Assessing Interval Estimation Methods for Hill Model Parameters in a High-Throughput Screening Context

Dustin F. Kapraun[1], Eric D. Watt[1, 2], R. Woodrow Setzer[1], Richard S. Judson[1]

[1]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 27711

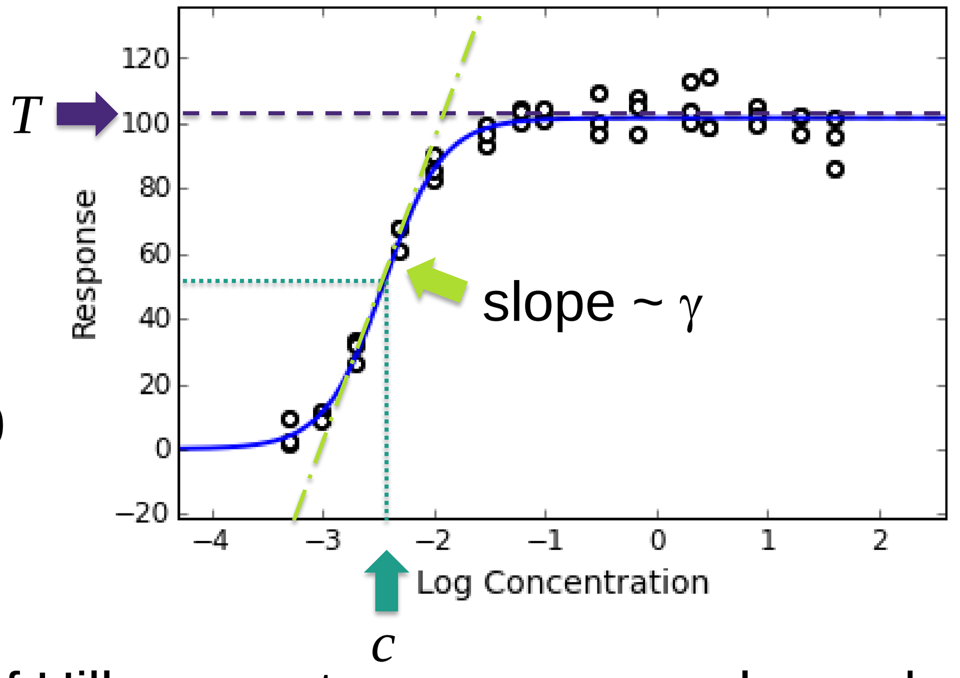
[2]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, 37831

Dustin F. Kapraun I kapraun.dustin@epa.gov I ORCID 0000-0001-5570-6383

Background

- The **Hill model** relates **response** y (of an assay or organism) to **concentration** x (of a chemical):

$$y = f(x; q) = \frac{T}{1 + 10^{\gamma(c-x)}}$$

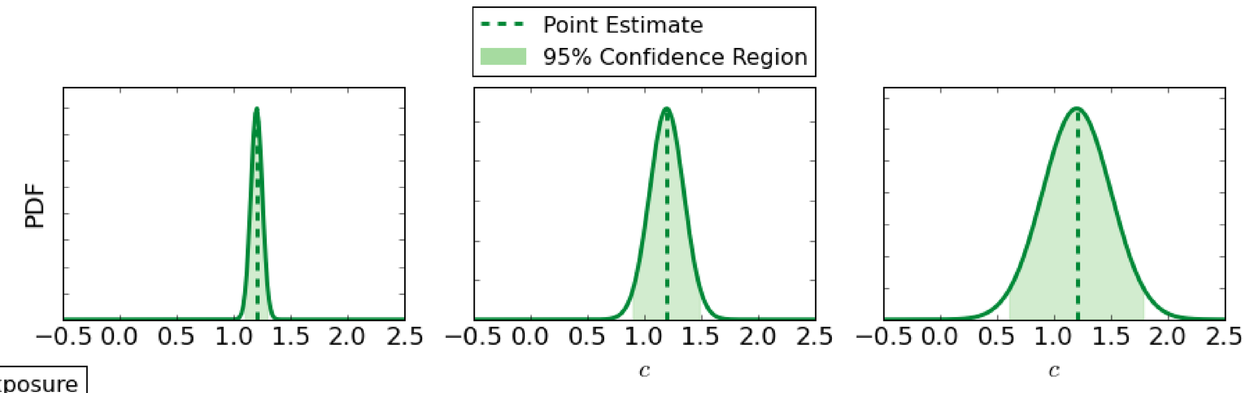


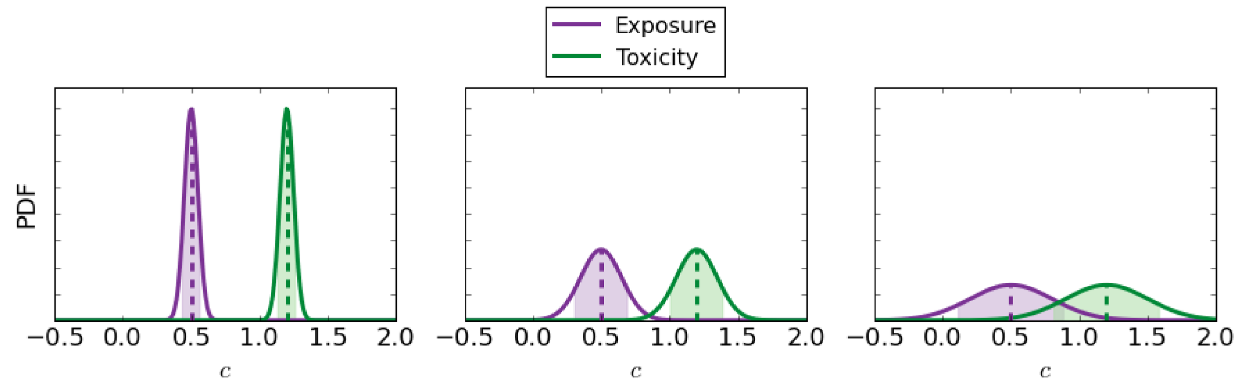
- Hill model **parameters** $q = (T, c, \gamma)$ provide useful metrics of **efficacy** (T), **potency** (c), and ligand binding **cooperativity** (γ).
- **Point estimates** (single values) of Hill parameters are commonly used in **high-throughput screening (HTS)** risk assessment.
- Point estimates **lack information** about precision (or confidence).
- **Interval estimates**, such as those produced by
  1) **Asymptotic theory** (Wald type),
  2) **Bootstrapping** (two different methods), and
  3) **Bayesian approach** (MCMC),

  provide information about **precision**.
- **Goal: Determine which of four interval estimation methods gives the most accurate information about the *precision* of toxicity parameter estimates; e.g., estimates of the AC_{50} (c).**

Motivation: Why interval estimation?

Interval estimates provide information about precision.



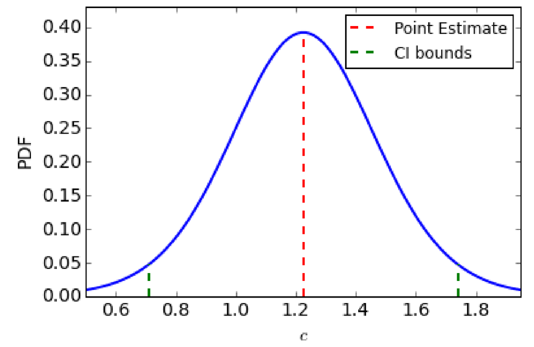


Interval estimates **allow for more informative comparisons** with other uncertain quantities.

Interval Estimation Methods

1) Asymptotic Theory (Wald)

- Hessian of log-likelihood function → FIM → **covariance matrix**.
- Diagonal entries of the covariance matrix are the **standard errors** for the parameter estimates.
- As $n \to \infty$, parameter distributions → **Gaussian**.

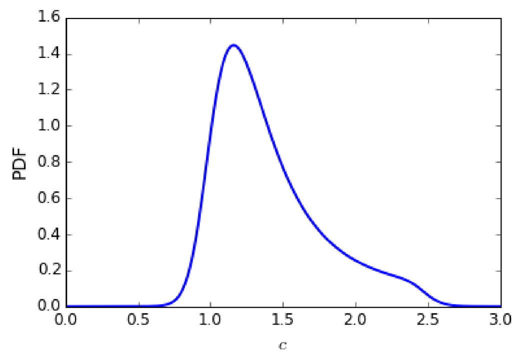


2) Bootstrapping (Boot & Wild)

- Refers to a procedure that relies on **sampling with replacement**.
- In the case of parameter estimation, sample **residuals**.
- Use the sampled residuals to create a **new data set**, then find a **new best fit model** and corresponding parameter values.
- Repeating this process many times, we obtain a **probability distribution** for the model parameters.
- For "Wild" method, keep residuals paired with concentration.

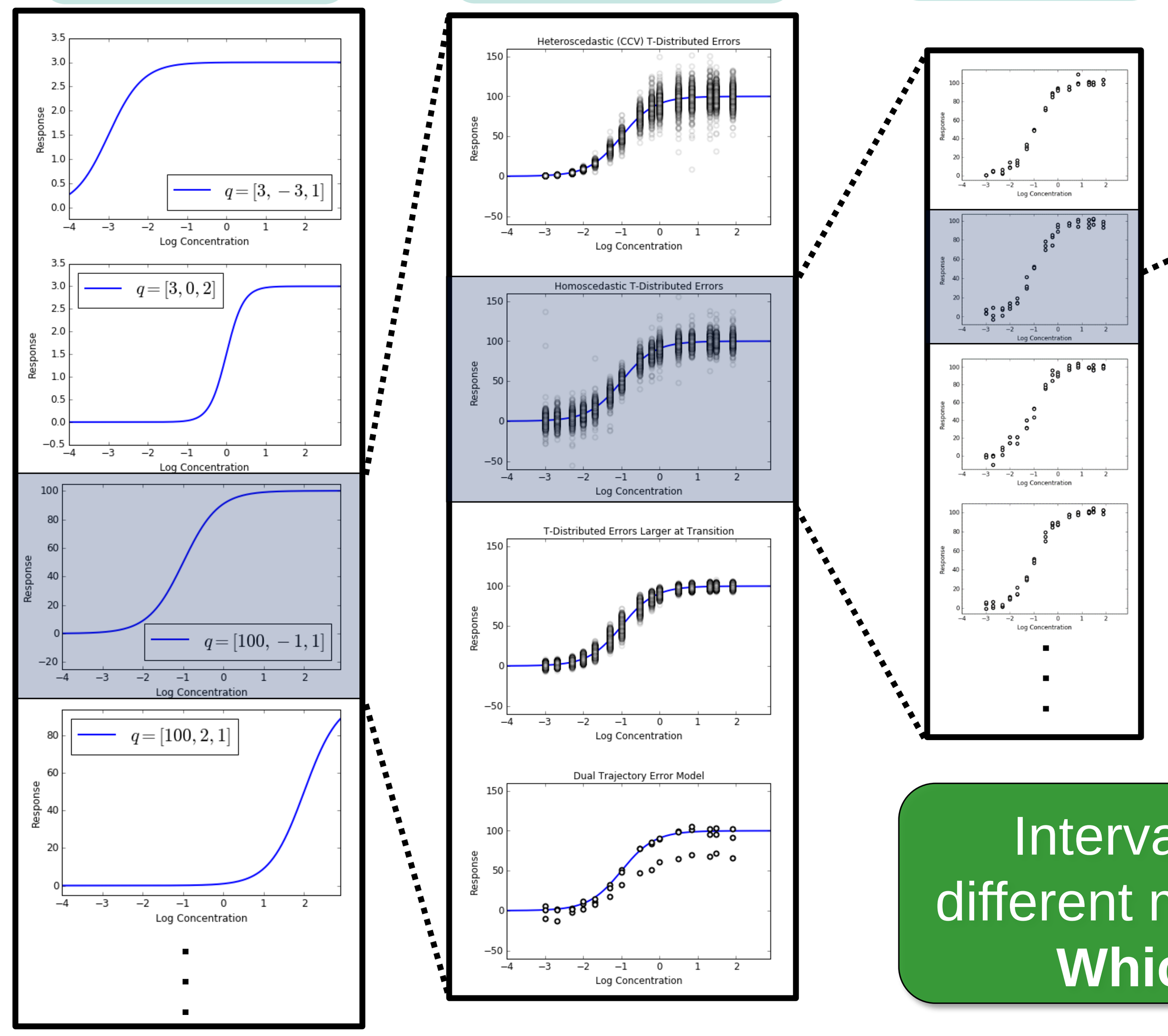
3) Bayesian Approach (Bayes)

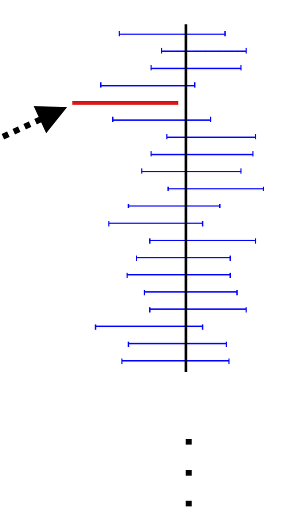
- Parameters are considered to be **random variables** rather than fixed quantities.
- We seek the **posterior density** that best reflects the distribution of parameter values based on the sampled observations.



Evaluating Interval Estimates: Simulation Study

Choose "true" parameter vectors. (Consider one at a time.) → Define several error models. (Consider one at a time.) → Generate many data sets (~1000). → Construct CIs using various methods. → Compare actual to nominal coverage.





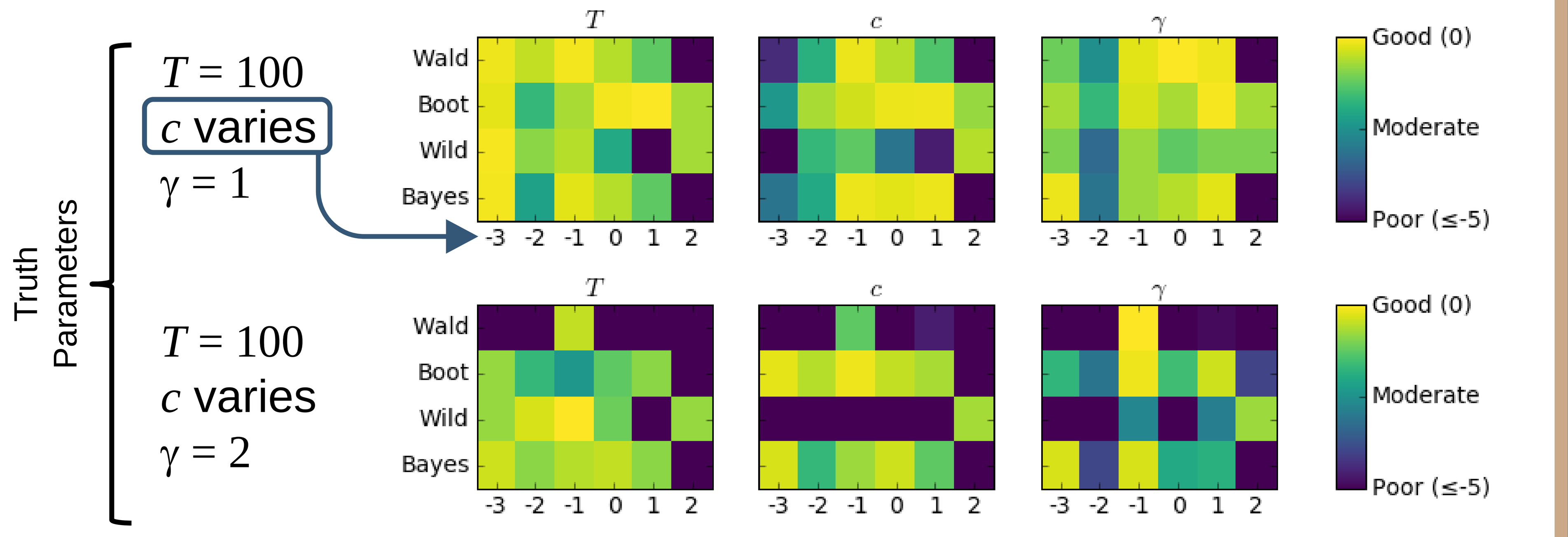
95%
87%
96%
74%
⋮

Interval estimates produced by different methods generally disagree. Which is method is best?

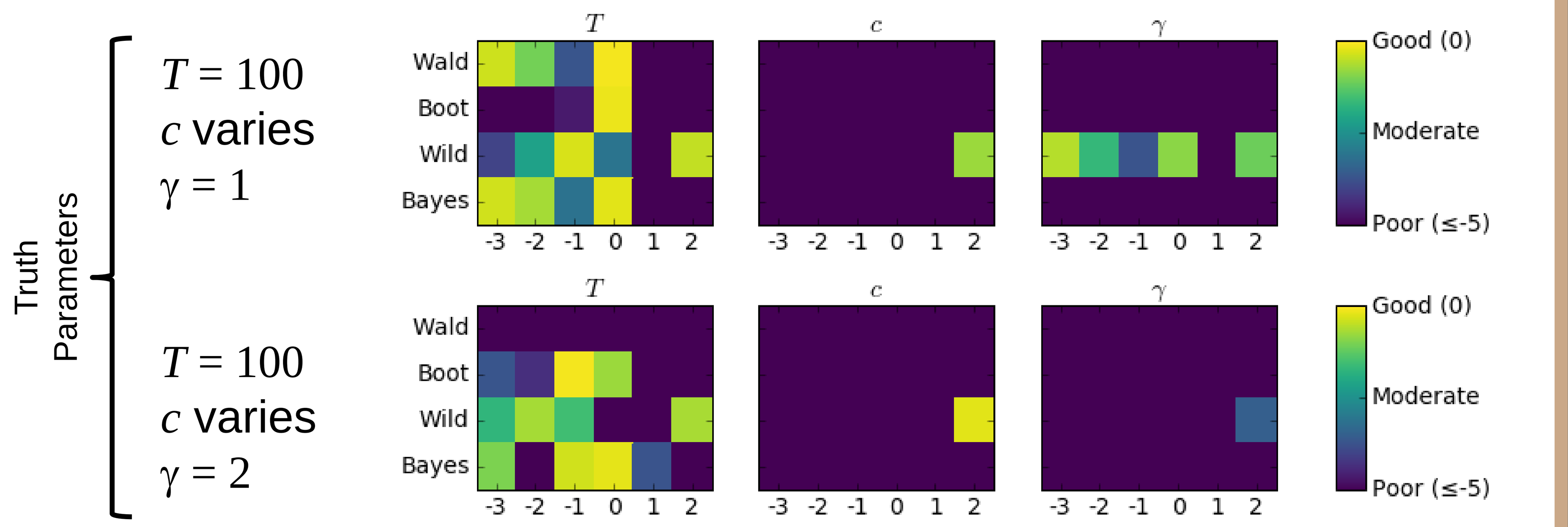
Results: Interval Estimation Performance

Log P-values Assuming H_0: 95% Coverage

(I) Homoscedastic Errors, SE = 3% T



(II) Errors Largest at Transition, SE(x) = 0.1, SE(y) = 2% T



Conclusions

- **Bootstrapping** tends to give **better coverage** than the other interval estimation methods when errors are homoscedastic.
- Data sets generated by HTS assays exhibit a **variety of error structures** (homoscedastic, error largest at transition, dual trajectory, etc.).
- Asymptotic theory confidence intervals generally do **not** give correct coverage.
- **None** of the methods give good coverage for non-homoscedastic error models, or when true value of c is >= largest measured concentration.