Field Guide to Bayesian Data Analysis Tools: JAGS, PyMC, Stan

Chris Grubb

MORS 84th Symposium

June 2016

Who I am…

- Data Scientist at Systems Planning and Analysis, Inc.
- M.S. Operations Research (George Mason University)
- You can find me on:
  - LinkedIn: https://www.linkedin.com/in/christophergrubb
  - Twitter: @channelgrubb

Approaches covered (and not covered) today

- JAGS
- PyMC
- Stan
- Bonus topics

- Languages not covered
  - Church – MIT project, see webChurch for active development
  - Figaro – a Scala-based project
  - Others – Probabilistic Programming Wiki for more information and resources (including references, summer schools, et al.)

Why am I doing this presentation?

- Promote awareness of and access to Bayesian data analysis methods/tools
- Give example uses of these methods/tools
- Provide resources for additional learning

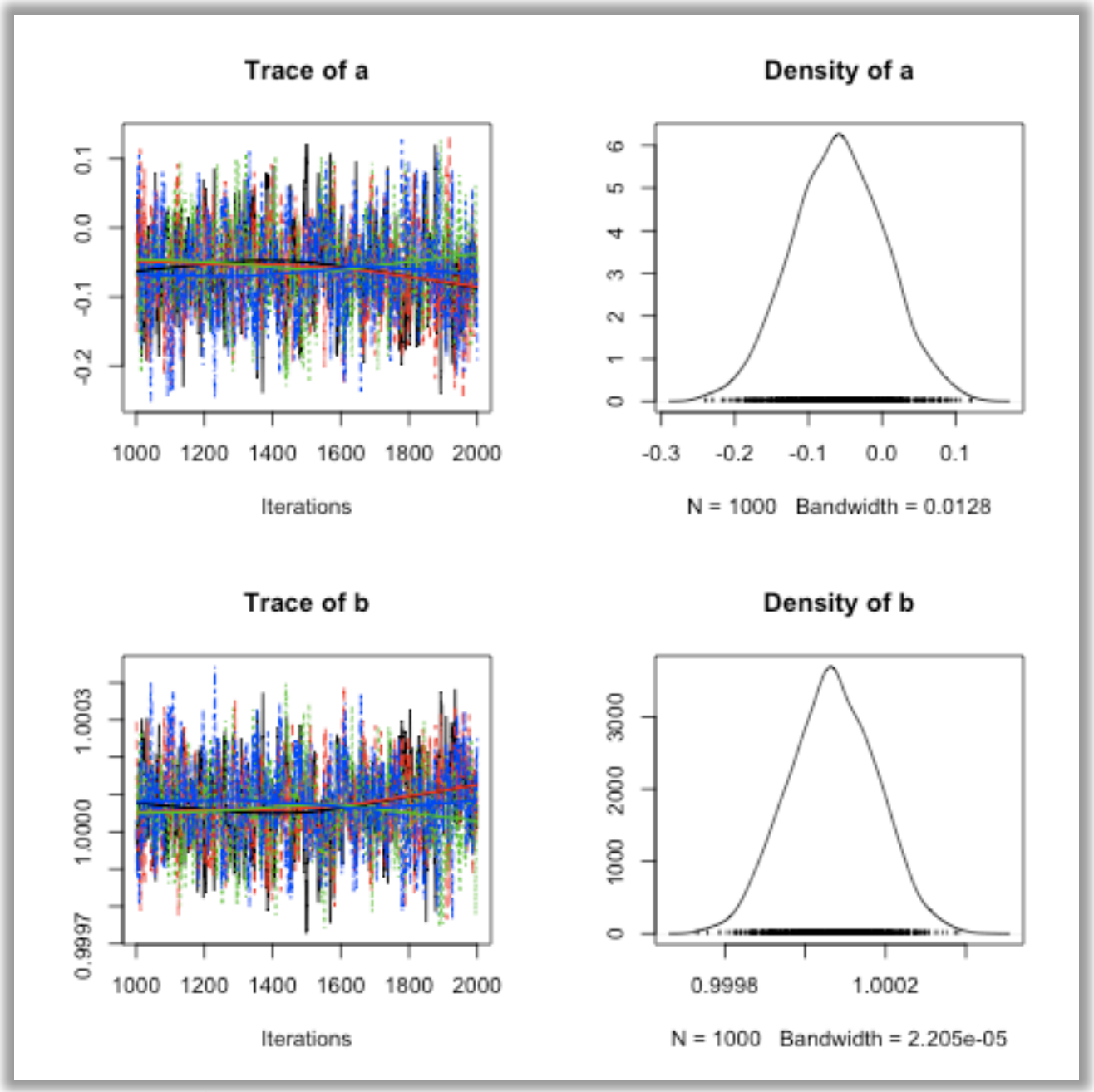
Landscape of Language Interfaces

	JAGS	PyMC	Stan
R	✓		✓
Python	✓	✓	✓
Java			✓
MATLAB	✓*		✓
Julia			✓
Stata			✓

JAGS

Just Another Gibbs Sampler

- Modeling language; dialect of BUGS
- Primarily used with R (has MATLAB & Python interfaces)
- Documentation and examples available with installer, available through SourceForge
- R interface requires rjags and coda package used for graphical results analysis



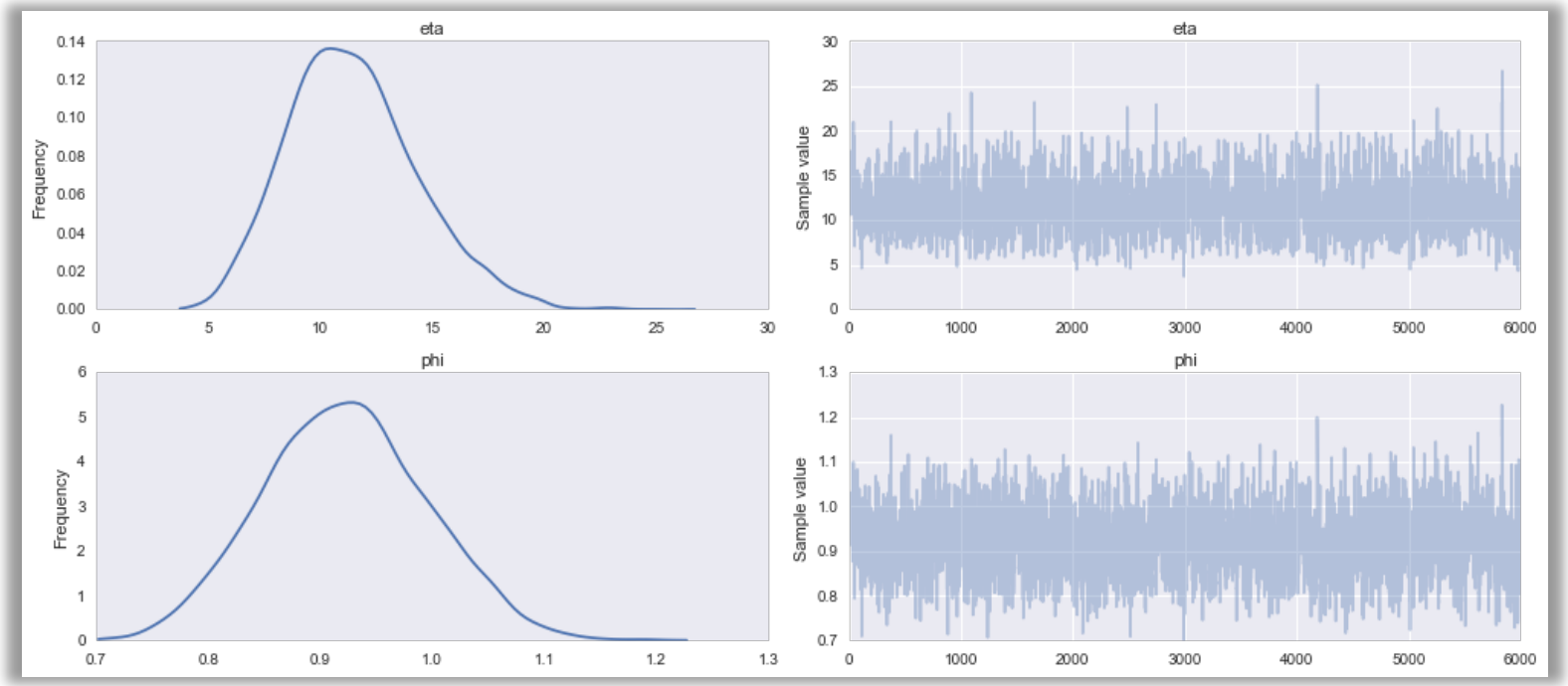
Graphics examples from:
http://www.johnmyleswhite.com/notebook/2010/08/29/mcmc-diagnostics-in-r-with-the-coda-package/

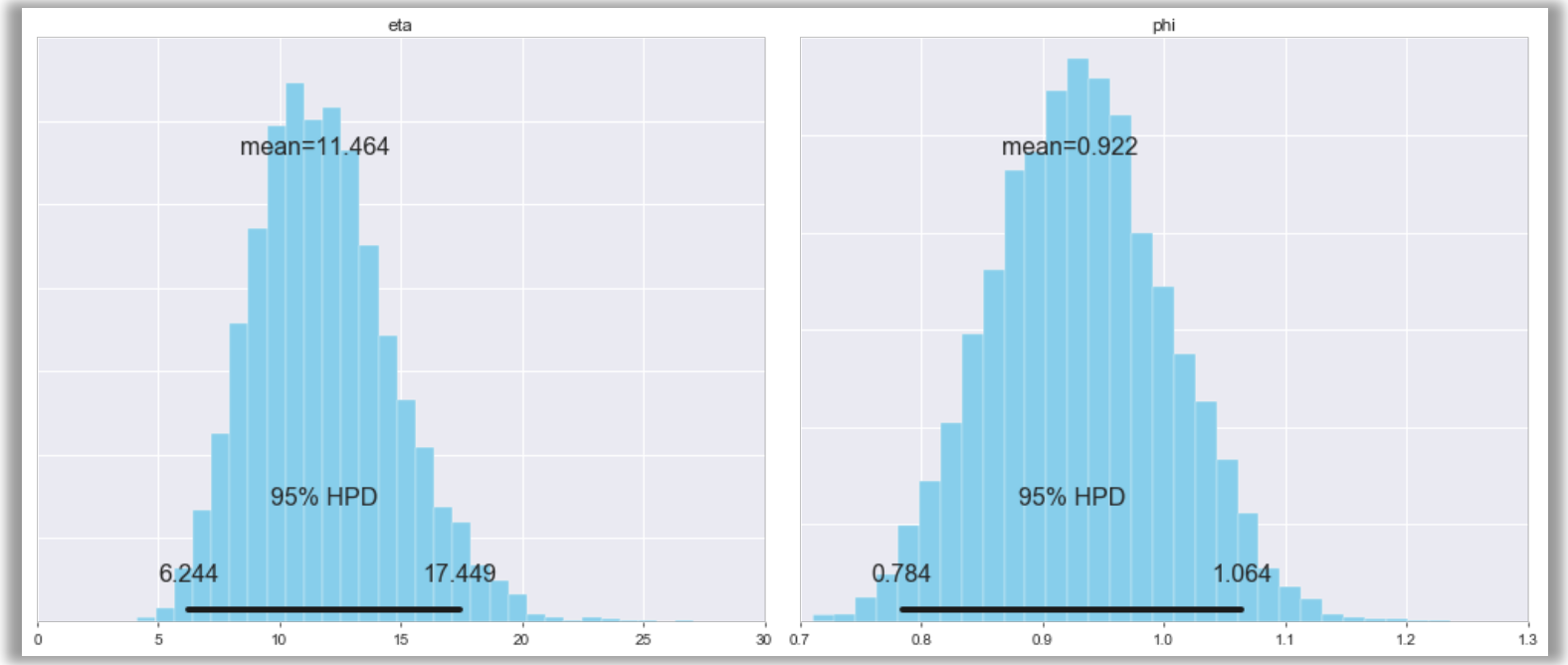
Stan

- Distinct modeling language with multiple interfaces
- **Excellent** documentation, including a detailed language manual, a large number of examples, and well-developed case-studies
- Robust R interface + Shinystan (online demo)
- Multiple inference methods (primarily NUTS/HMC)
- Interfaces through many other languages
- Strong Bayesian data analysis community support

PyMC 3

- Bayesian modeling package written entirely in Python
- PyMC 2 still available; PyMC 3 under active development; API documentation here
- Rich (and growing) set of examples
- Multiple sampling methods: MH, NUTS/HMC, et al.
- Improved model building syntax
- Out-of-box support for model diagnostics and graphics

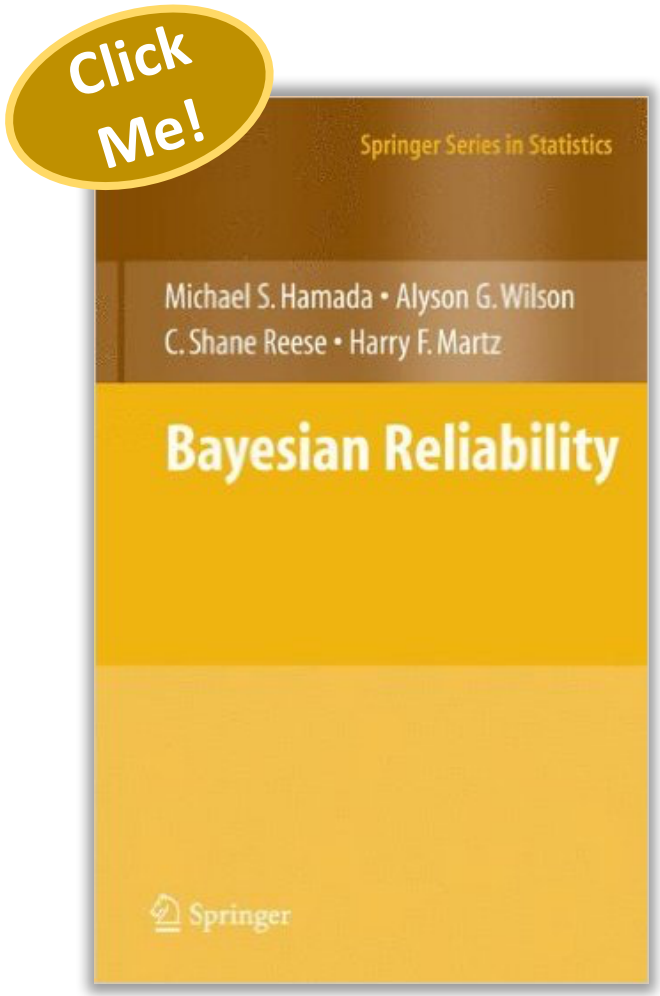




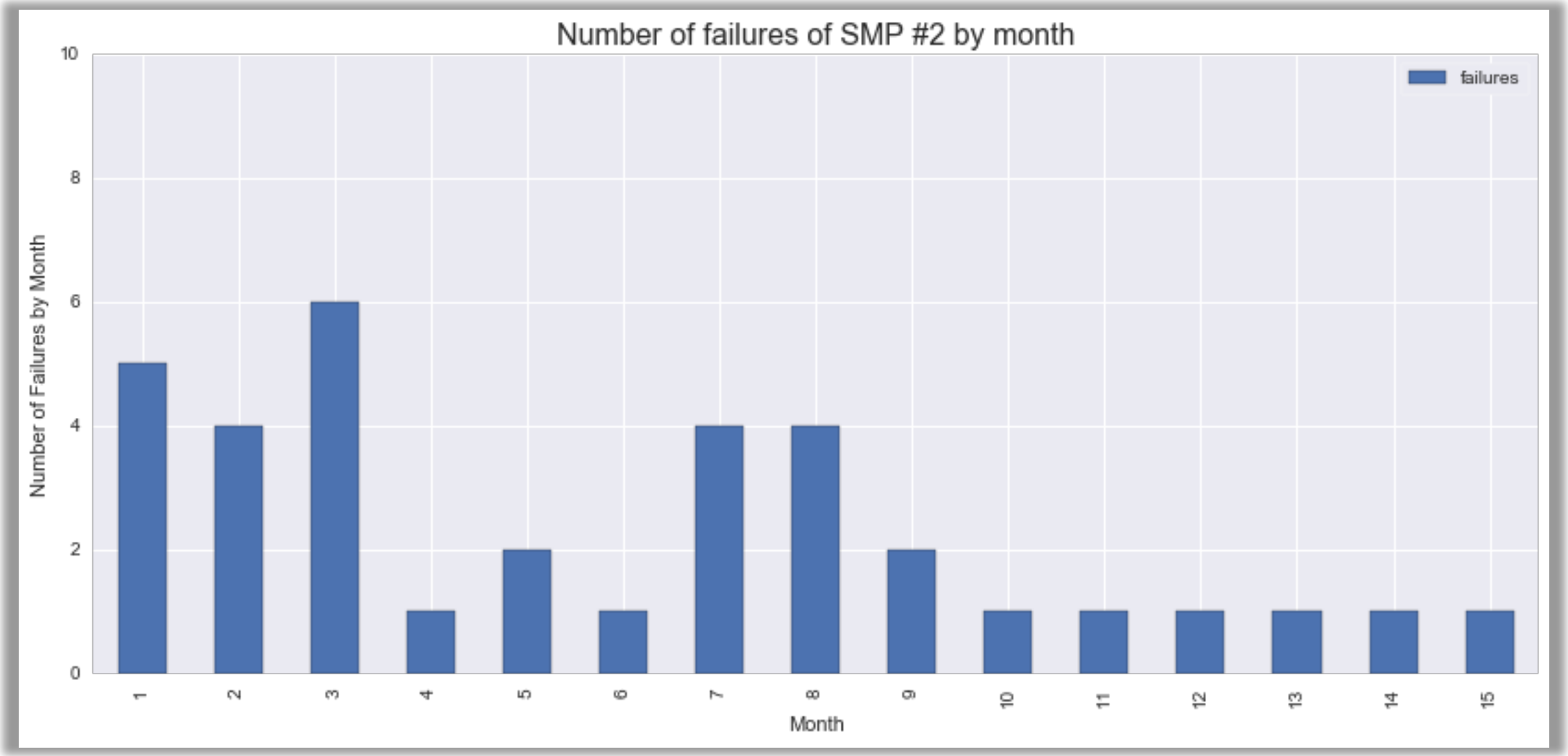
Bayesian Reliability

Hamada, Wilson, Reese, and Martz (2008)

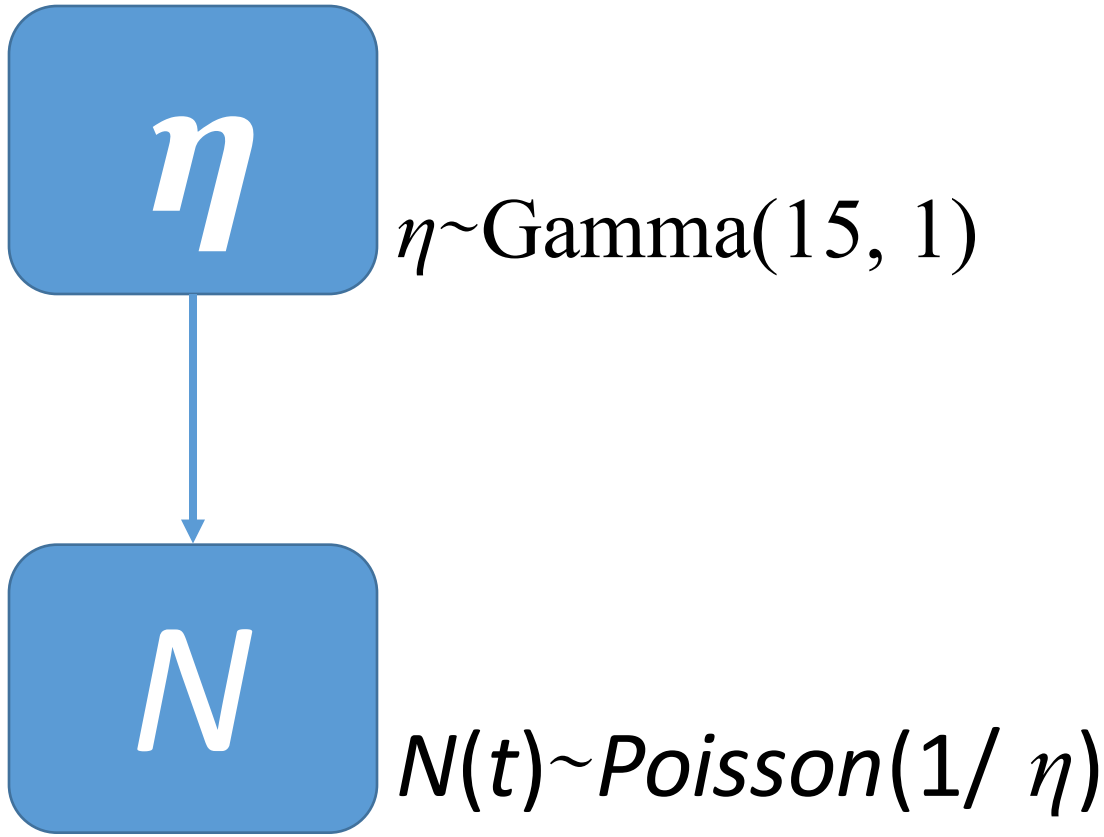
- The following examples are based on *Bayesian Reliability*
- Modeling the failure counts of a single processor
- Problem basics: failure data were collected from 47 separate shared memory processors (SMP) over 15 months of their operation; we'll focus on SMP #2
- We'll show excerpts of implementations in PyMC3



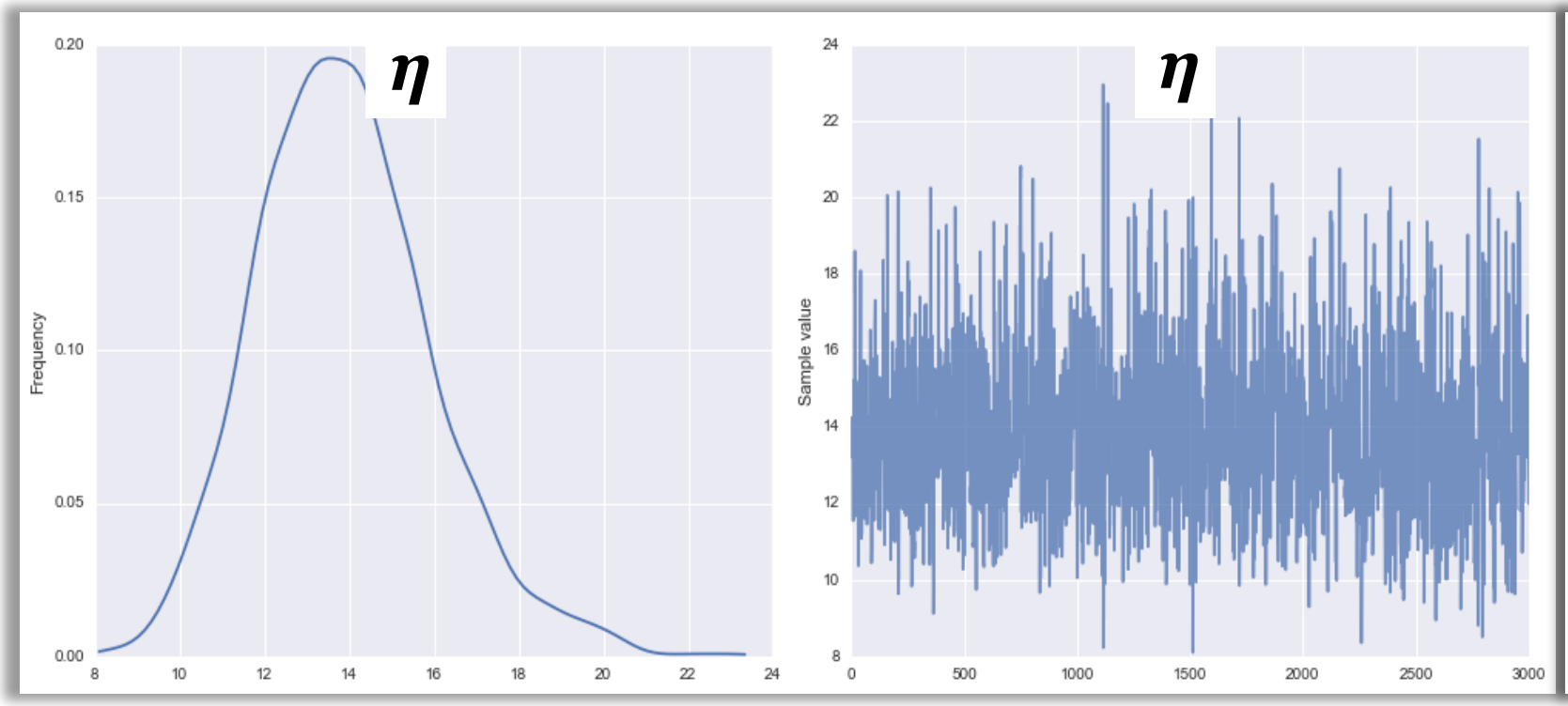
Failures by month for SMP #2

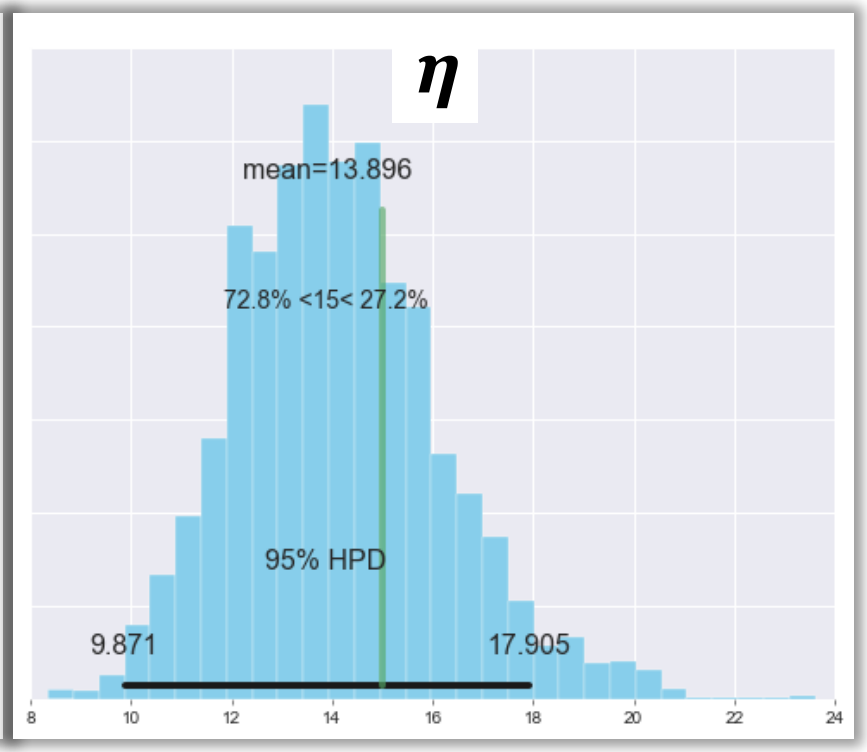


Model 1: Homogeneous Poisson Process (HPP)

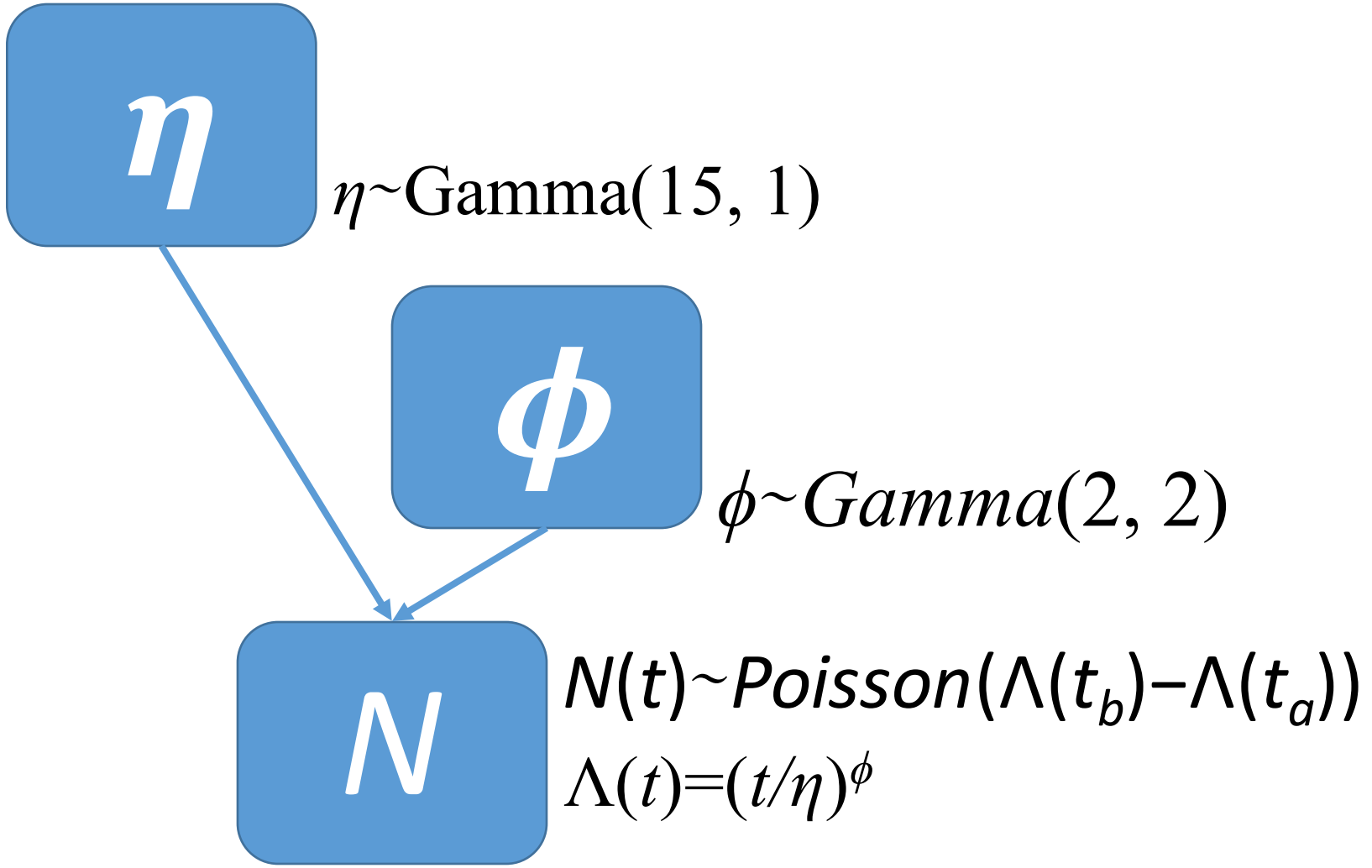


Model 1: Homogeneous Poisson Process (HPP)





Model 2: Non-Homogeneous Poisson Process using a Power Law Process model (graphical)

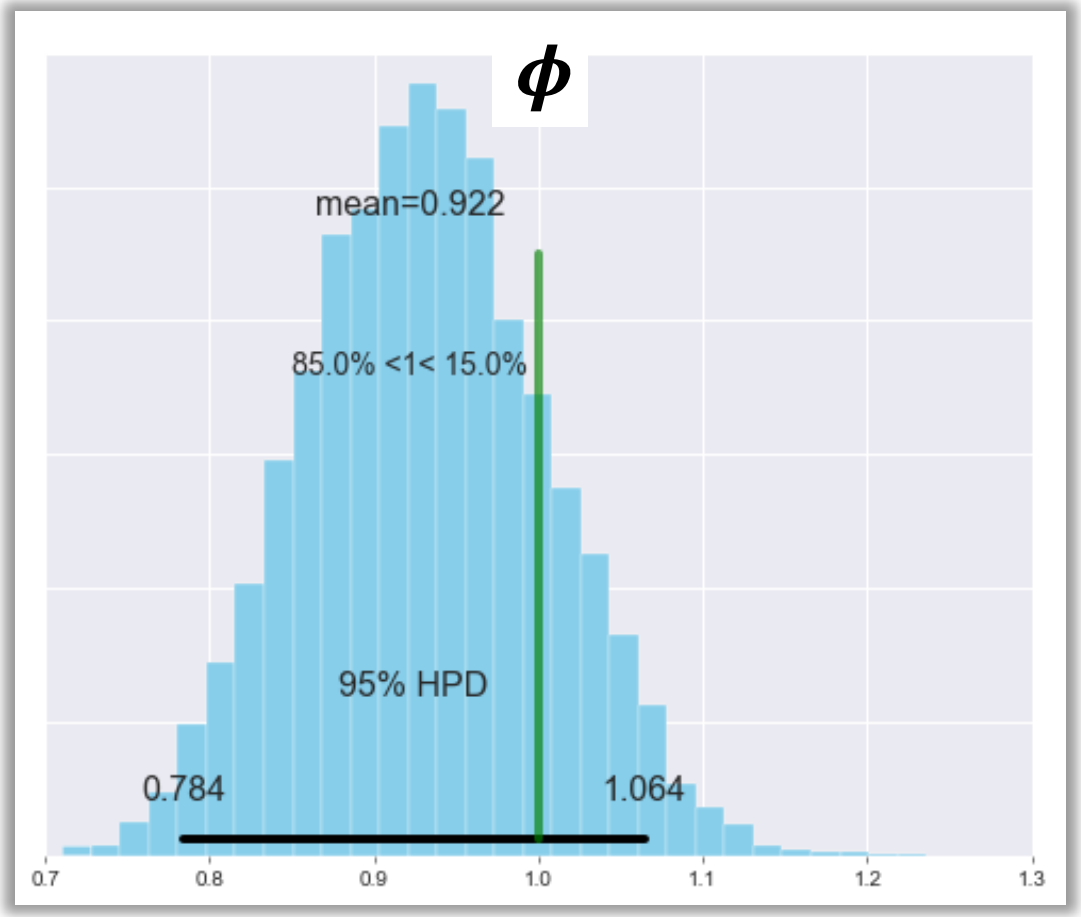


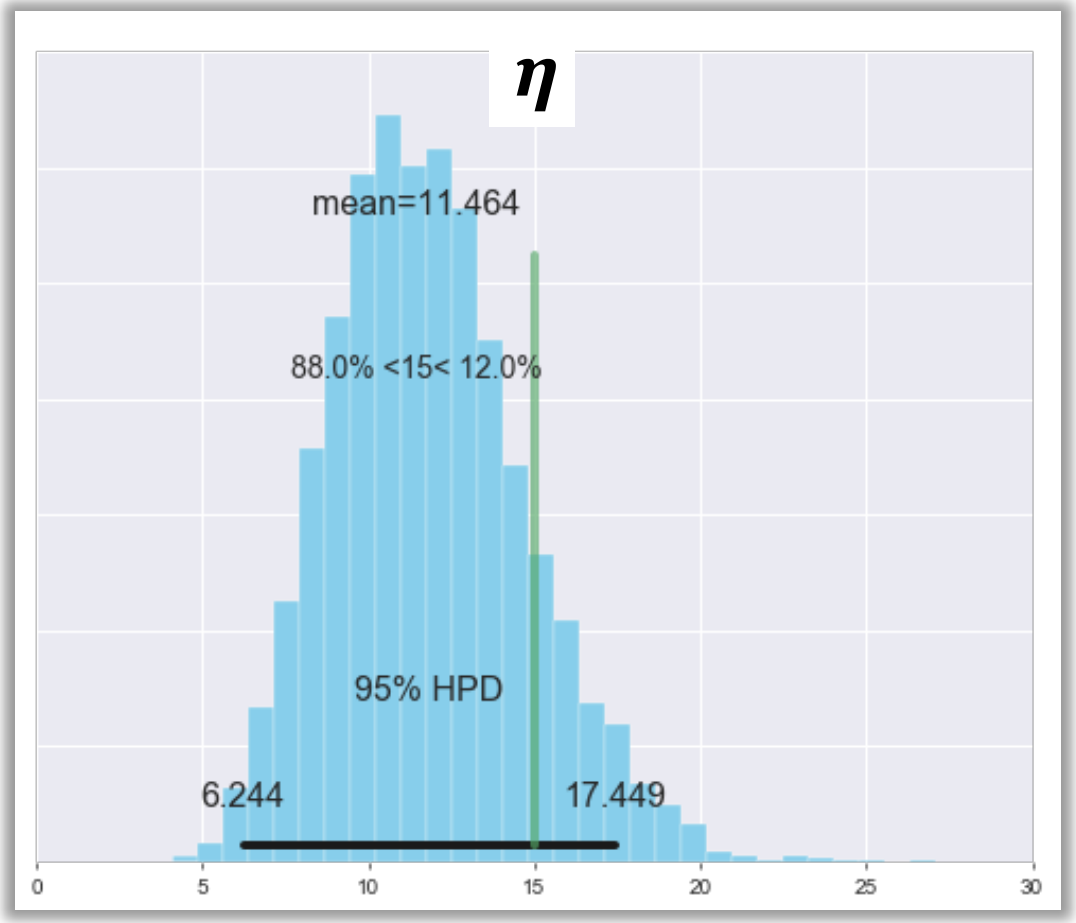
Model 2: Non-Homogeneous Poisson Process using a Power Law Process Model (PyMC3)

```
with pm.Model() as SMP2_model:

    # priors
    eta = pm.Gamma('eta', alpha=15, beta=1)
    phi = pm.Gamma('phi', alpha=2, beta=2)

    # Model and model error
    intensity_1 = ((SMP2_failures.b.values**phi) - (SMP2_failures.a.values**phi))/(eta**phi)

    # Data likelihood
    y = pm.Poisson('y',mu=intensity_1, observed=SMP2_failures.failures)

    SMP2_trace = pm.sample(6000, random_seed=20160622)
```

```
Applied log-transform to eta and added transformed eta_log to model.
Applied log-transform to phi and added transformed phi_log to model.
Assigned NUTS to eta_log
Assigned NUTS to phi_log
 [-----------------100%-----------------] 6000 of 6000 complete in 6.7 sec
```

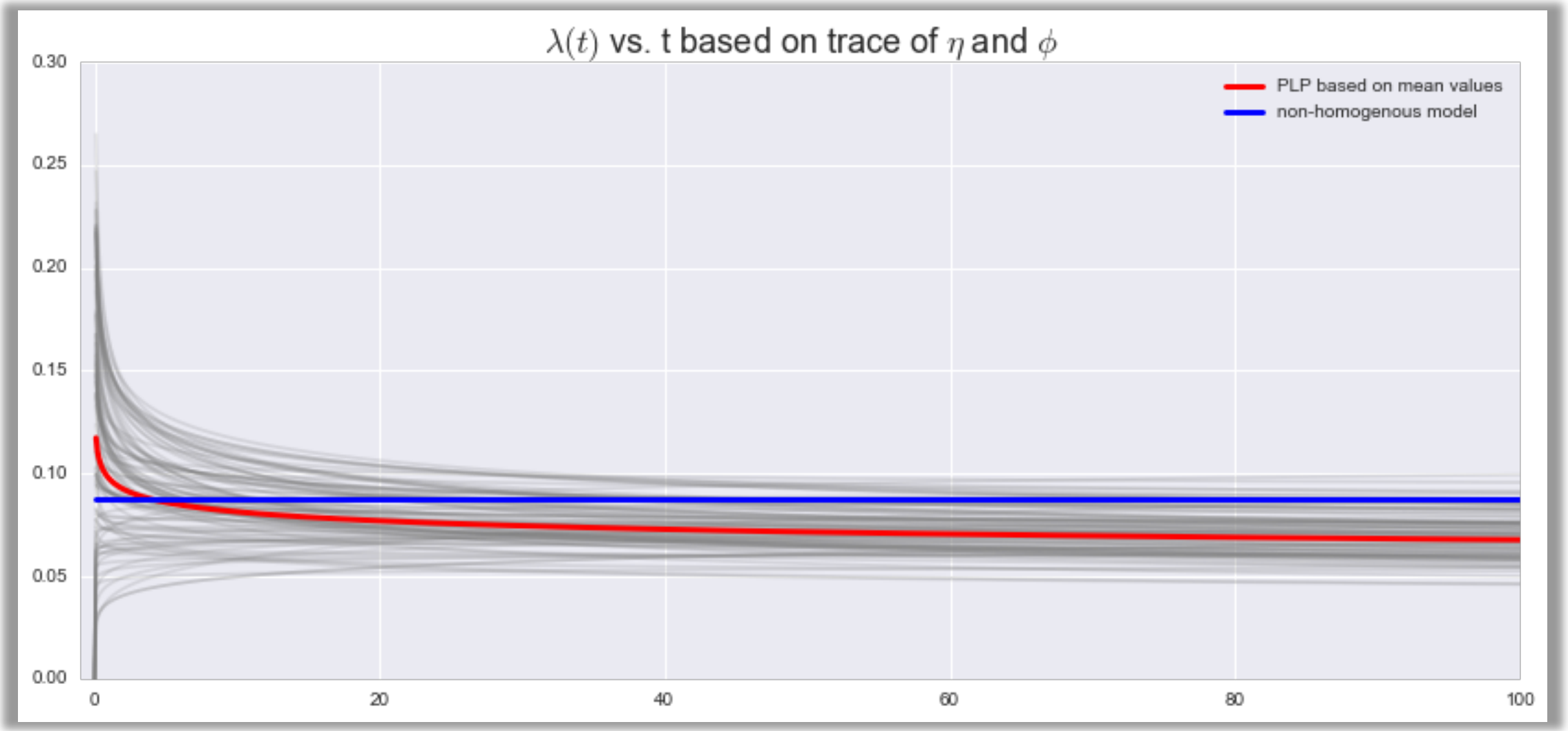
Results – Marginal Posteriors of ϕ and η





Plotting $\lambda(t)$

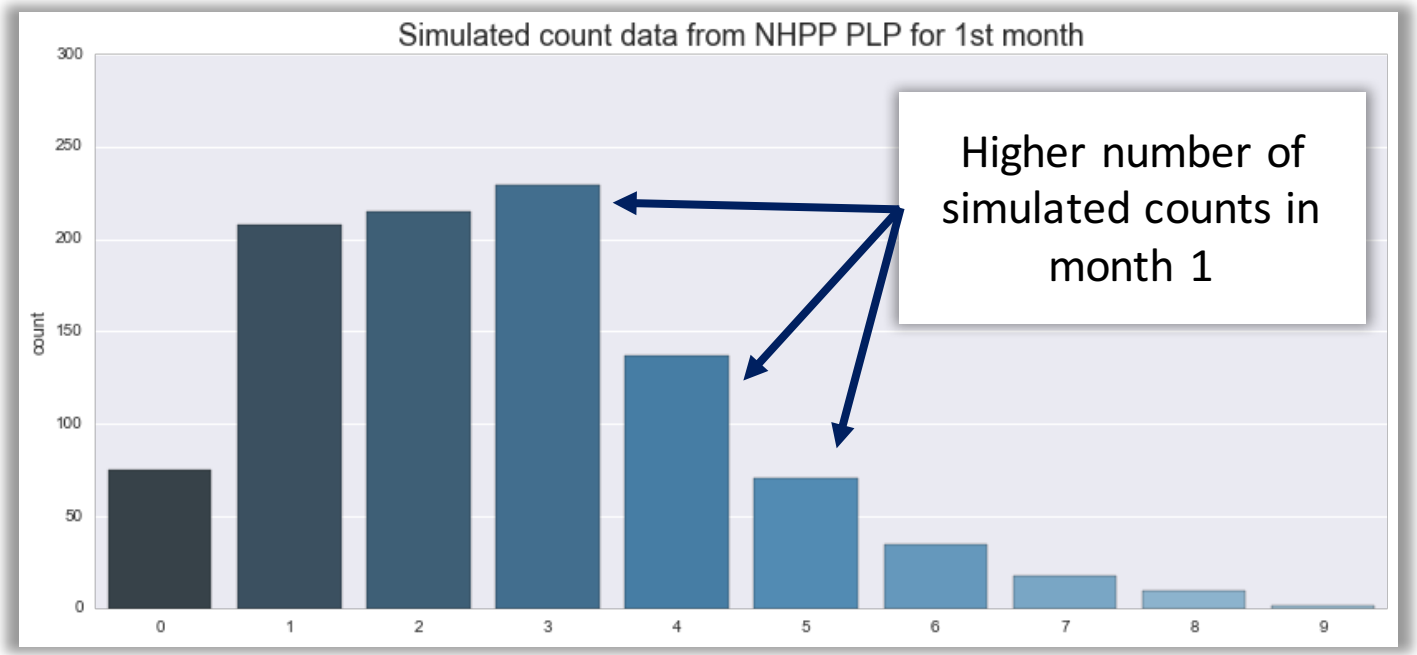
based on last 100 samples of MCMC trace

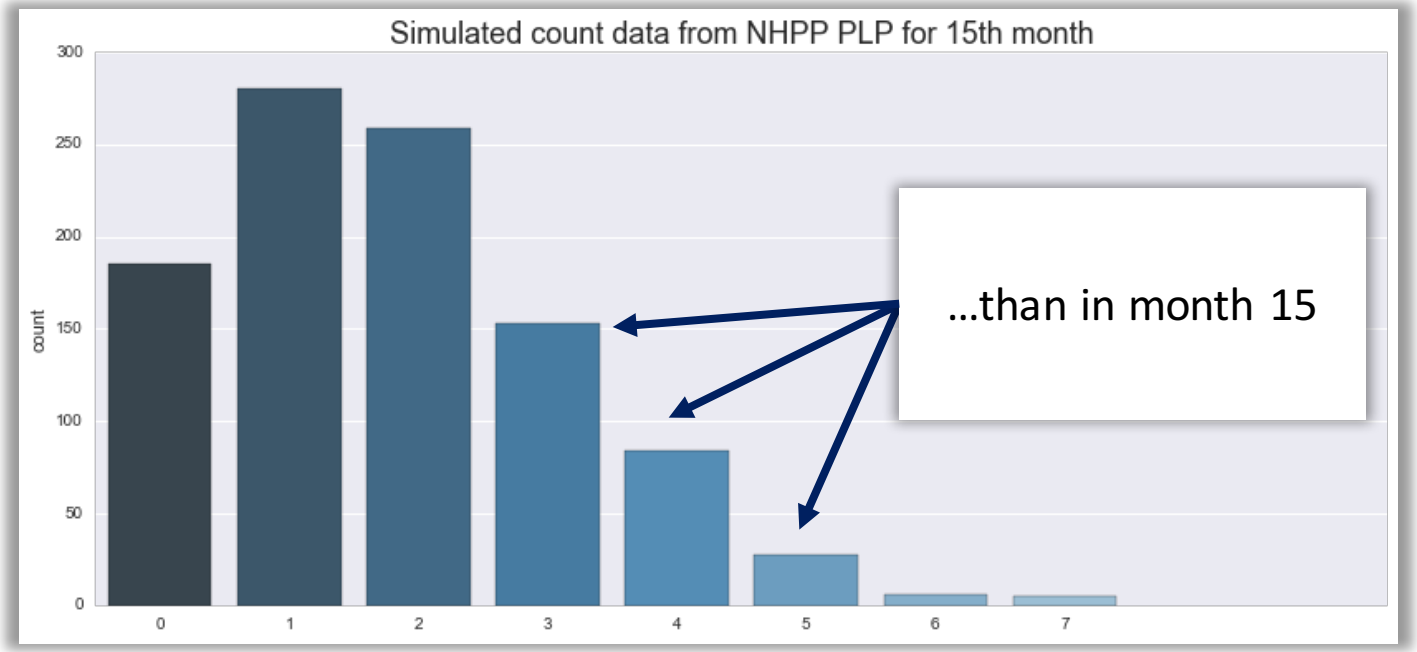


Simulated failure count data

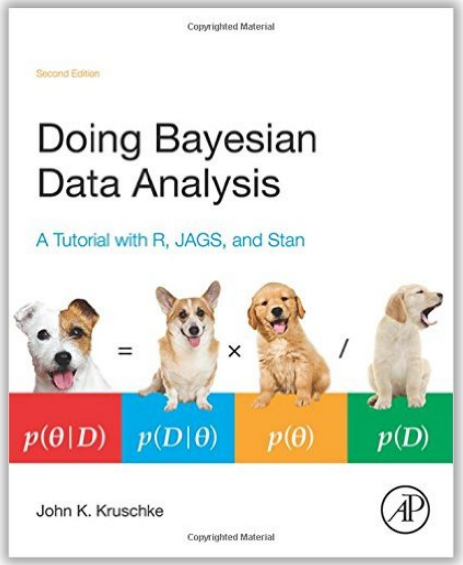
- Simulated data show that in the first month there are more failures than in the 15 month

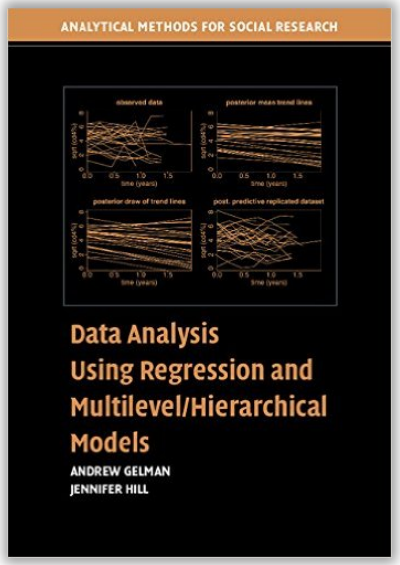
- Given that E[ϕ | Data] ~ 0.92, we see that the failure rate improvements are modest over time

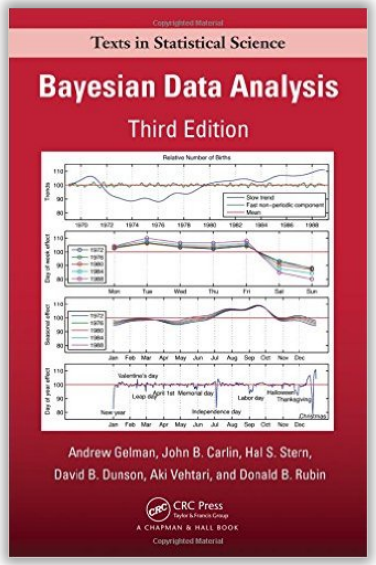


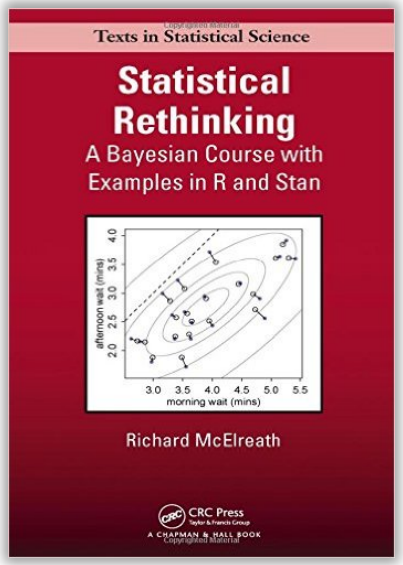


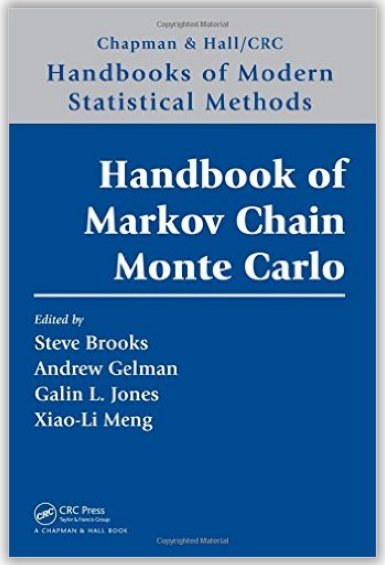
Books for your bookshelf

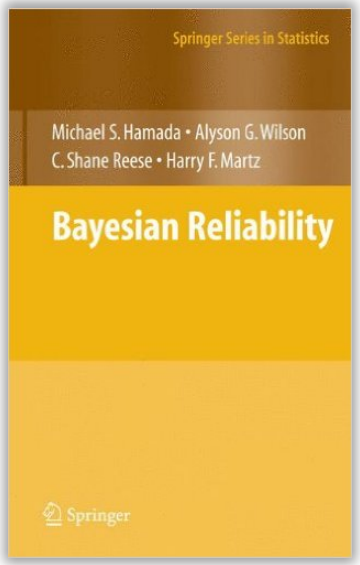












Sample of Web Resources

Links in addition to those directly to PyMC3, Stan, and JAGS pages

- Dr. Kathryn Laskey (GMU) posts the slides and assignments for her entire course on Bayesian Inference and Decision Theory (<-- great course!)
- Dr. Andrew Gelman's blog (see e.g. post on Stan and PyMC)
- The Stan Google Group is a treasure trove of discussions on not only Stan, but Bayesian modeling in general. Also, core developers and other experts are very engaging and willing to answer questions!
- Dr. Thomas Twiecki (Quantopian, key contributor to PyMC3) has a great blog with examples of PyMC3 implementations
- Duke offers a course in Computational Statistics based in Python, which has nice resources to Python stats, PyMC 2 and 3, and adjacent projects
- Bayesian Methods for Hackers: a popular introduction to Bayesian methods
- Martyn Plummer's JAGS News blog

Thank you!