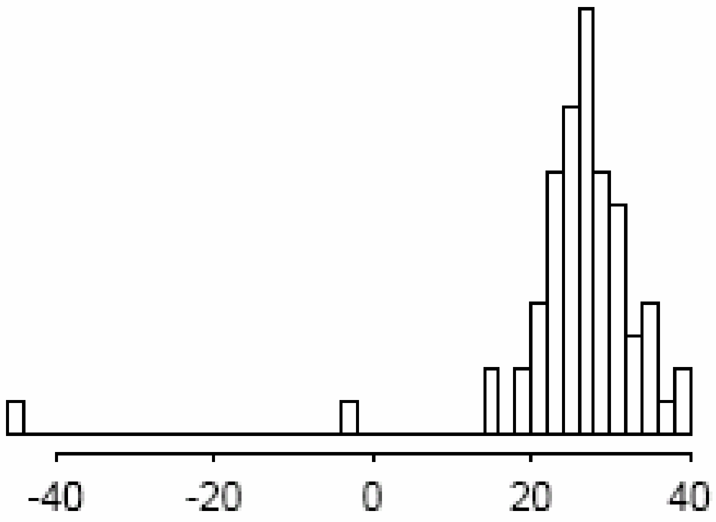


Model checking

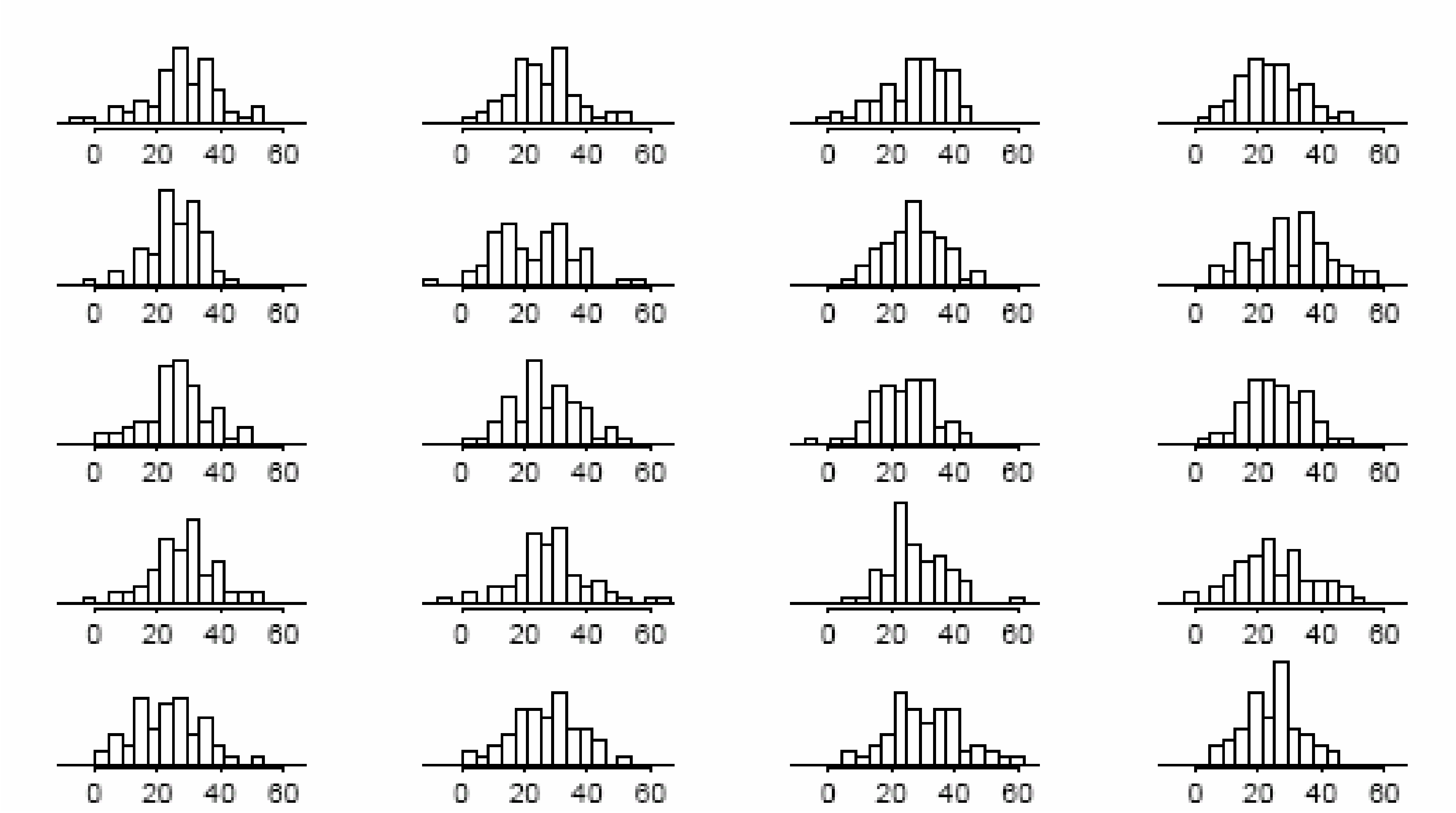
- Basic idea:
  - Display observed data (always a good idea anyway)
  - Simulate several replicated datasets from the estimated model
  - Display the replicated datasets and compare to the observed data
  - Comparison can be graphical or numerical
- Generalization of classical methods:
  - Hypothesis testing
  - Exploratory data analysis
- Crucial “safety valve” in Bayesian data analysis

Model checking: simple example

- A normal distribution is fit to the following data:



20 replicated datasets under the model



Comparison using a numerical test statistic

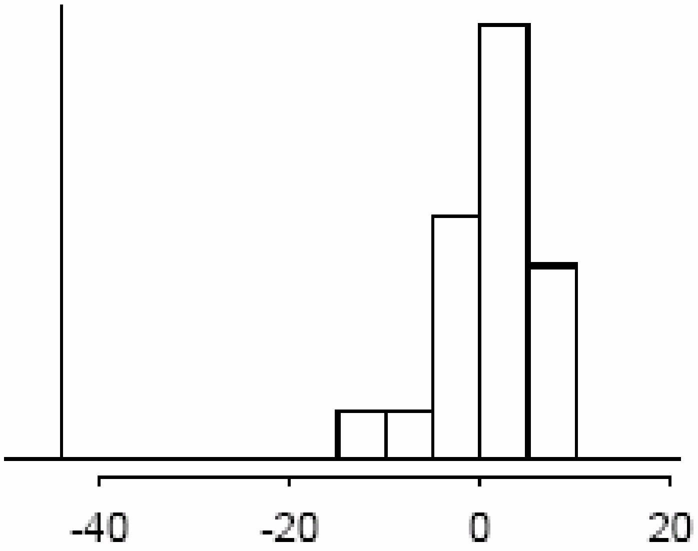
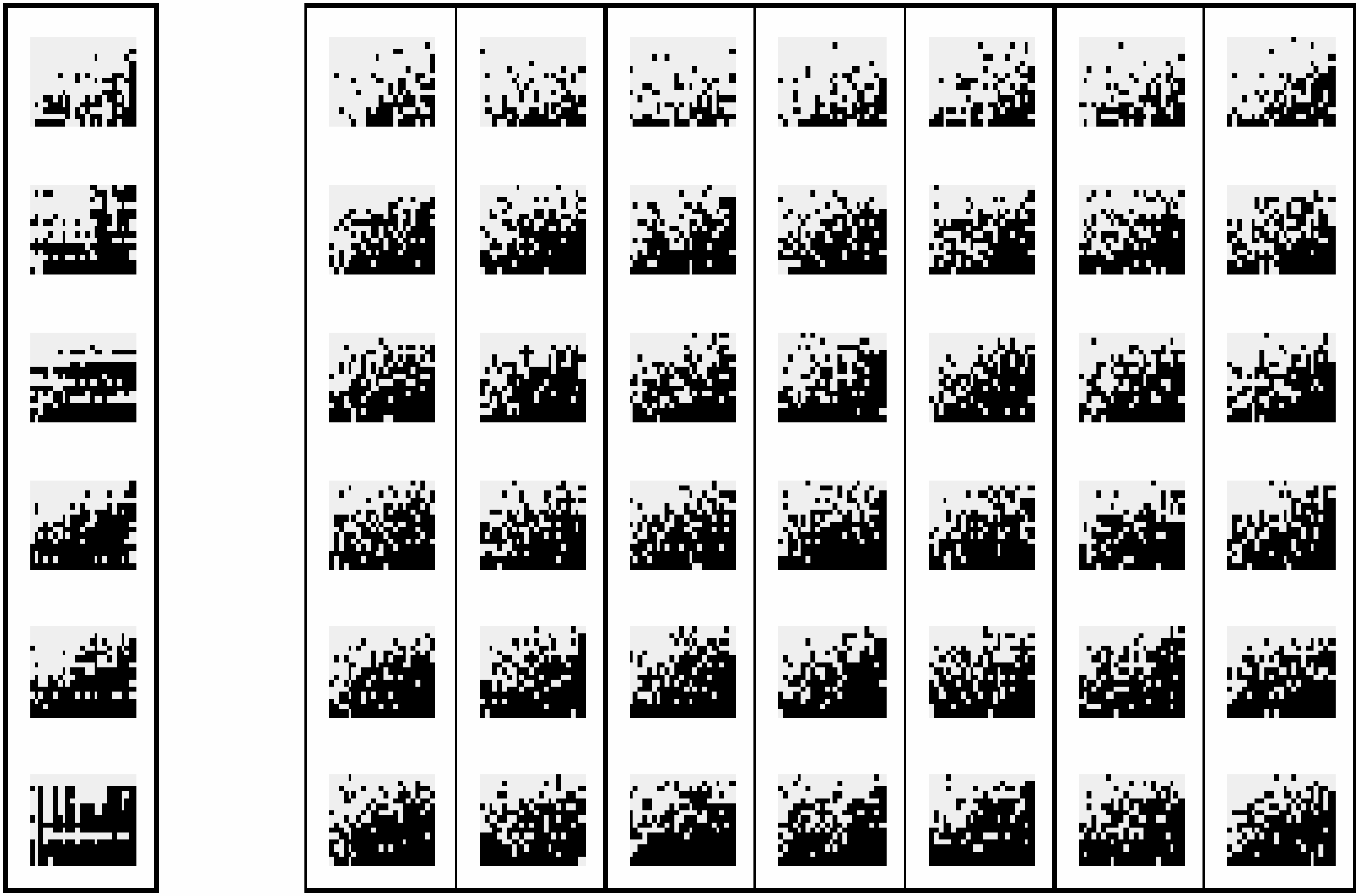


Figure 6.3 *Smallest observation of Newcomb's speed of light data (the vertical line at the left of the graph), compared to the smallest observations from each of the 20 posterior predictive simulated datasets displayed in Figure 6.2.*

Model checking: another example

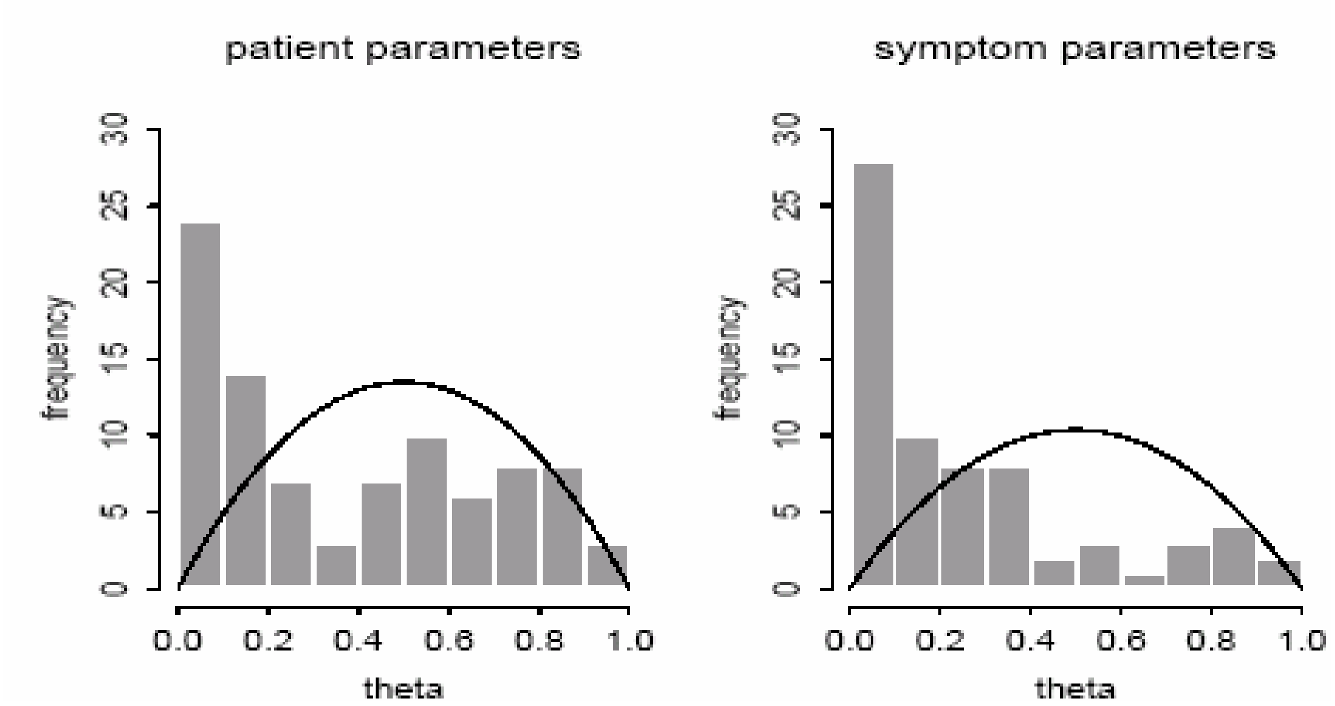
- Logistic regression model fit to data from psychology: a 15 x 23 array of responses for each of 6 persons
- Next slide shows observed data (at left) and 7 replicated datasets (at right)

Observed and replicated datasets



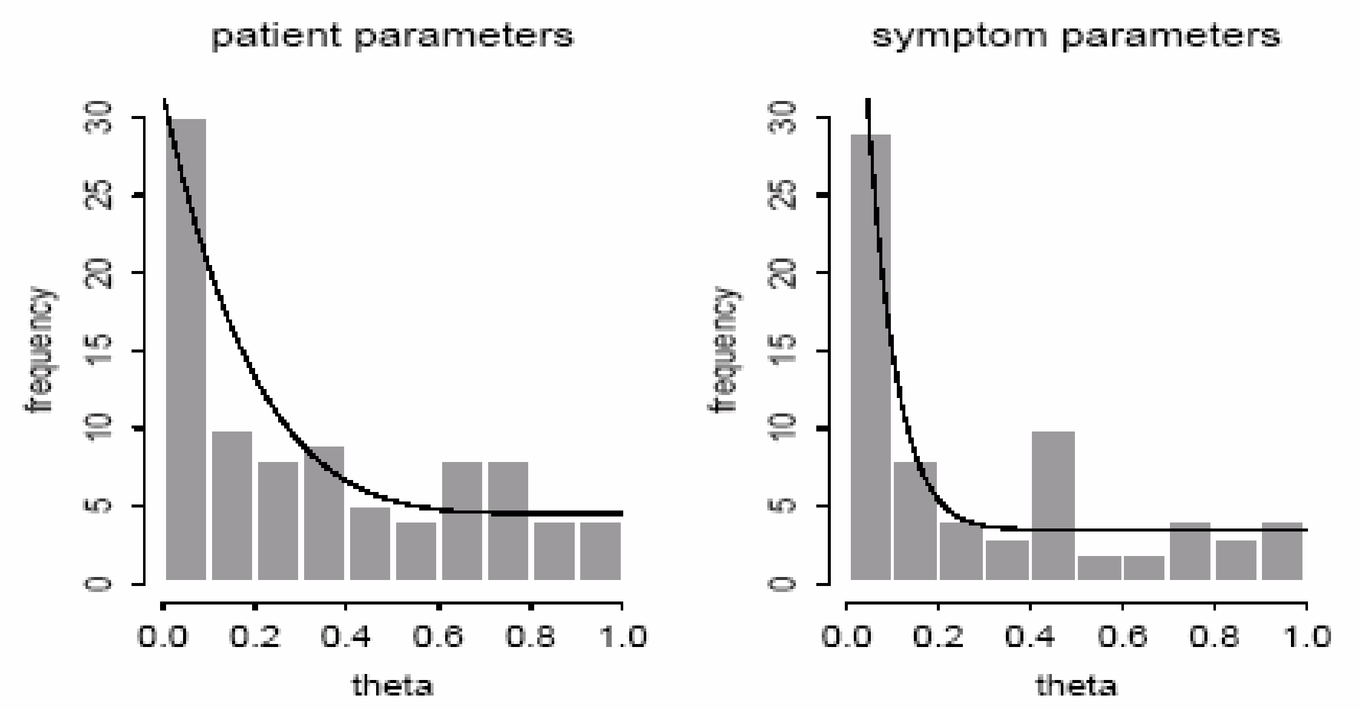
Another example: checking the fit of prior distributions

- Curves show priors for 2 sets of parameters; histograms show estimates for each set



Improve the model, try again!

- Curves show new priors; histograms show new parameter estimates



Summary

- Bayesian data analysis is about modeling
- **Not** an optimization problem; no “loss function”
- Make a (necessarily) false set of assumptions, then work them hard
- Complex models --> complex inferences --> lots of places to check model fit
- Prior distributions are usually not “subjective” and do not represent “belief”
- Models are applied repeatedly (“beta-testing”)

Some open problems in Bayesian data analysis

- Complex, highly structured models
  - (polling example) interactions between states and demographic predictors, modeling changes over time
  - Need reasonable classes of hierarchical models (with “just enough” structure)
- Computation
  - Algorithms walk through high-dimensional parameter space
  - Key idea: link to computations of simpler models (“multiscale computation”, “simulated tempering”)
- Model checking
  - Automatic graphical displays
  - Estimation of out-of-sample predictive error

Special thanks

- The examples were done in collaboration with Phillip Price, Thomas Little, David Park, Joseph Bafumi, Ginger Chew, Michael Shnaidman, Iven Van Mechelen, and Michel Meulders