Environmental Data and How They Speak

Michael J. Messner, PhD

U.S. Environmental Protection Agency - retired

Outline

- Flavors of This and That
  - Flavors of Environmental Data
  - Statistical Flavors
- THE Key to Understanding Statistics (!?!)
- Environmental Applications With:
  - Presence/Absence Data
  - Sample from Normal Population
  - Sample from Lognormal Population
  - Censored Data

Flavors of This and That

- Ice Cream – comes in many flavors. The only one I didn’t like (until recently) was coffee. My favorite is peanut butter fudge (swirled in either vanilla or chocolate).
- Pies – come in many flavors, too. Not wild about mincemeat, but enjoy any fruit pie. My favorite is apple crumb.
- Pie with Ice Cream – Apple pie with chocolate ice cream is fine, but apple pie with peanut butter fudge isn’t.
- Statistics lectures – Simple or complex, theoretical or applied, Bayesian or classical (frequentist)

Flavors of This and That

- Ice Cream – comes in many flavors. The only one I didn’t like (until recently) was coffee. My favorite is peanut butter fudge (swirled in either vanilla or chocolate).
- Pies – come in many flavors, too. Not wild about mincemeat, but enjoy any fruit pie. My favorite is apple crumb.
- Pie with Ice Cream – Apple pie with chocolate ice cream is fine, but apple pie with peanut butter fudge isn’t.
- Statistics lectures – Simple or complex, theoretical or applied, Bayesian or classical (frequentist)

Flavors of Environmental Data

- Flavors of Environmental Data:
  - Discrete
    - Point counts of asbestos in insulation material (of points examined, the percentage identified as asbestos)
    - Counts of *Cryptosporidium parvum* recovered from a volume of water
    - Number of experts deciding a contaminant belongs on a candidate list, per number of experts reviewing the contaminant
    - Of samples tested for total coliform or *e. coli*, the percentage testing positive
  - Continuous
    - Instrument calibration / linear regression
    - Measured concentration of contaminant per unit mass or volume assayed
    - Concentration reported only as “less than” some censoring limit
    - Concentration measured over time
  - Other
    - Hot spot detection
    - Microbial and chemical dose-response modeling

By the way,

- By the way, **the best thing about the field of environmental statistics is the great variety of applications**.
  - There's always something new and challenging.
  - It never gets old.
- **The greatest challenge for the environmental statistician is getting involved up front** - in planning the project that produces data. In planning, we ask key questions like "What do you wish to estimate?" or "What hypothesis do you wish to test?" And follow-up questions like "How much error (in your estimate or hypothesis test) can you live with?" **Too often, we're called when it is too late.**

Statistical Flavors

- Google search results were interesting:
  - I did a Google search for “What **are** statistics?” and got the Oxford Languages answer to “What **is** statistics?”: “the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a while from those in a representative sample.”
  - I searched for “What **is** statistics?” and got this from the UCI Statistics Department: “Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.”
- The Google search also returned (under “People also ask”) “What are the 3 types of statistics?” and the answer:
  - Descriptive statistics.
  - Inferential statistics.

I think I can do better.

Statistics (singular) and Statistics (plural)

- Statistics (plural) are numbers derived from data, often to summarize features of the data. For example, if you were to roll a pair of dice 10 times, the observed values comprise your sample (of size 10) and you might report that the total over the 10 rolls was 75. That's a statistic. You might report that the minimum was 3 (another statistic) and that the range was 9 (yet another).

- Statistics (singular) is the discipline, practice, or science that we bring to bear in understanding what data (and the statistics derived from data) are trying to tell us.

The Key!

Barb Forsyth, teaching graduate level statistics at Ohio State: “If there is only one thing you remember from this course, let it be this”:

Statistics are random variables!

THE Key to Understanding Statistics!!!

- Data are random variables. Until they’re observed, think of them as being drawn from some parent distribution.
- Statistics are random variables, too. *Because they’re derived from data, how could they not be!*

And after you observe the data?

- At that point, the data are the data.
  - No longer random variables.
  - Once “realized”, they’re fixed.
- The data are ready to speak.
  - They can tell us about how they were generated.
  - This requires knowledge or assumptions about their statistical model.
  - They speak one way to classical (a.k.a. frequentist) statisticians and another way to Bayesian statisticians. **Statisticians come in different flavors!**
  - They speak through something called “likelihood”
    - Classical statistician works to understand likelihood of observing data more extreme than what was observed, as a function of model parameters.
    - Bayesian statistician works to understand likelihood of the data, as a function of model parameters. Because this statistician realizes the data are fixed, she/he thinks of this as the likelihood of the model parameters.

Simple Example

- Spin a coin many times and observe the outcomes (heads or tails). Repeat and observe the outcomes.

- Environmental Problem #1: Test each of many 100 ml volumes of drinking water for presence of total coliform (TC) bacteria. A city might test 100 volumes each month. For this example, 100 are tested and 2 are found to be TC positive.

- The model: Each test (or coin spin) is TC positive (heads) with shared probability P.
- When the data “speak”, they tell us about parameter P.

Simple Example, continued

- Data: **2** of 100 volumes were TC positive and **98** were negative.
- A good point estimate of the positive rate is P = 2/100 = 0.02 = 2%
- Likelihood = probability of 2 positive out of 100 $\propto P^2 * (1-P)^{98}$
- Likelihood is a binomial probability mass function.
- The Bayesian result is called a posterior distribution.
  - Posterior ∝ Prior * Likelihood
- Prior is what you know beforehand. Likelihood is what the data speak.
- For this kind of problem, my prior might be uniform over the range from 0 to 1. This is a weak prior. The data will speak much more loudly than this prior. This prior is the same as a Beat distribution with parameters 1 and 1
- For these data, the posterior is the Beta distribution with parameters 1 + **2** **and** 1 + **98**).

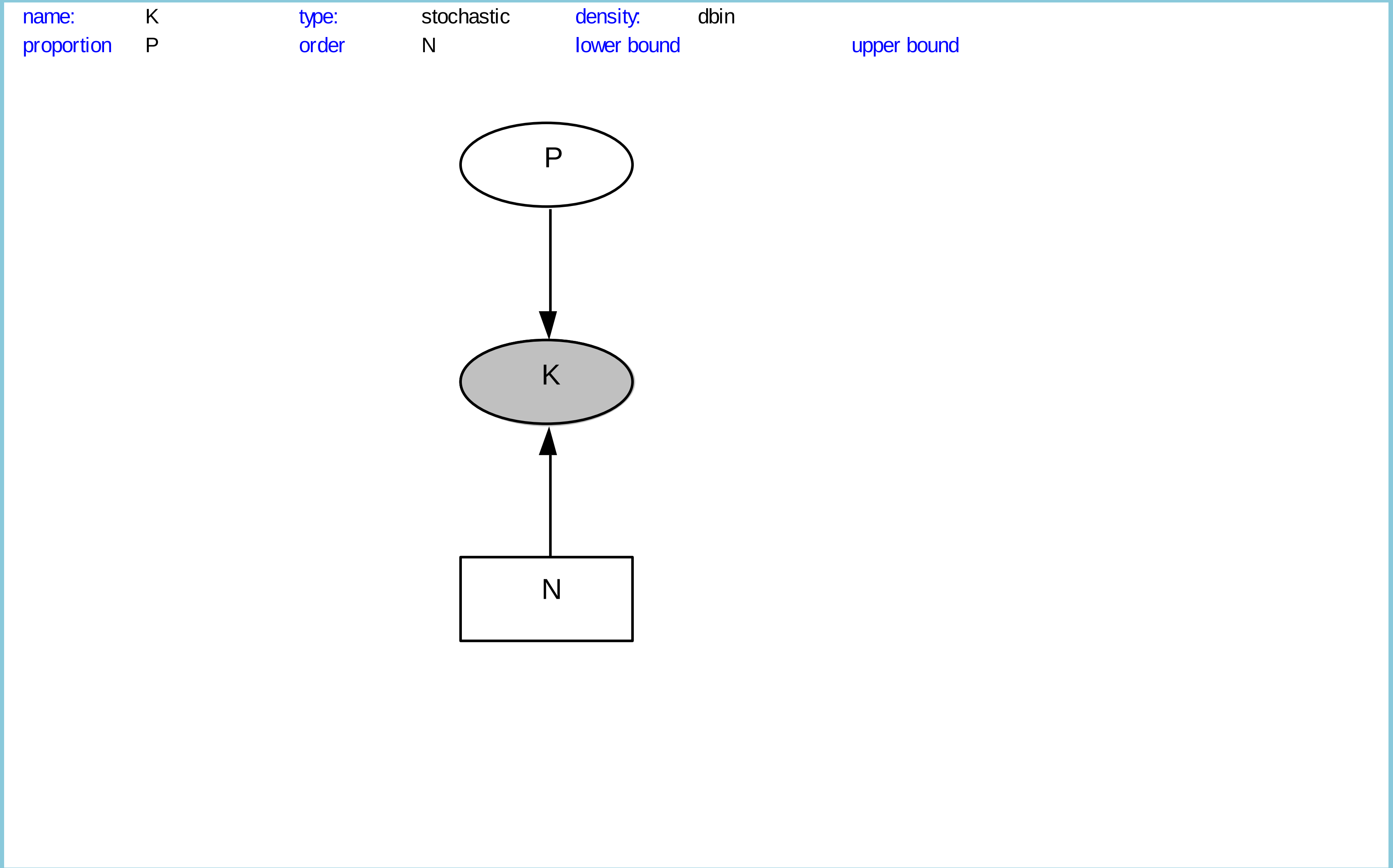
Why start with this example?

- This is exactly the kind of problem solved by Thomas Bayes, in his 1763 publication “An Essay towards solving a Problem in the Doctrine of Chances”. Bayes showed that a uniform prior leads to a posterior for success probability that is Beta distributed with parameters K + 1 and N – K + 1, where K is the number of successes and N – K is the number of failures.
- This example illustrates the kind of problem that could be solved until the Bayesian revolution that began ~ 35 years ago.
- I’ll use this problem to introduce software that helped fuel the Bayesian tidal wave that continues today.

OpenBUGS

- BUGS = “Bayesian Inference Using Gibbs Sampling”: Open-source software produced by the MRC Biostatistics Unit in Cambridge and Imperial College, London.
- From late 1990s, BUGS was THE software for doing Bayesian analysis.
- Although recently surpassed by other software (like STAN for multilevel modeling), OpenBUGS remains a great tool for teaching and learning Bayesian statistics.

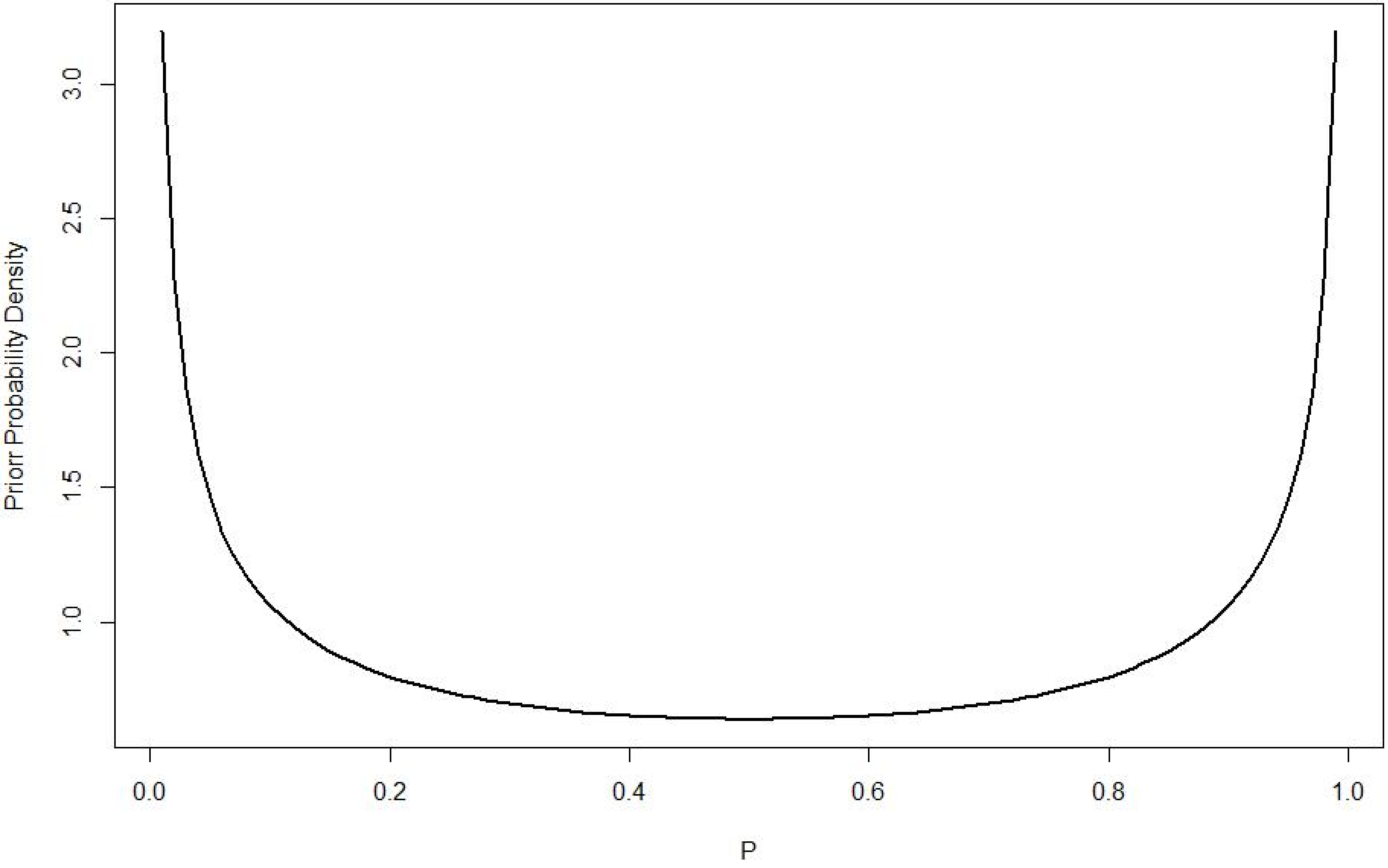
The Simple TC Problem with OpenBUGS



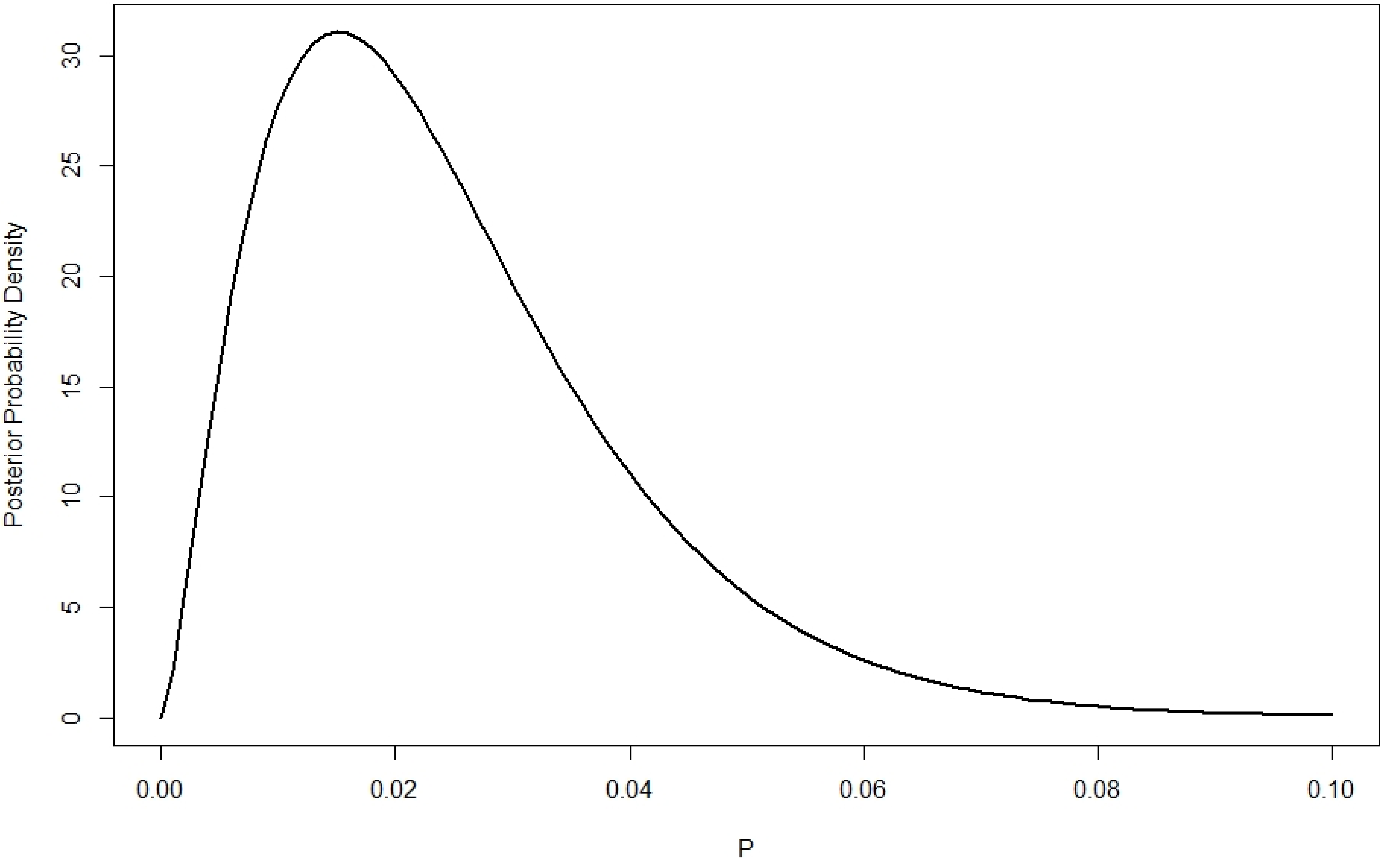
name: K type: stochastic density: dbin

proportion P order N lower bound upper bound

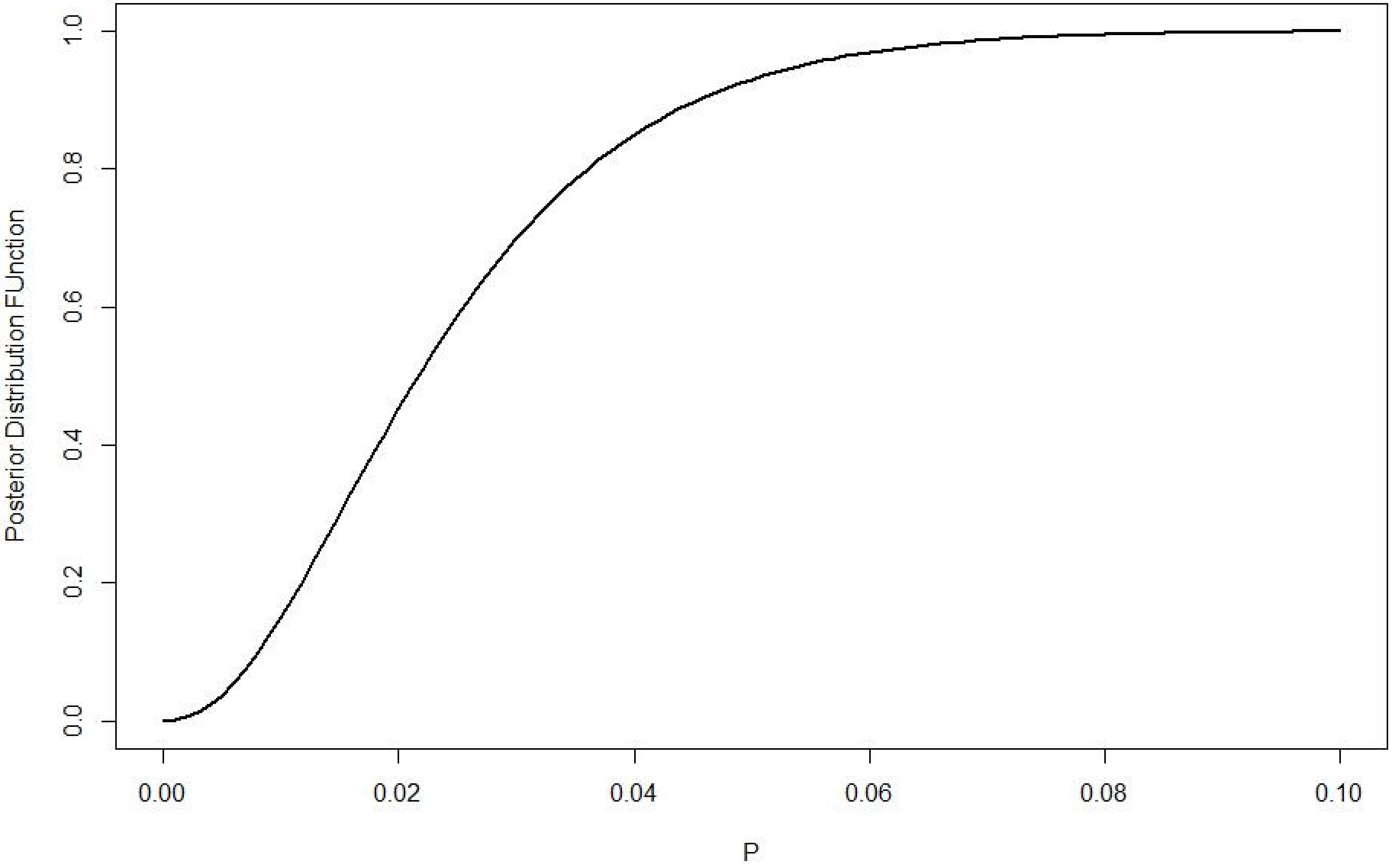
Prior = Beta(0.5, 0.5)



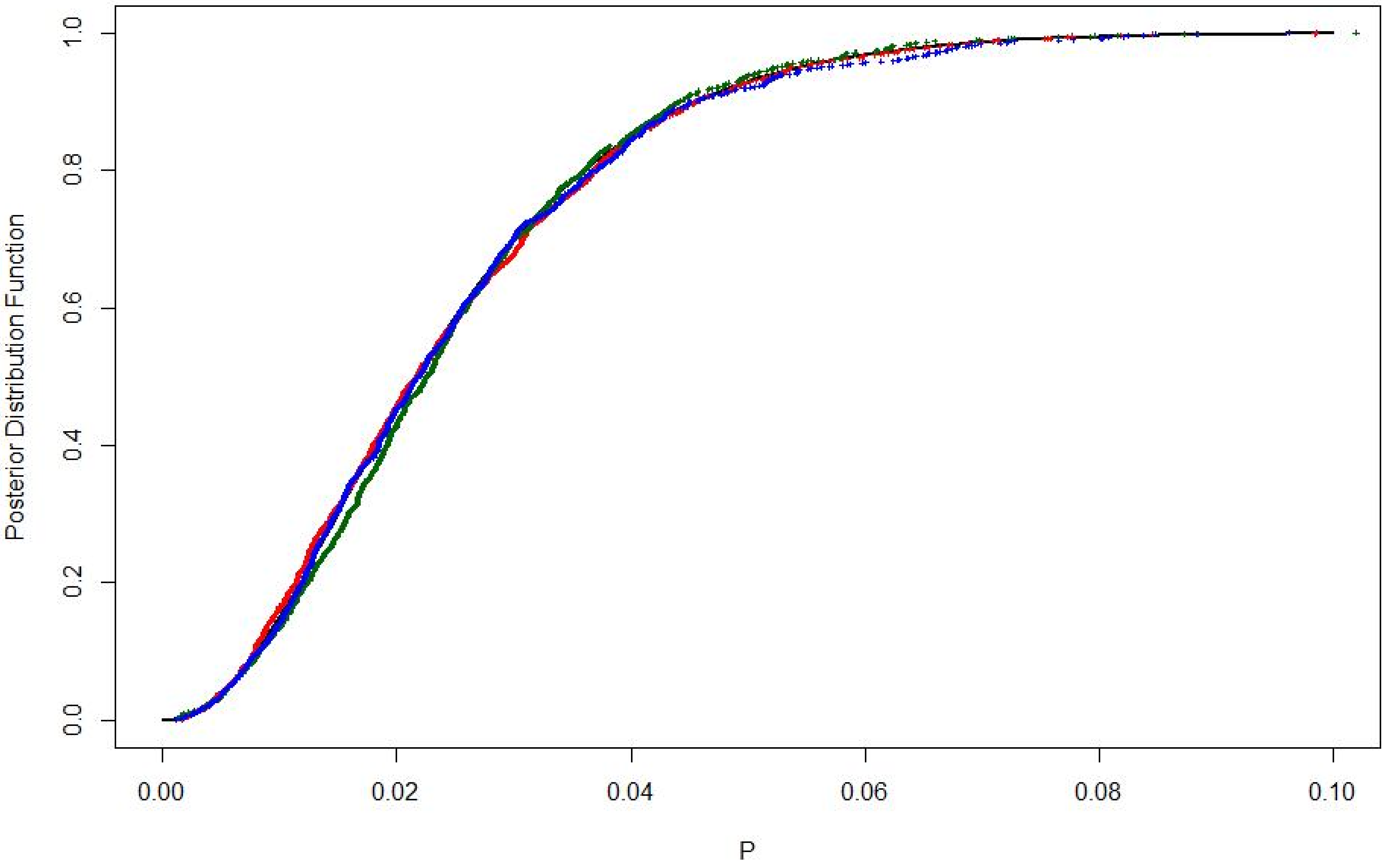
Posterior = Beta(2.5, 98.5)



Posterior = Beta(2.5, 98.5)



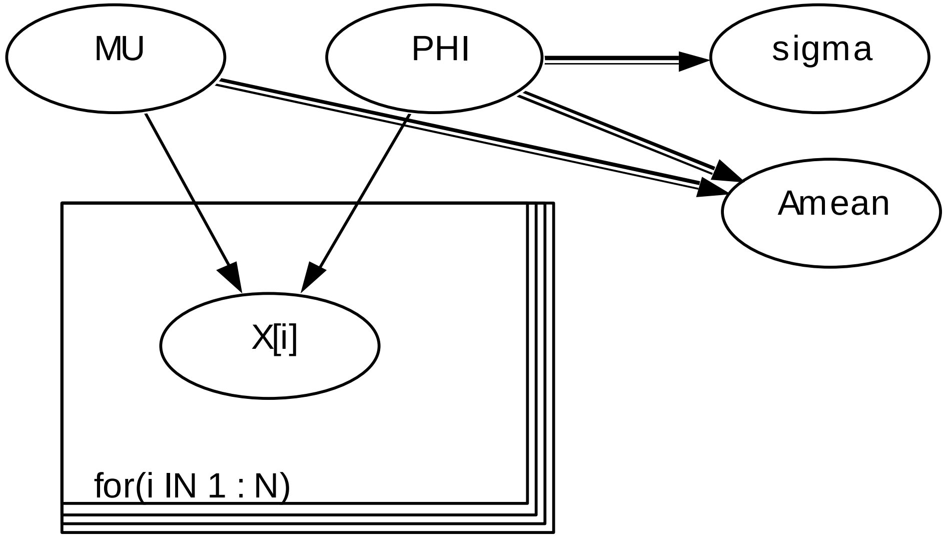
MCMC samples ~ Beta(2.5, 98.5)



Environmental Problem #2: Sampling the Lognormal Distribution

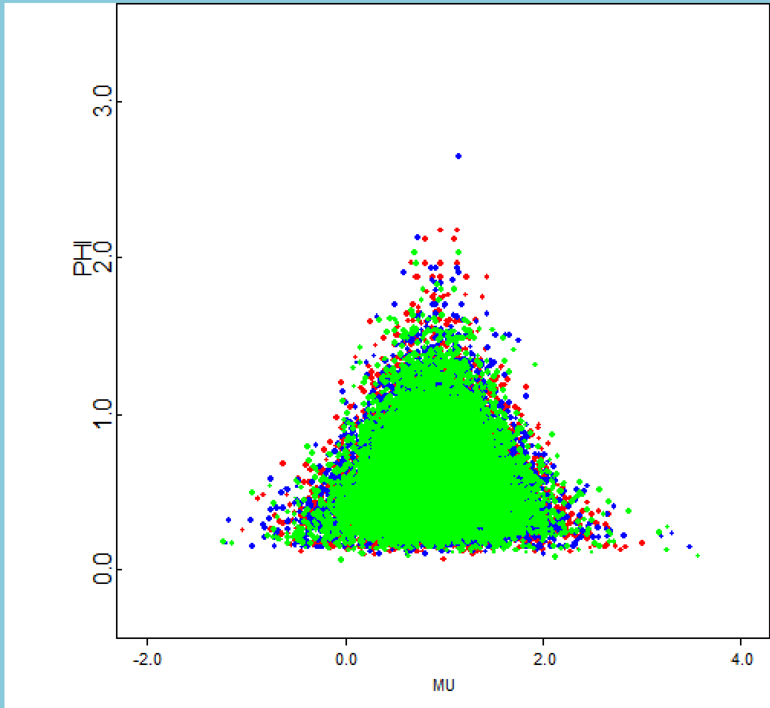
- Contaminant concentration data are constrained by zero and tend to be right-skewed. We often assume concentrations vary as lognormal random variables.
- Some examples:
  - Cryptosporidium concentration at intake to a drinking water treatment plant.
  - Lead concentration in the distribution system
- To illustrate, we'll grab a sample of size 10 from a lognormal distribution.
  - Mean concentration = 10 ppb
  - σ = standard deviation of log(Concentration) = 1.8
  - μ = mean of log(Concentration) = log(10) – σ^2 / 2 = 0.6826

OpenBUGS



Results

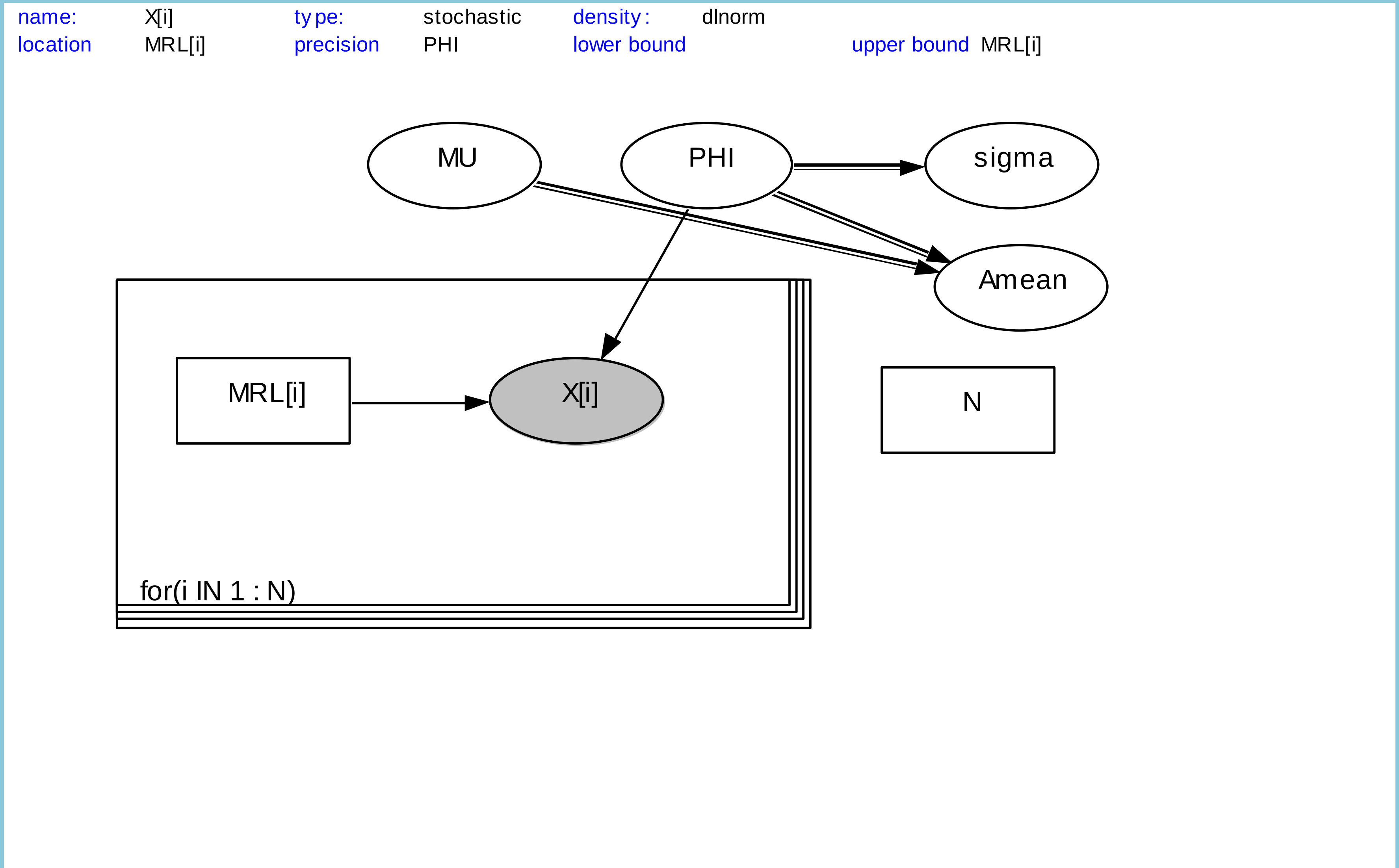
	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
Amean	10.88	108.4	0.6695	2.529	6.005	36.11	1	30000
MU	0.9206	0.4462	0.00252	0.02399	0.9203	1.818	1	30000
PHI	0.6184	0.2642	0.003536	0.2119	0.5845	1.233	1	30000
deviance	54.53	1.988	0.02159	52.59	53.92	59.9	1	30000
sigma	1.368	0.3257	0.004242	0.9008	1.308	2.173	1	30000



Environmental Problem #3: Lognormal with Censoring

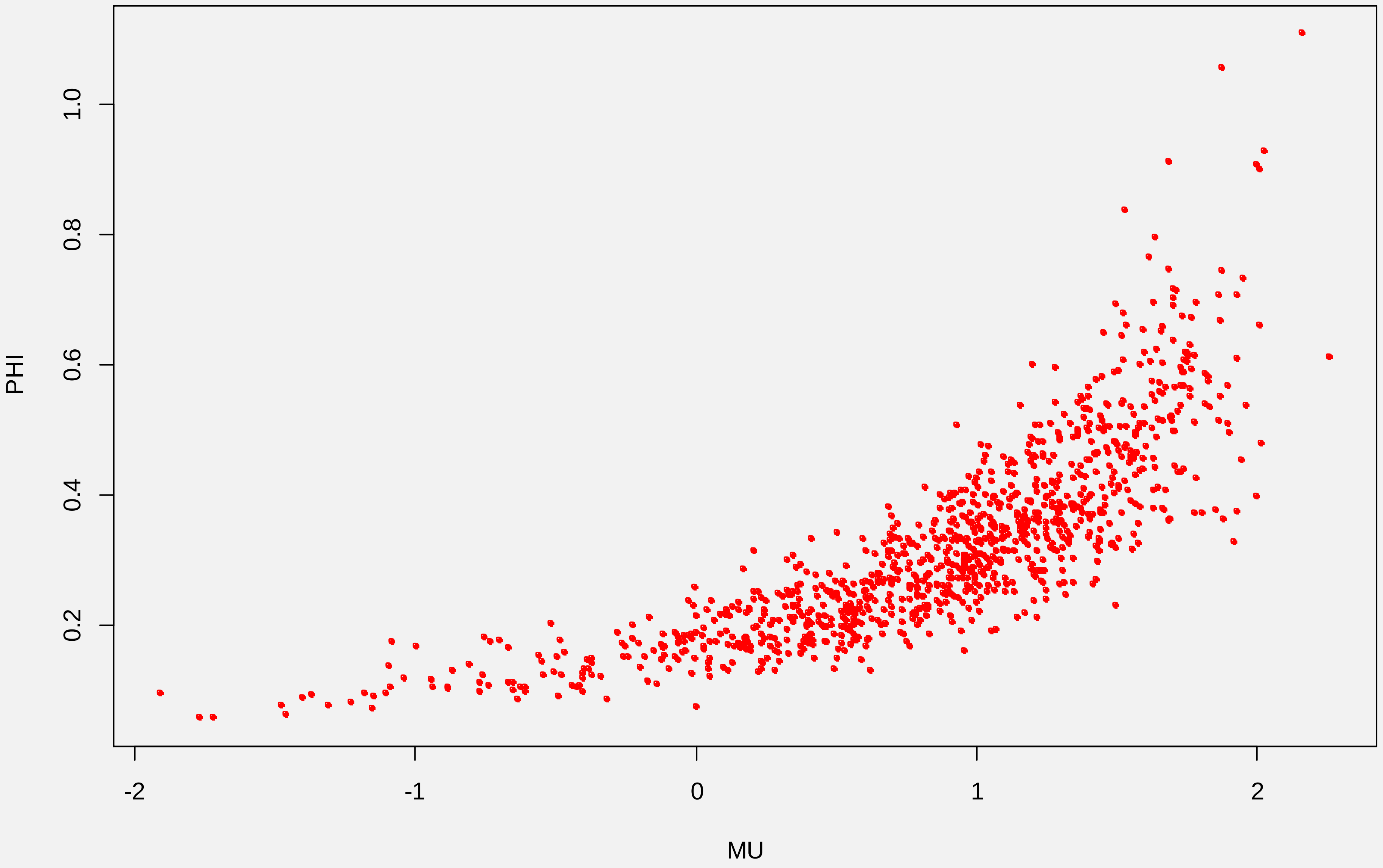
- In drinking water, contaminants of human health concern may be present at levels so low that they are difficult to detect.
- Current practice with drinking water: When contaminants are detected, but at levels that fall below minimum reporting limits (MRLs) the measured values are censored (hidden from end users) and simply reported as being less than MRLs. I worked unsuccessfully to change this practice in cases where my unit was a critical end user.
- This example: Sample of size 100 from lognormal distribution, with 87 censored and reported as “less than 20”. The other 13 were greater than 20 and not censored.

OpenBUGS



name:	X[i]	type:	stochastic	density:	dlnorm		
location	MRL[i]	precision	PHI	lower bound		upper bound	MRL[i]

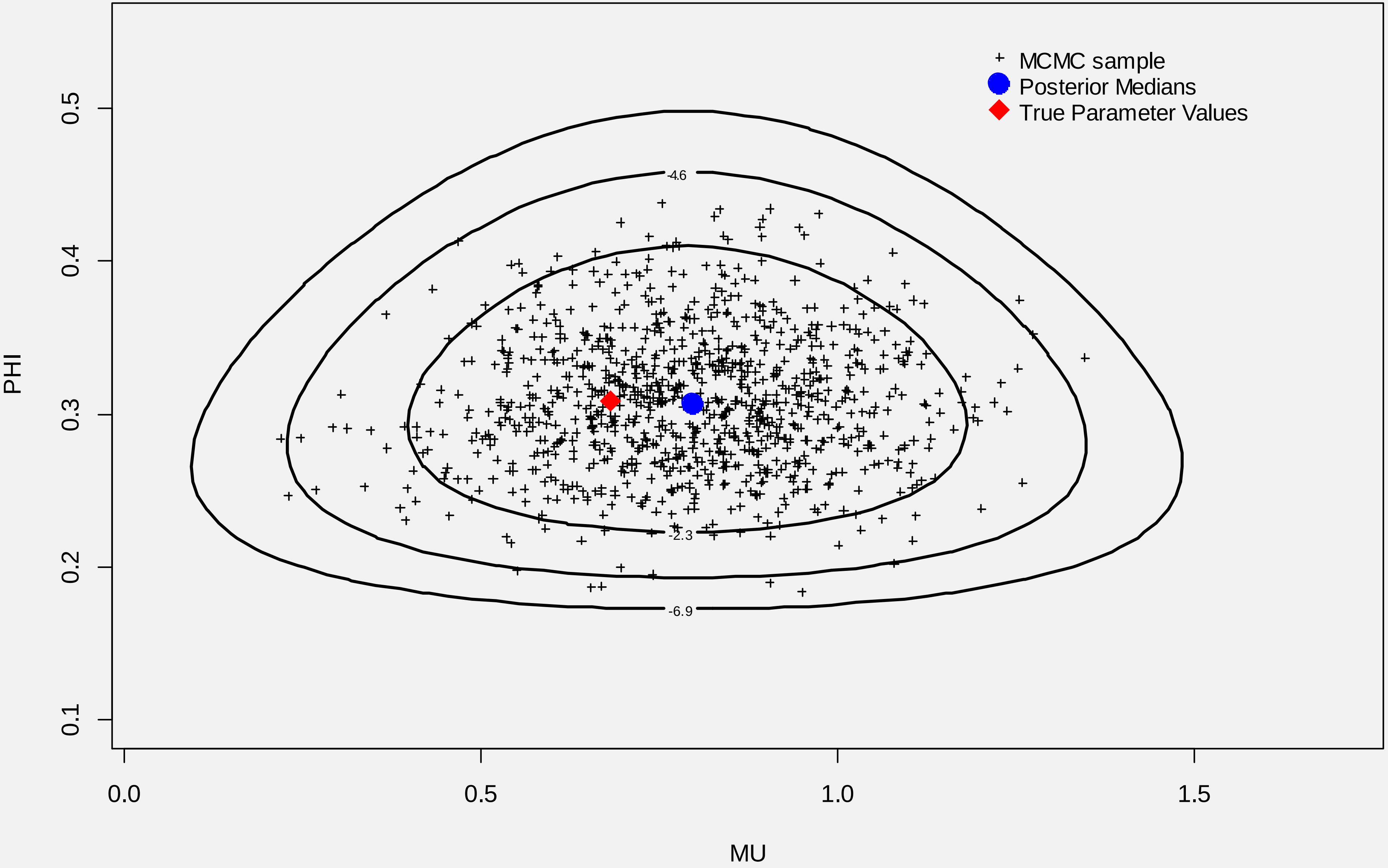
MCMC sample based on censored data

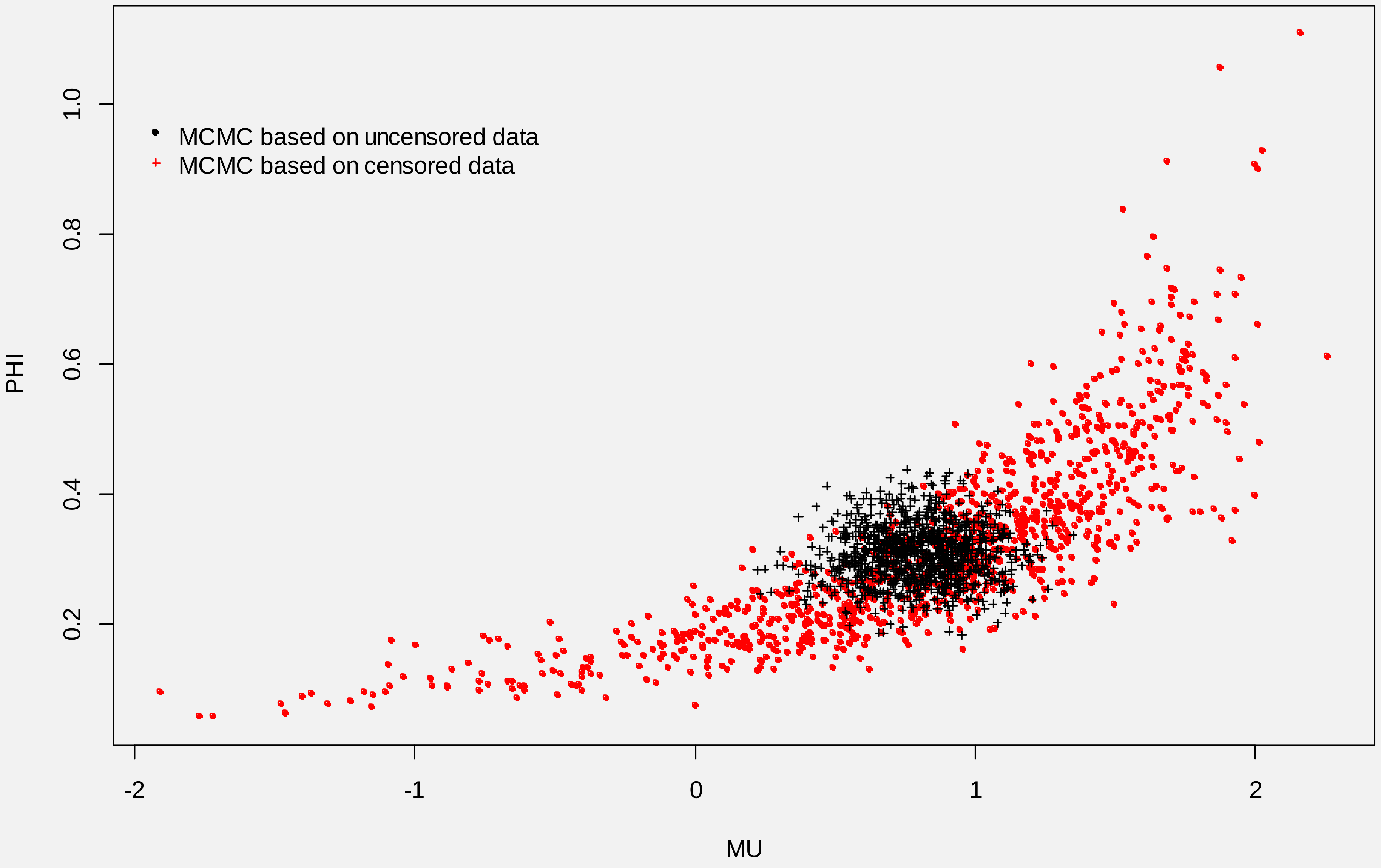


What if Data Weren’t Censored

- Because I’m working with simulated data, I can easily do this.
- With real environmental measurements, I would need to be concerned with measurement error, which tends to increase as concentration falls below the critical level for detection.
- I this case, I’ll use another good tool for learning Bayesian statistics: Jim Albert’s *LearnBayes* package (in R).
  - Function *mycontour()* produces a contour plot of the joint posterior distribution.
  - Function *simcontour()* generates an approximate MCMC sample from that posterior distribution.

Letting the Uncensored Data Speak

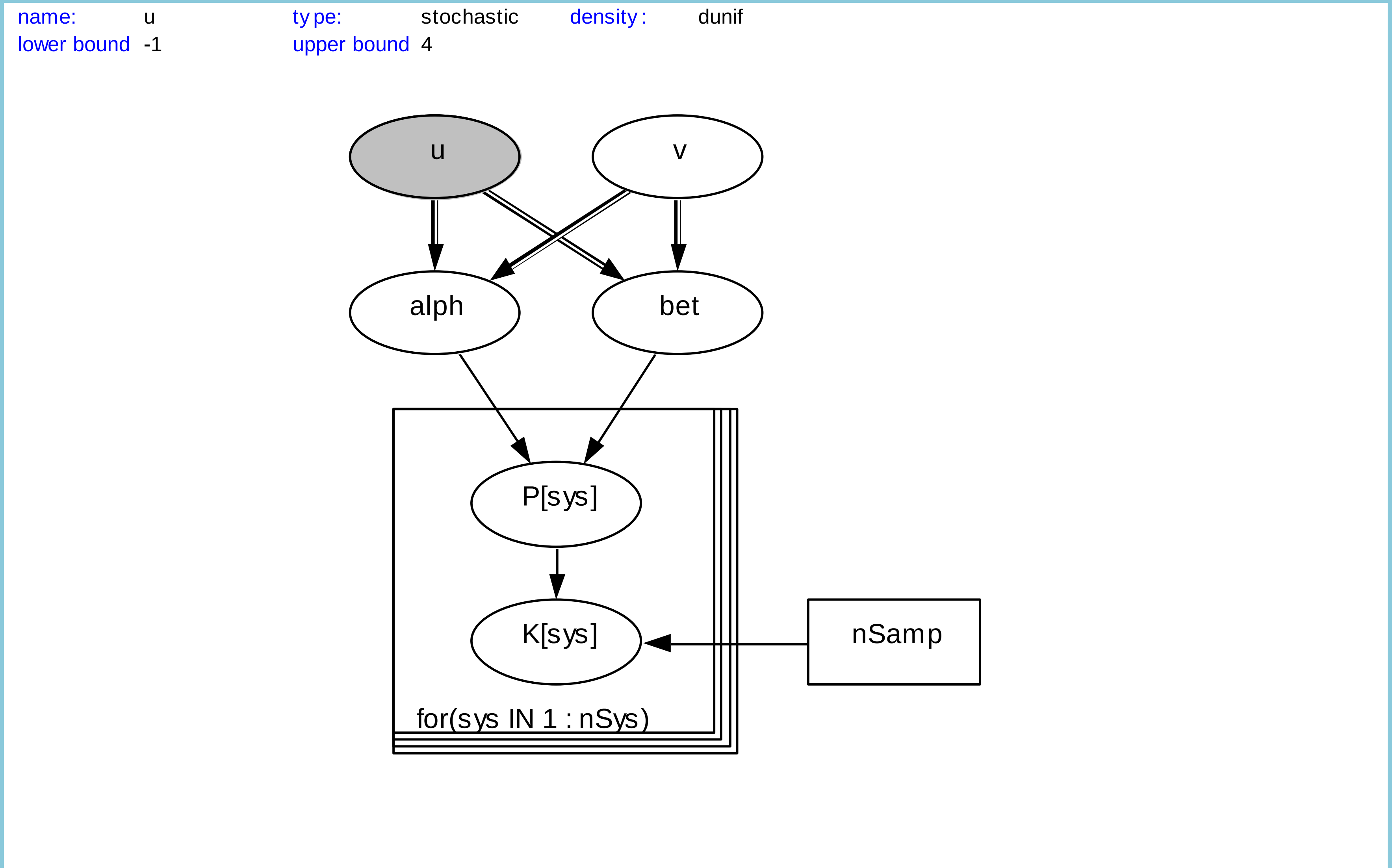



● MCMC based on uncensored data
+ MCMC based on censored data
PHI
0.2
0.4
0.6
0.8
1.0
MU
-2
-1
0
1
2

Environmental Problem #4: Multilevel Total Coliform

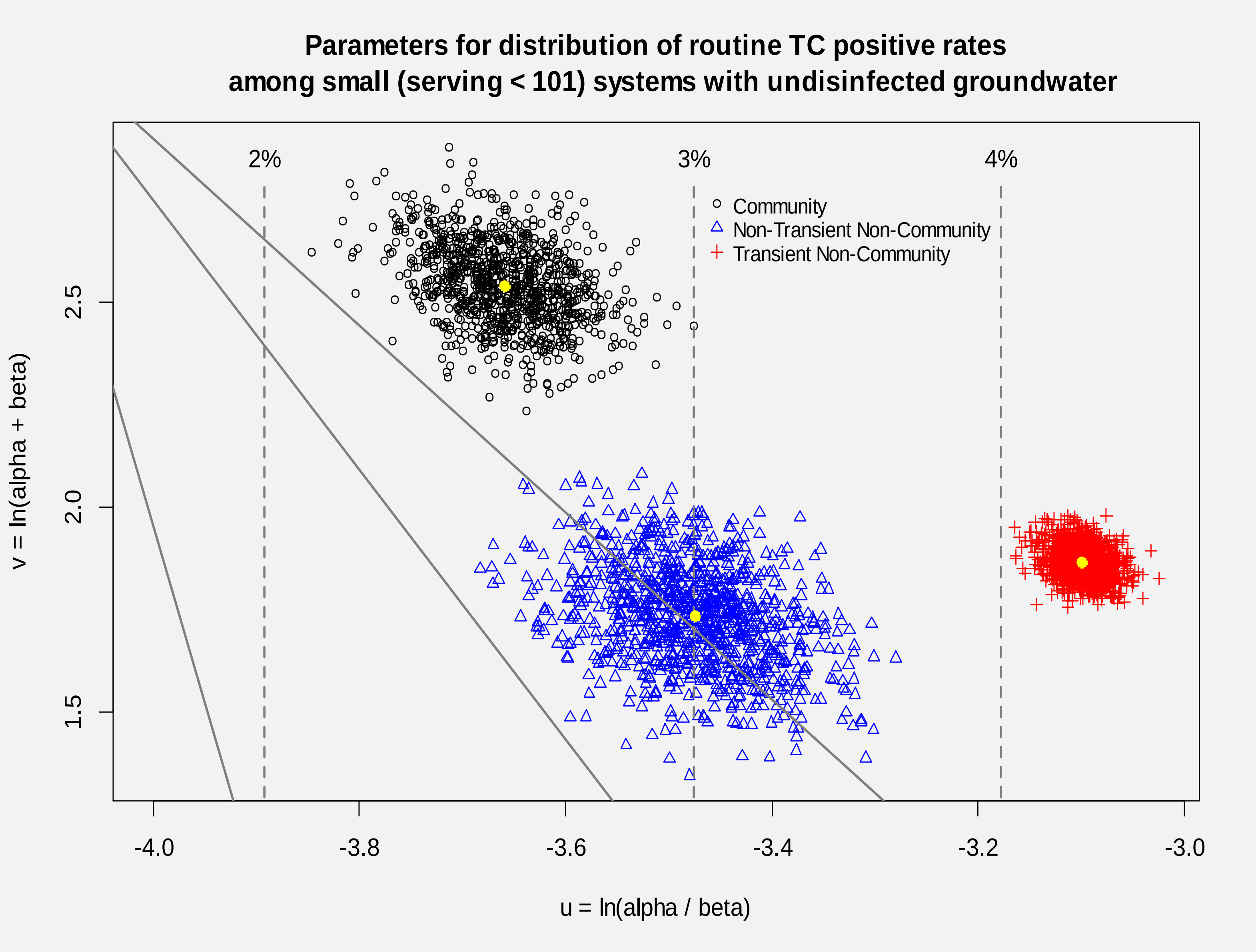
- This is an extension of the simple problem with one system.
- In this simulated data set, a sample of nSys = 200 public water systems conduct one TC assay each month for a year. At the end of the year, we have the number positive (of 12) for each system.
- The sample of systems is drawn from a larger population of systems, all of similar size, type and disinfection practice.
- We use the data to estimate the parent distribution of positive rates.

OpenBUGS



name: u type: stochastic density: dunif
lower bound -1 upper bound 4

From our paper:



Why give Bayes a try?

- Puts the analyst’s focus where it belongs: on likelihood and modeling.
- Allows inclusion of prior information.
- Simplifies addition of model complexities, like censoring and hierarchical structure.
- Results are easy to understand and interpret probabilistically.
- Output can directly inform decision makers and decision-analytic modeling.

Final Example (not environmental)

Monte Hall & Let’s Make a Deal, from Marilyn vos Savant’s 9/9/1990 Parade Magazine column.

1. You (the game show contestant) pick door #1, knowing that only one of the three doors hides the grand prize of a new car.
2. Game show host Monte Hall knows that two doors hide only goats, so he’ll can always shoe you one in a door you didn’t select. He wants to put the pressure on you, so shows you that door #3 hid a goat.
3. Monte now asks if you want to pick door #2.
4. Should you switch your selection from door #1 to door #2?

To obtain files used in this presentation:

Send email to Mike Messner (mjaymessner@gmail.com), with the subject line “file request”.

References

- Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique and future directions (with discussion), *Statistics in Medicine* **28**: 3049-3082.
- Albert, J.H., Bayesian Computation with R (2nd edition), Springer-Verlag (UseR! Series), New York, 2009.
- Messner M.J., Berger P., Javier J. Total coliforms and *E. coli* in public water systems using undisinfected groundwater in the United States. Int. J. Hyg Environ. Health. 2017;220(4):736–743.

Highly recommended reading: Peter Lee’s preface to the first edition of his book Bayesian Statistics (https://www.york.ac.uk/depts/maths/histstat/pml1/bayes/).