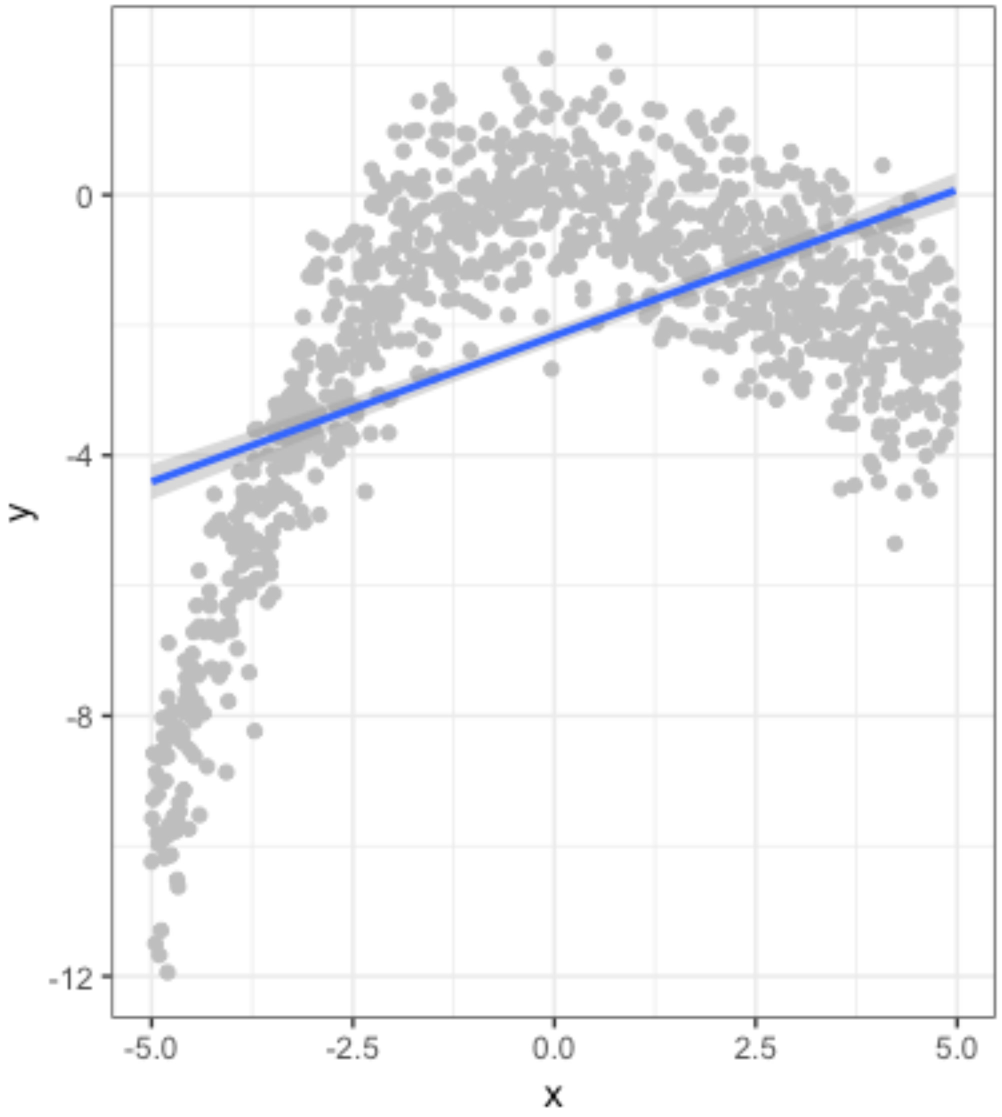
SOMETIMES ALL WE HAVE LEFT ARE PICTURES AND FEAR

Daniel Simpson
Department of Statistical Sciences
University of Toronto

Joint with: Lauren Kennedy, Aki Vehtari, Andrew Gelman, Michael Betancourt, Sean Talts, Bob Carpenter, Yuling Yao, Jonah Gabry

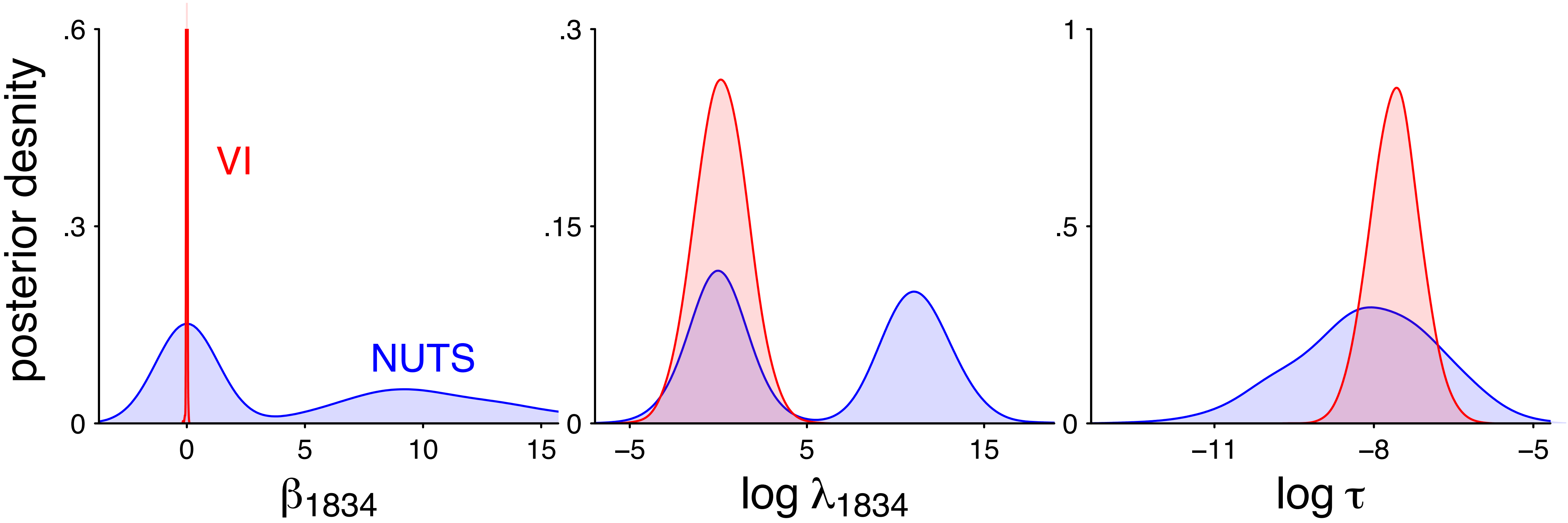
OF GODS AND MONSTERS

WHAT SCARES YOU THE MOST?



MODEL MISSPECIFICATION

WHAT SCARES YOU THE MOST?



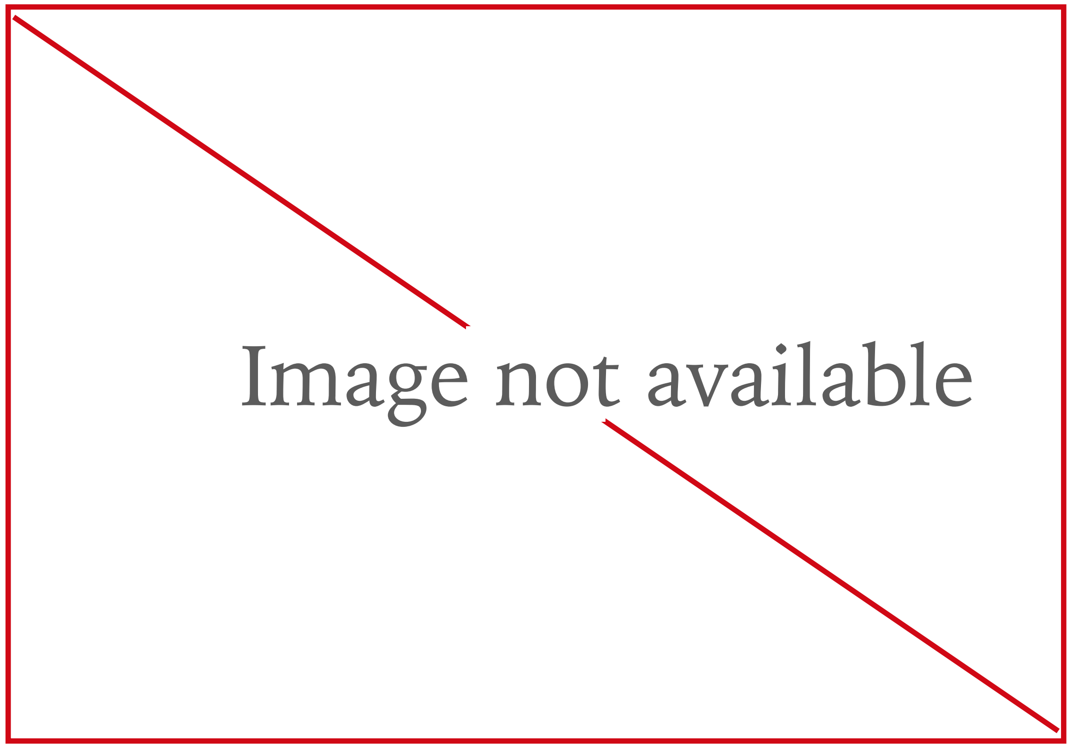
COMPUTATIONAL DISASTER

WHAT SCARES YOU THE MOST?

n ⛷ 🏊 🏇 ⛔ ↶ ∞

GETTING THE ASYMPTOTIC REGIME WRONG

WHAT SCARES YOU THE MOST?



THAT IT'S IMPOSSIBLE, ON THE AVAILABLE EVIDENCE, TO TELL THE DIFFERENCE BETWEEN A VAMPIRE AND A CANNIBAL WHO JUST HAS SPECIFIC PREFERENCES

YOU'RE ALL WRONG!

EVERYTHING IS TERRIBLE.

WE DON'T KNOW HOW ANYTHING WORKS.

ASYMPTOPIA IS JUST A CONSPIRACY OF CARTOGRAPHERS.

LET US SIT AND TALK OF HELEN OF TROY

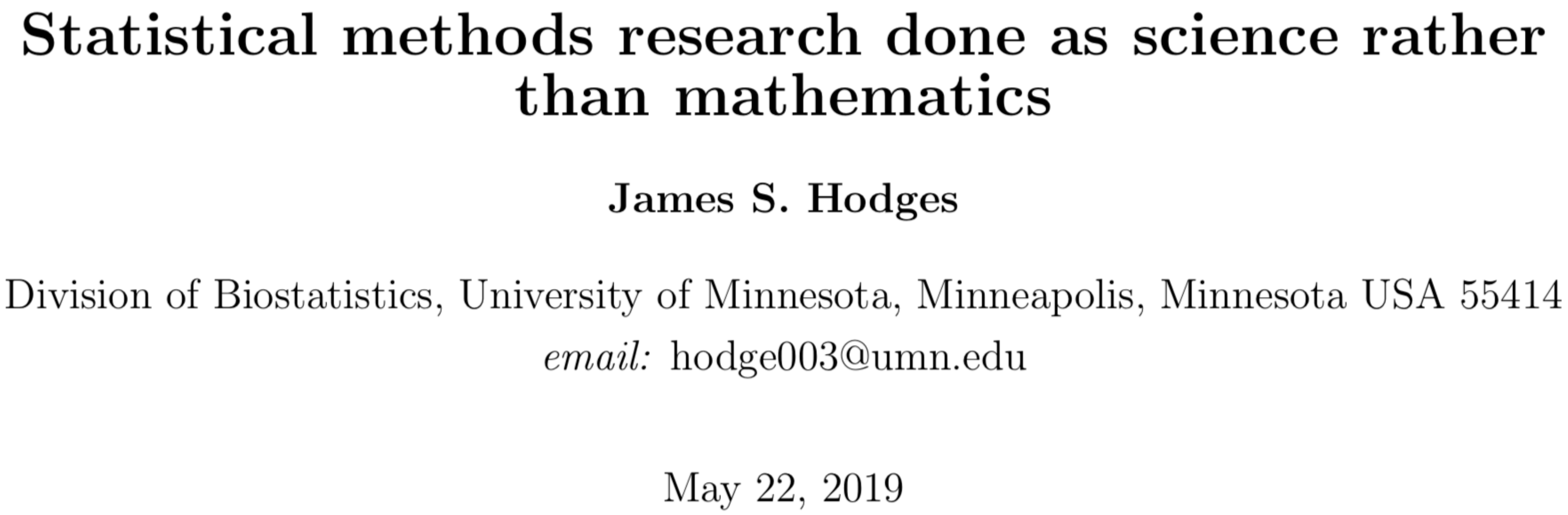
HELEN WAS PARIS'
PRIZE FOR PROCLAIMING
APHRODITE THE FAIREST

HOW CAN I TELL IF MY MODEL FITS

- Firstly, **how *DARE* you???**
- But really, the best way that we can do this is by looking at pictures
- This is STAT100 regression type of stuff:
  - Check normality (Q-Q plot or histogram)
  - Check for serial autocorrelation (time series plot)
  - Check for homoskedasticity (scatterpolot)
  - Check for high leverage points and outliers (magic plot)
- But what do you do when you are no longer fitting straight lines though things?

WHAT IF ALL WE REALLY KNOW HOW TO DO IS PUT A STRAIGHT LINE THROUGH THINGS?

- In a recent article, Jim Hodges makes the argument that even moving slightly beyond this framework leads to underdeveloped model checking and model understanding.



Statistical methods research done as science rather than mathematics

James S. Hodges

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota USA 55414
email: hodge003@umn.edu

May 22, 2019

OUR QUESTIONABLE SAVIOUR: PREDICTION

- If we have some **observable** that we are interested in, we can look at how well this is predicted by the data.
- Cross-validation is our only all-purpose **(????)** tool for checking prediction quality.
- The basic idea is that if we split our data into training and test sets, it's better to do it multiple times.
- The danger is that this **strongly** leans on the assumption that current data is exchangeable with future data.

TYPICAL USE OF CROSS VALIDATION

- Take your data y and pull out k observations (here either k is 1 or around 10% of your observations)
- Fit your model on the remaining observations.
- Compute the **log-score** of your remaining observation

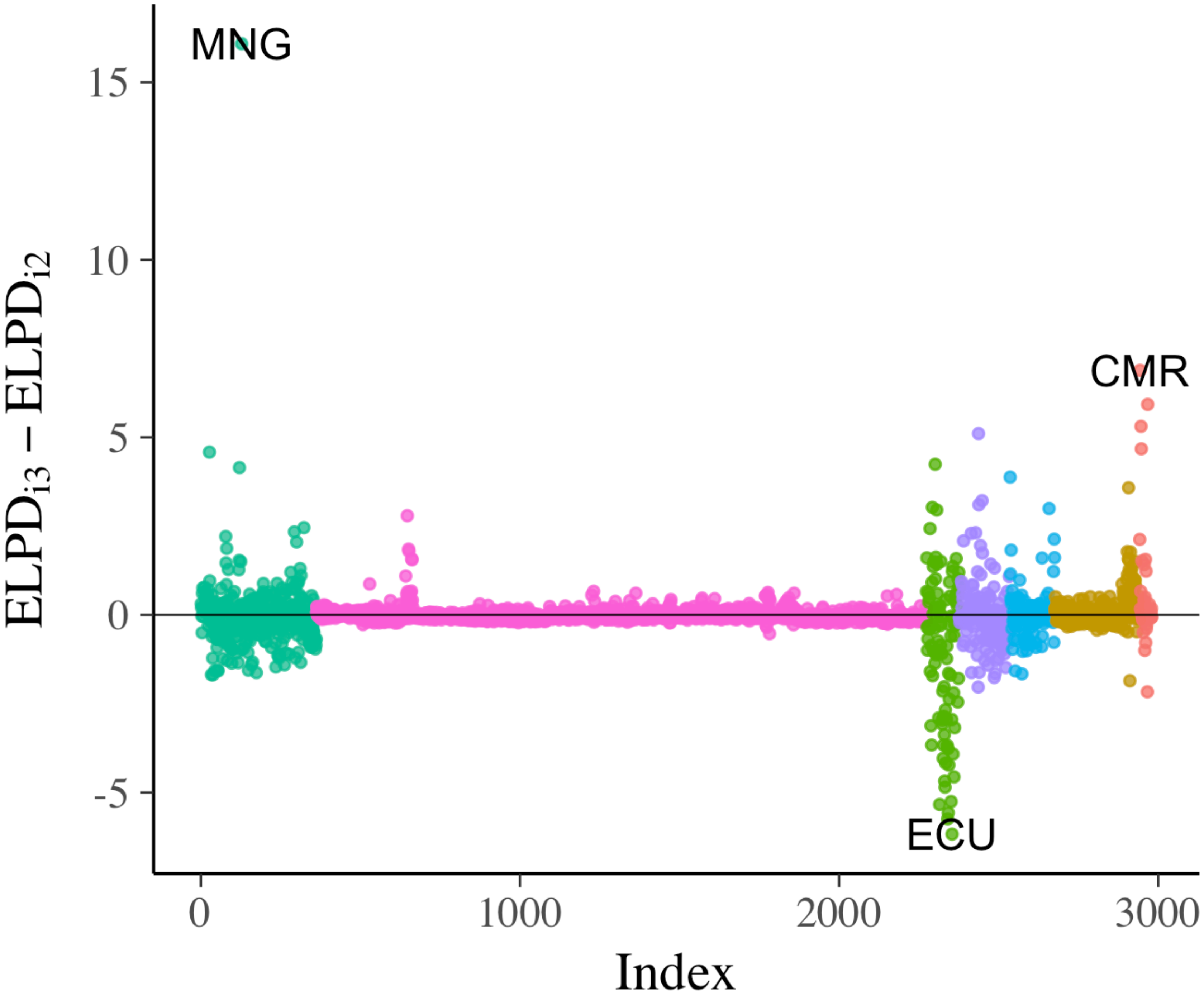
$$\text{elpd}_j = \frac{1}{k}\sum_{i=1}^{k} \log p(y_i \mid y_{-k})$$

- Repeat this as many times as possible and report the average elpd.

BIGGER IS ALWAYS BETTER

- This is then typically used to compare between two models
- The assumption is that bigger is better.
- Why? Because elpd converges (under independence assumptions) to a constant minus the Kullback-Leibler divergence between the prediction and the data generating distribution.
- But there’s more information than just the sum…

MORE INFORMATION THAN JUST COMPUTING A SUM

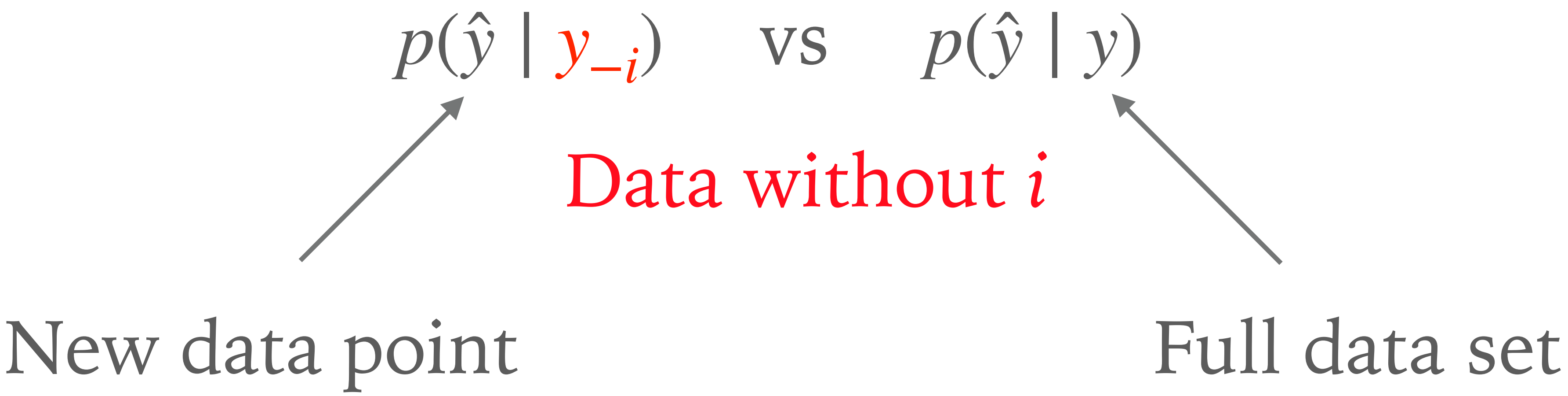


THE SECRET UTILITY OF LEAVE ONE OUT DISTRIBUTIONS

- But what if we didn’t care about comparing with another distribution?
- There’s still value here.
- Why? Because what does it mean if the leave-one-out predictive distribution is different from the whole-data predictive distribution?
- It is **important** to know about these influential points!
- They are very similar to high-leverage points in linear regression.

JUST A MOMENT

- For observation i, we want to compare



$$p(\hat{y} \mid y_{-i}) \quad \text{vs} \quad p(\hat{y} \mid y)$$

Data without i

New data point

Full data set

- How do we tell if these things are similar?
- **Idea:** The full data predictive distribution should be a good importance sampling proposal for the loo distribution, ie

$$\mathrm{Var}_{\theta \sim p(\theta \mid y_{-i})} \frac{p(\theta \mid y_{-i})}{p(\theta \mid y)} < \infty$$

OK THIS IS HARD TO CHECK IN GENERAL

- In general, we only have access to a sample of these importance weights.
- This makes things hard.
- But some classical statistics comes to the rescue:

The extreme tail of a distribution converges
to a Generalized Pareto Distribution

- So while we might not be able to check analytically if the importance weights has a finite variance, we can estimate the distribution of the extreme weights, **which gives us the same information.**

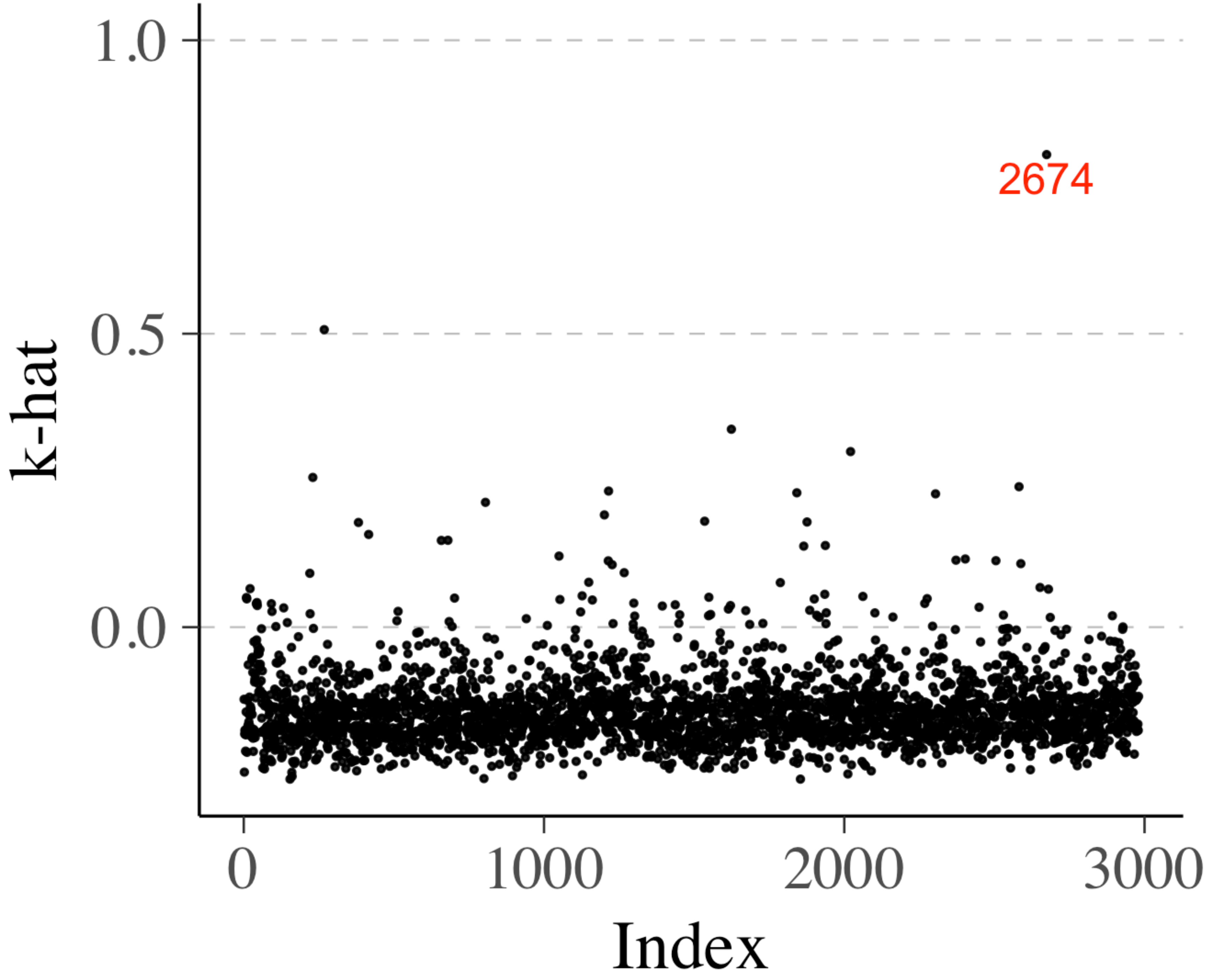
THE GENERALIZED PARETO DISTRIBUTION

- The generalized Pareto distribution has the

$$p(z) = \frac{1}{\sigma}\left(1 + kz\right)^{-1/k-1}$$

- The key parameter is *k*, which controls how many moments the tail distribution has.
- We can estimate *k* by k-hat, which tells us how many moments a specific sample appears to have.
- This is an extremely useful, and easy to compute, quantity. Because if k-hat is large, then the LOO predictive distribution is **very** different from the full data predictive distribution!

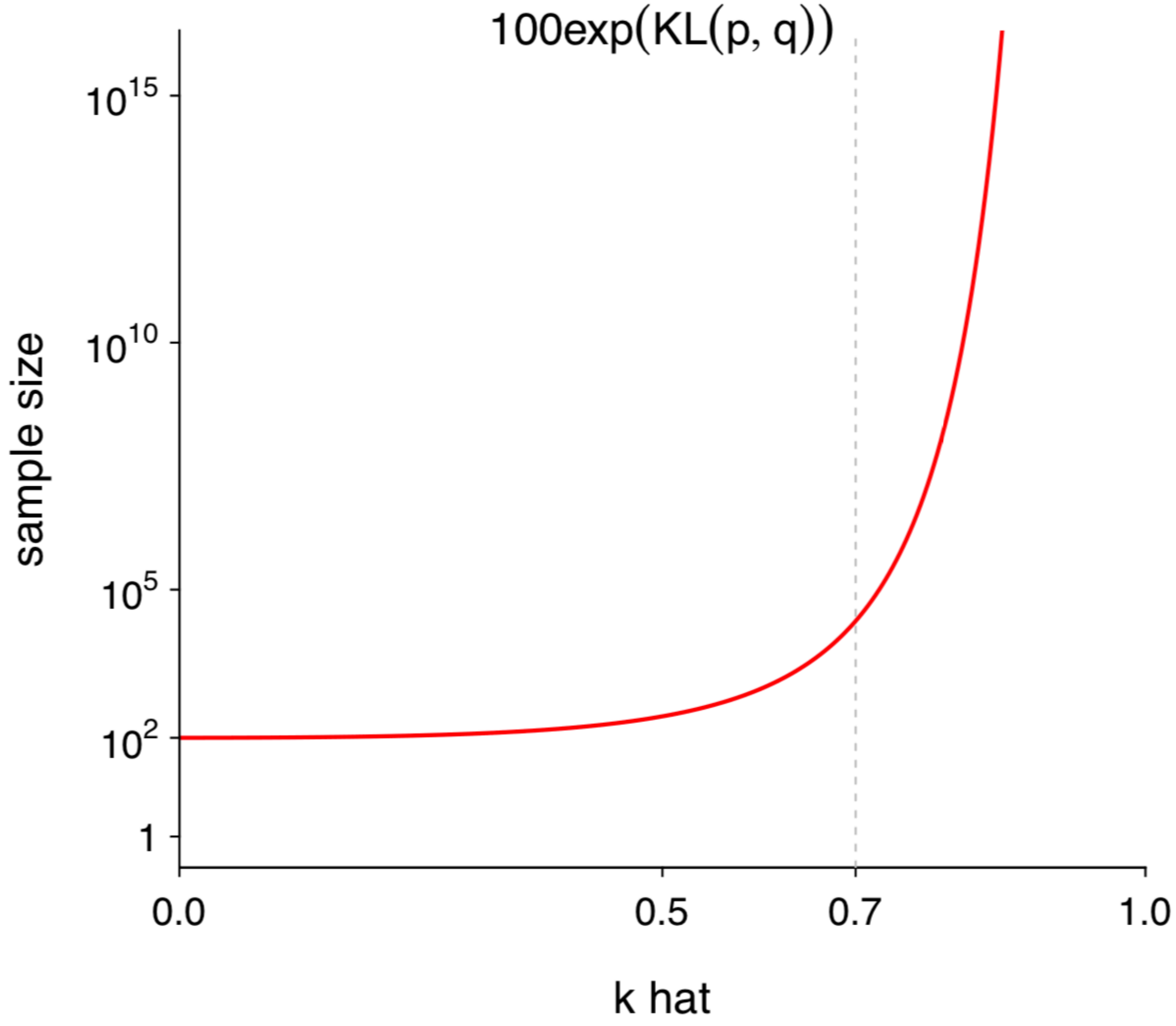
DIAGNOSTICS (K–HAT: A PREDICTIVE LEVERAGE)



Mongolia

BUT THIS IS AHISTORICAL

- The k-hat diagnostic did not come as an attempt to generalize the concept of leverage.
- It came as a way to make a version of importance sampling that verified its own assumptions.
- Essentially, k-hat is related to the number of samples needed for an accurate importance sampling estimate.



ATHENA AND HERA WERE FURIOUS AND CONSPIRED WITH HERMES

MY ALGORITHM SAMPLES FROM YOUR POSTERIOR

WHAT ABOUT OTHER FIELDS?

- ***A posteriori* error estimates:** Hard to be both tight and rigorous (see variance of importance samplers [not rigorous], coupling arguments *a la* Jacob [not obviously tight])
- **Benchmarking**: Hard to do here, You get the "NeurIPS problem"

HORNIMAN WALRUS

(CAN) I BIND YOU NANCY!(?)

BUT HOW DO YOU UNIT TEST A STOCHASTIC ALGORITHM?

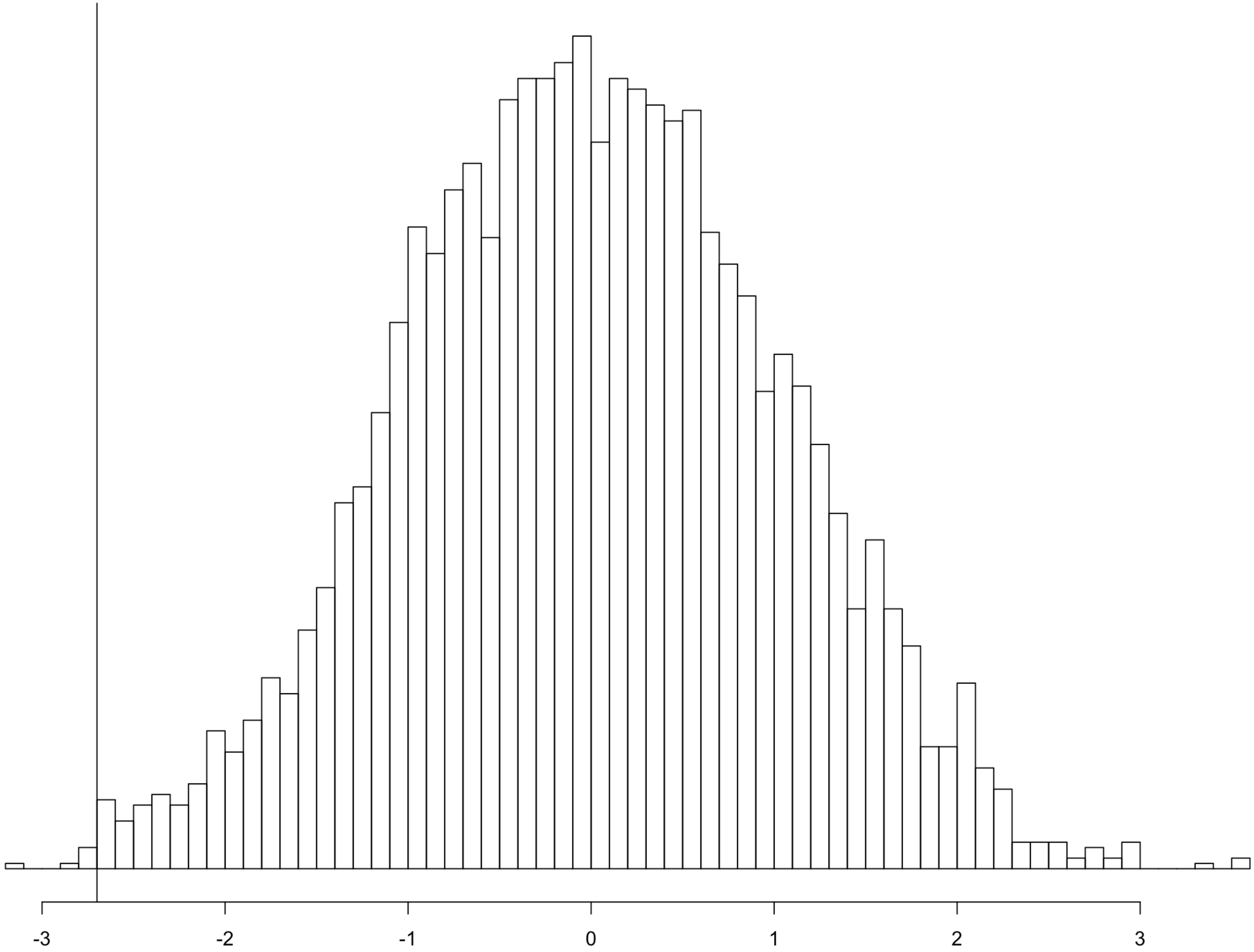
- Any reliable software is built on a foundation of crushed dreams and unit tests.
- But how do we do this for implementations of algorithms?
- Actually this is very hard in general, but there is a way in the promised land of Bayes!

A MORE GENERAL IDEA

- Idea: Run the algorithm on simulated data.

1. Pick a parameter value θ_0
2. Generate data from $p(\mathbf{y} \mid \boldsymbol{\theta}_0)$
3. Fit model to data
4. Compare the posterior to the known true value

OKAY! IS THIS RIGHT?

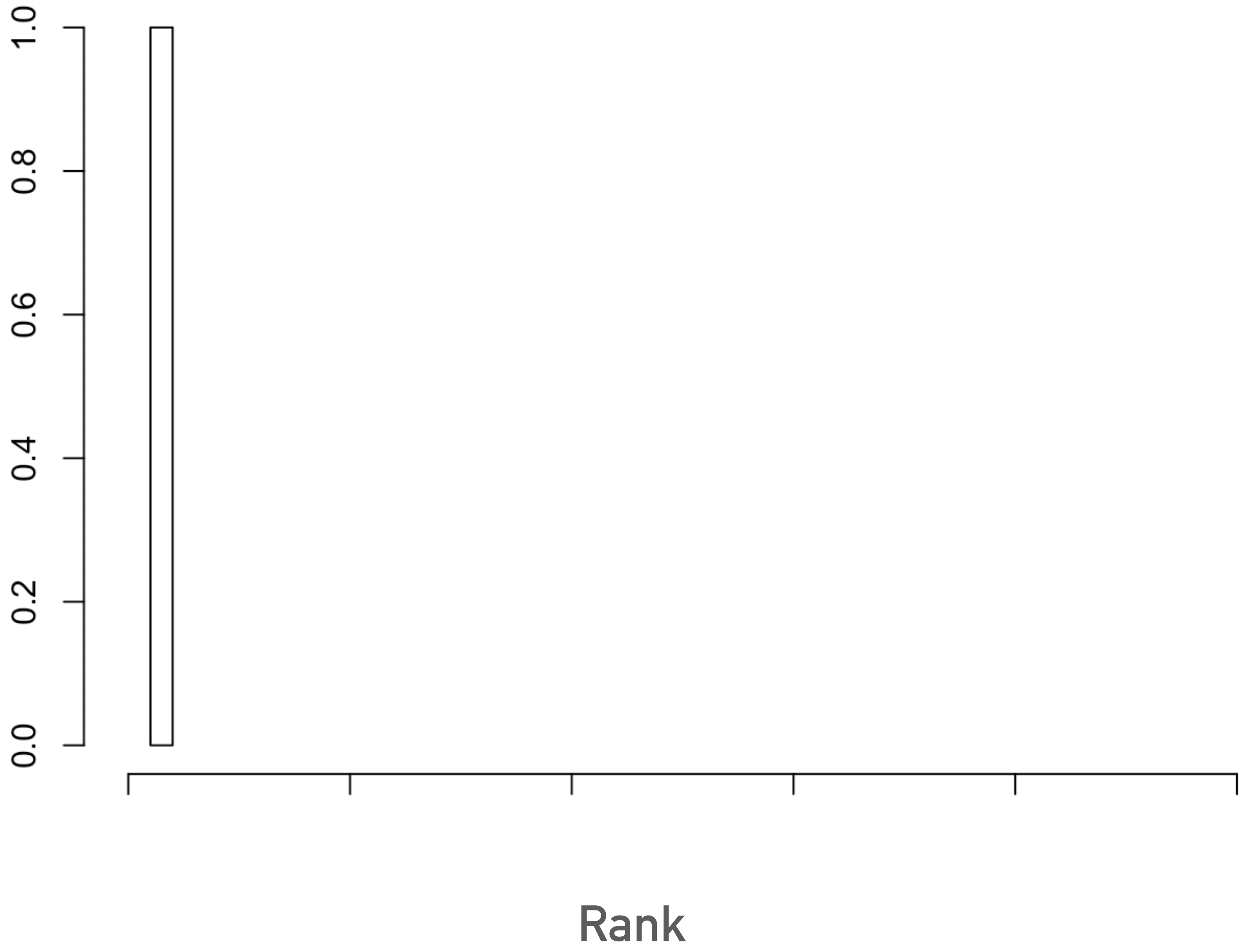


HOW DO YOU TELL IF IT FITS?

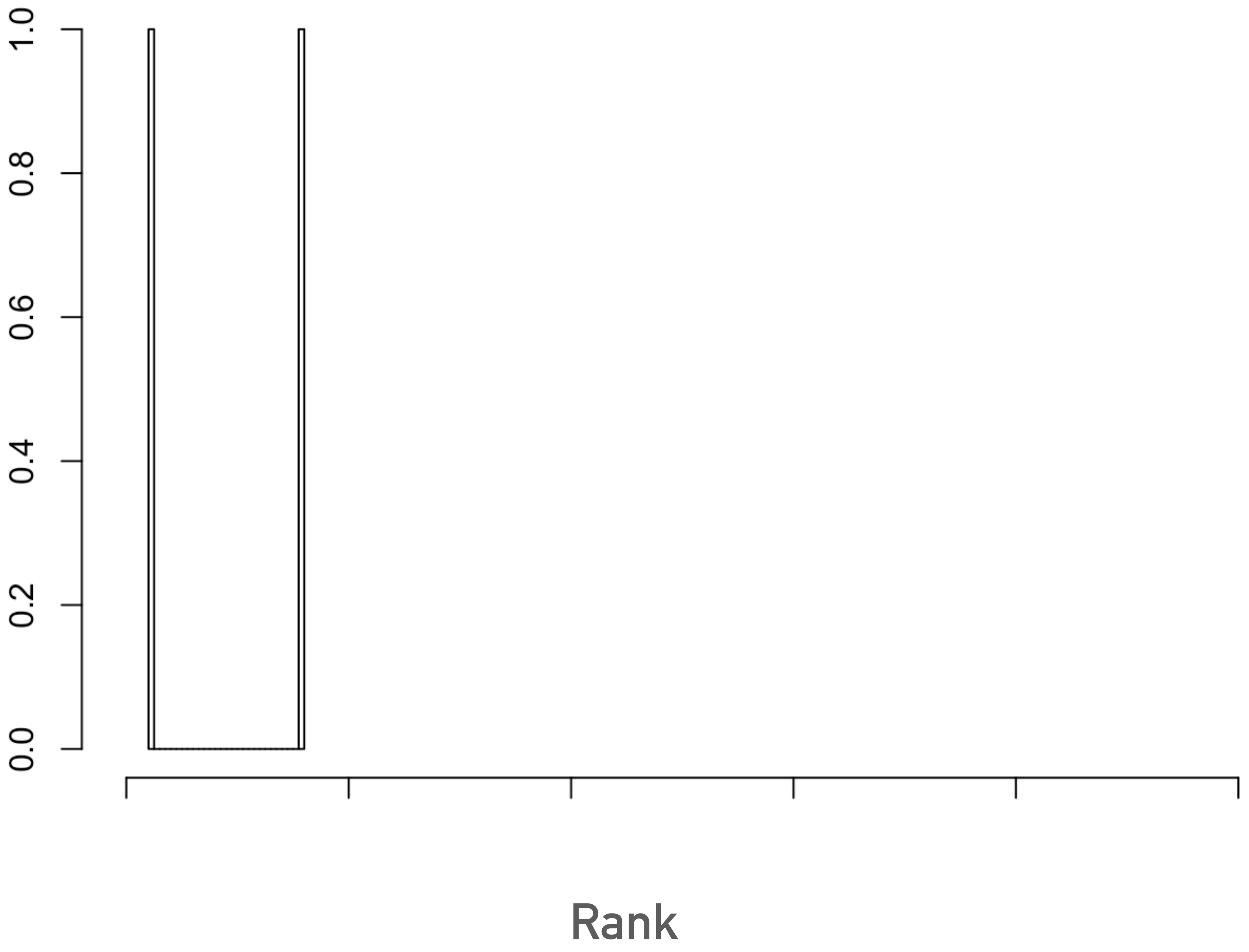
- We have a true value
- We have a bag of (approximately independent) posterior samples
- We can just look at where the true value lies in the bag of samples
- We look at the **rank** of the true value within the sample
- What happens when we do it a lot of times?

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, Andrew Gelman (2018)
Validating Bayesian Inference Algorithms with Simulation-Based Calibration
arXiv preprint: https://arxiv.org/abs/1804.06788

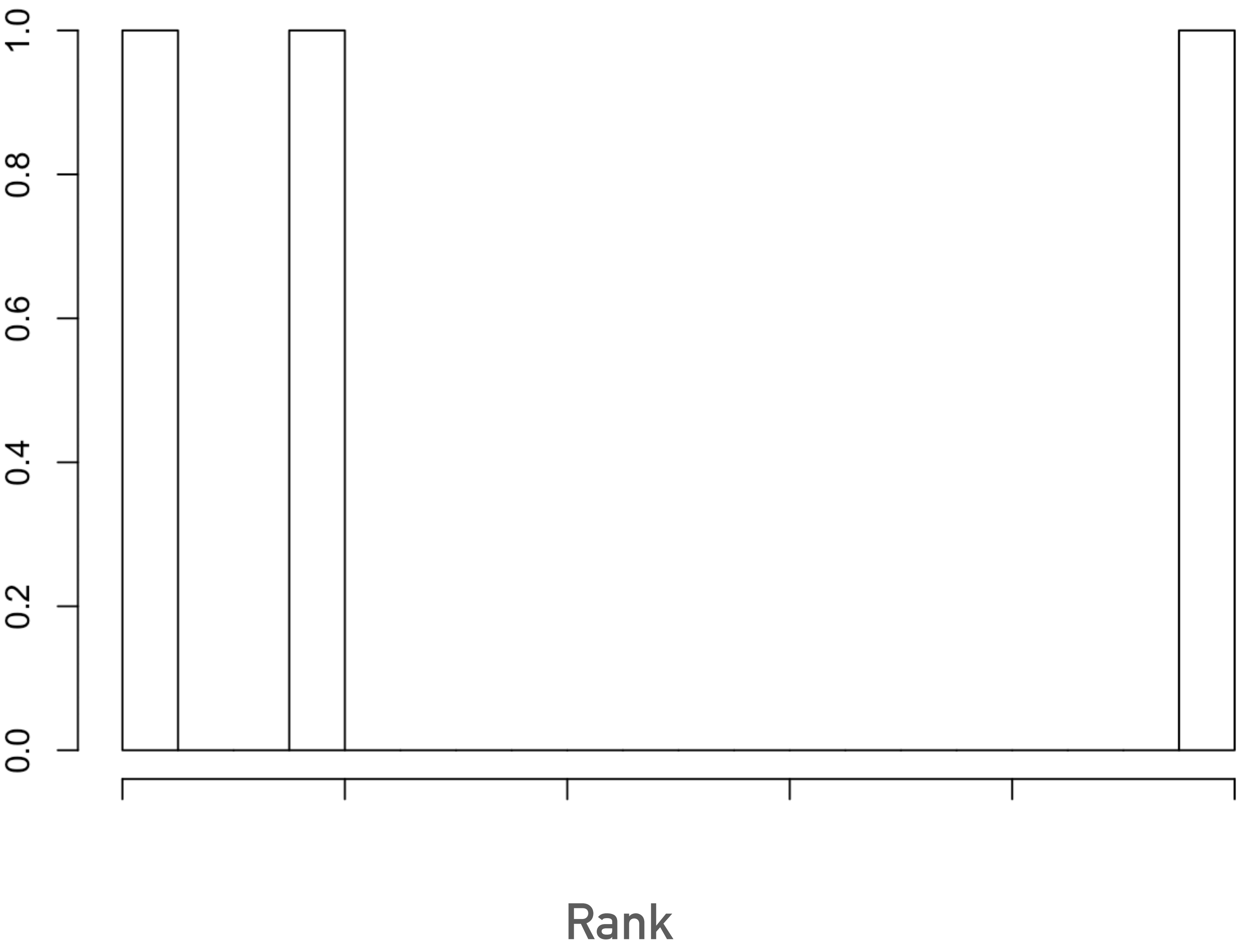
SINGLE RECOVERY



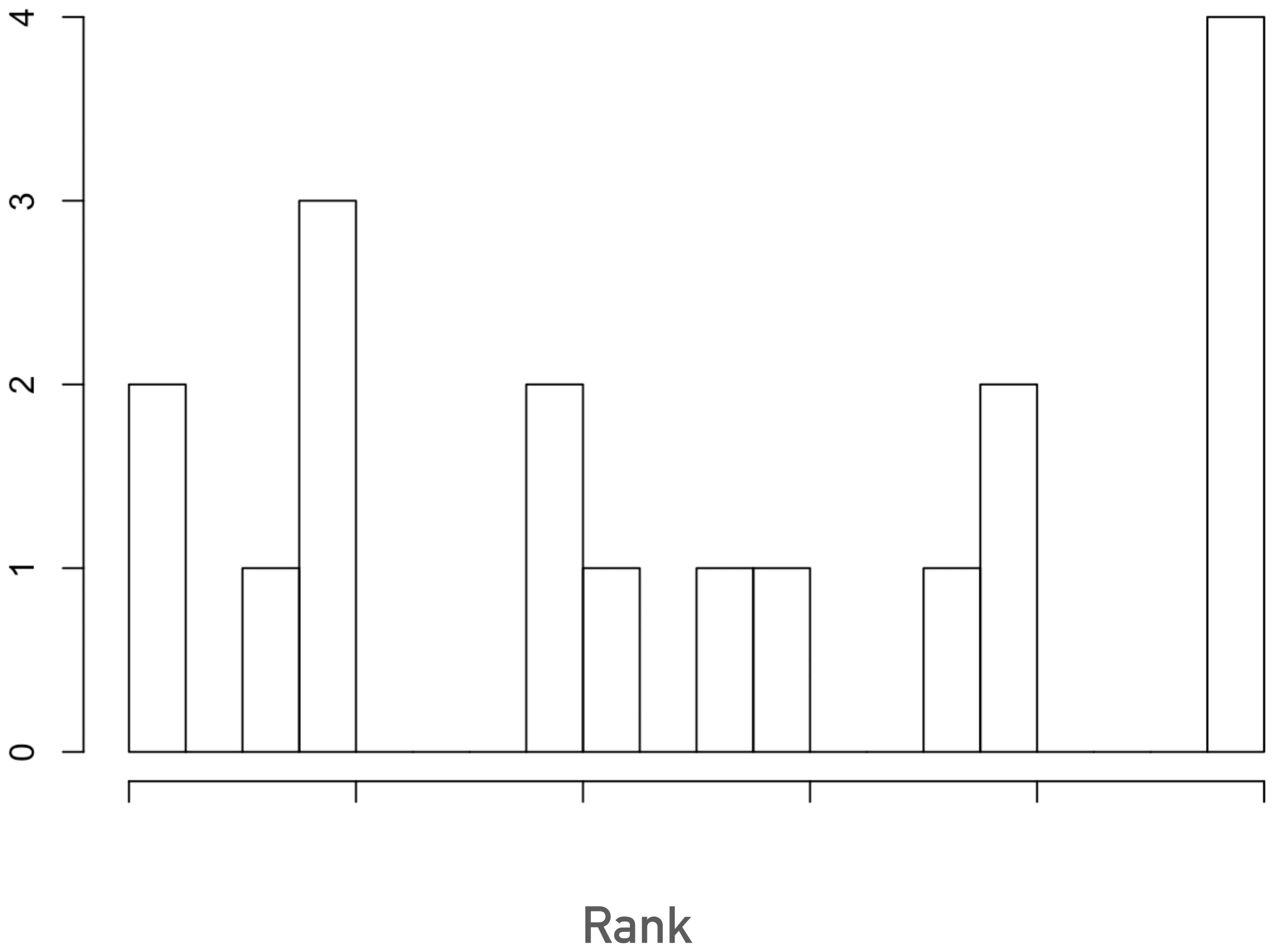
MULTIPLE RECOVERY



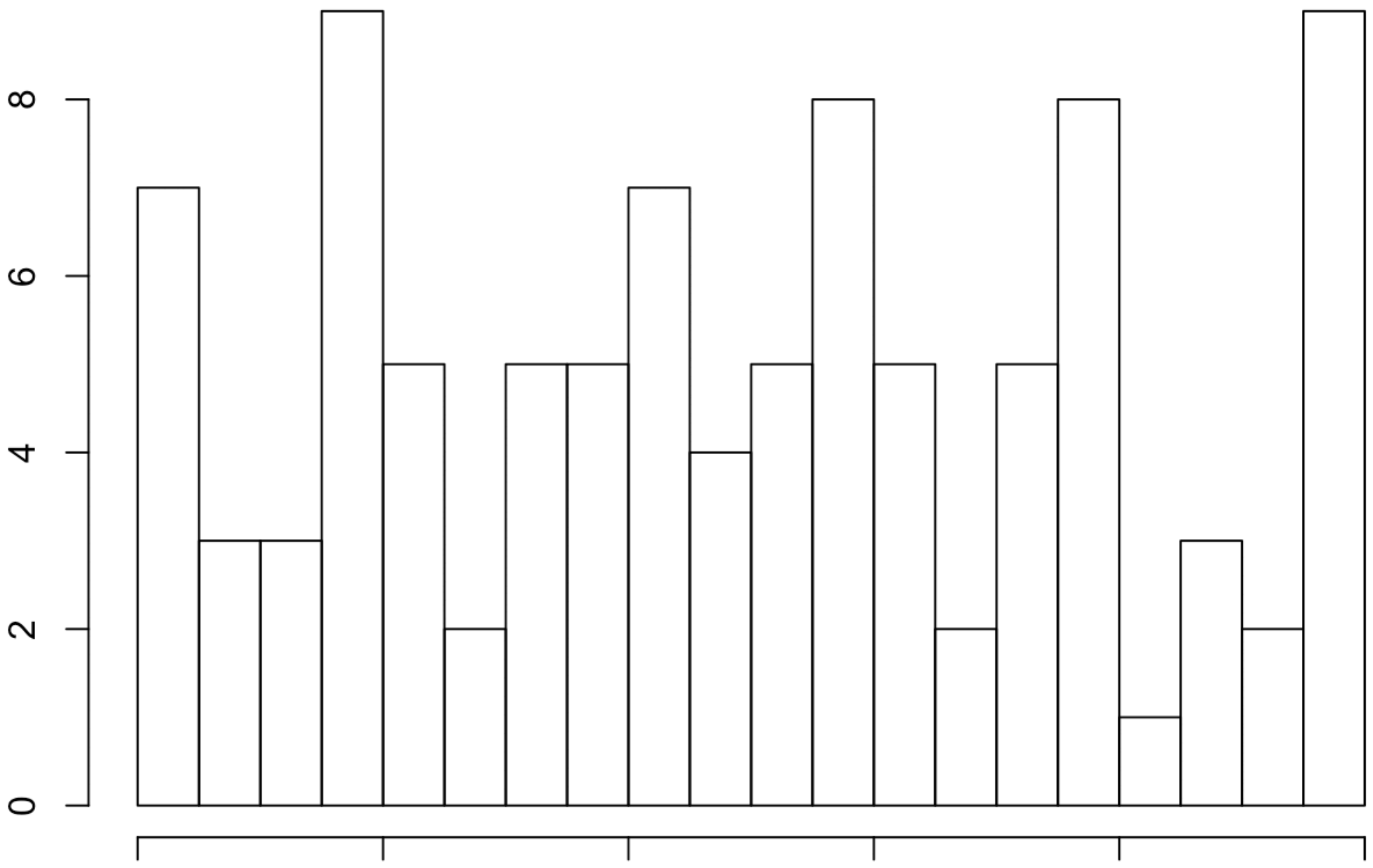
MULTIPLE RECOVERY



MULTIPLE RECOVERY

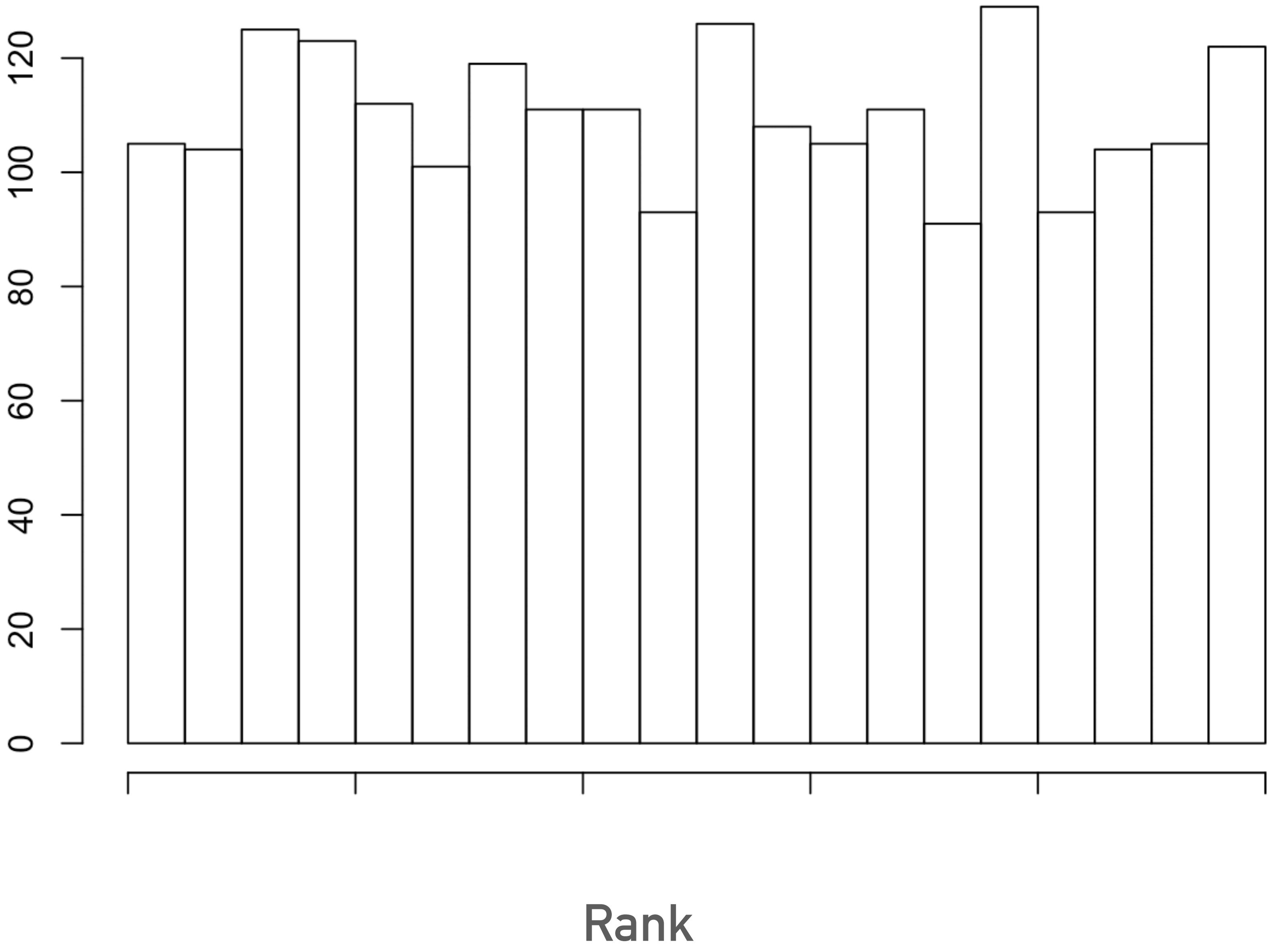


MULTIPLE RECOVERY



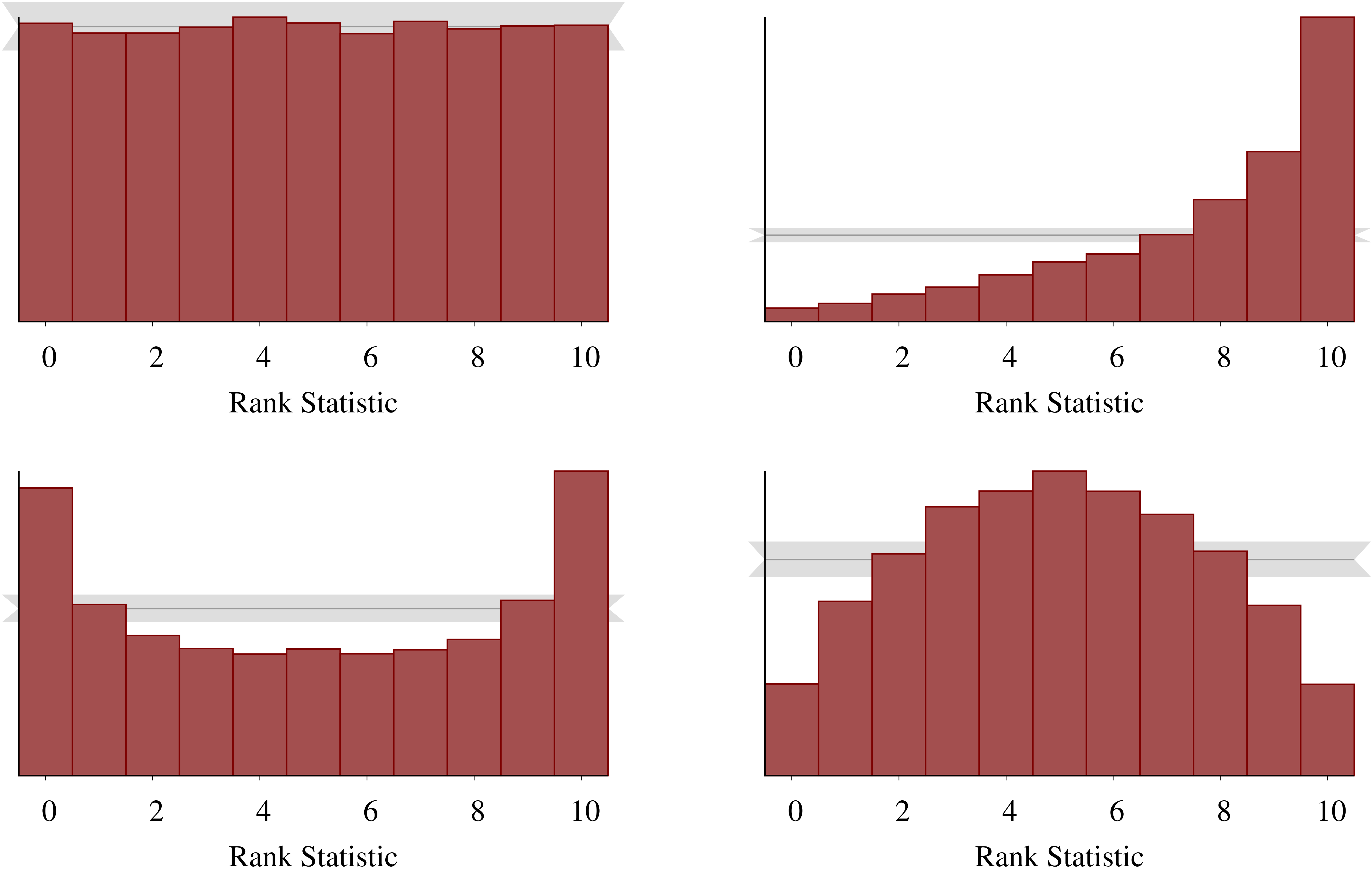
Rank

MULTIPLE RECOVERY



x

THE DEVIATION FROM UNIFORMITY IS MEANINGFUL



BUT THAT CAN BE QUITE EXPENSIVE

- We call this method **Simulation-Based Calibration (SBC)**
- **We** can run this on a cluster and it’s not too bad, but it is a problem
- But this is still a super-expensive thing to do!
- And it doesn’t guarantee that any particular run is reliable?

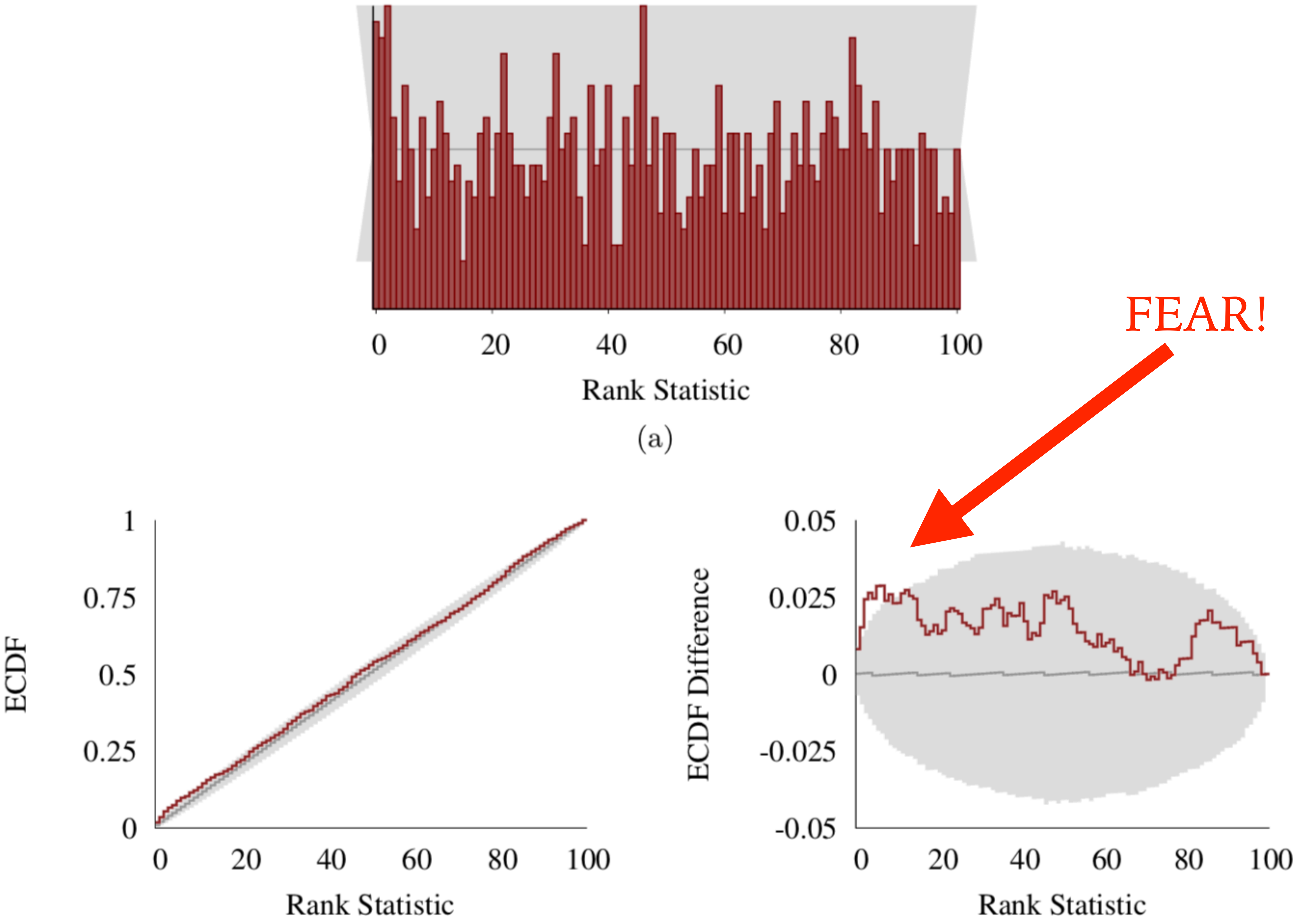
WHAT ABOUT MULTIPLE PARAMETERS?

- Two options: Random linear combinations or **substantively important quantities**.
- **Example:** Spatial mapping of HIV prevalence from survey data (Wakefield, Simpson, Godwin, 2016).
- Spatial binomial regression (simplest case intercept + GP)
- We care about area averages:

$$\frac{1}{|A|}\int_A \operatorname{logit}^{-1}(\beta_0 + S(x))\,dx\,.$$

- We fit the model with INLA, which is known to be a bit tetchy with rare data and binomial likelihoods

THE PICTURES DIDN'T GIVE ENOUGH FEAR

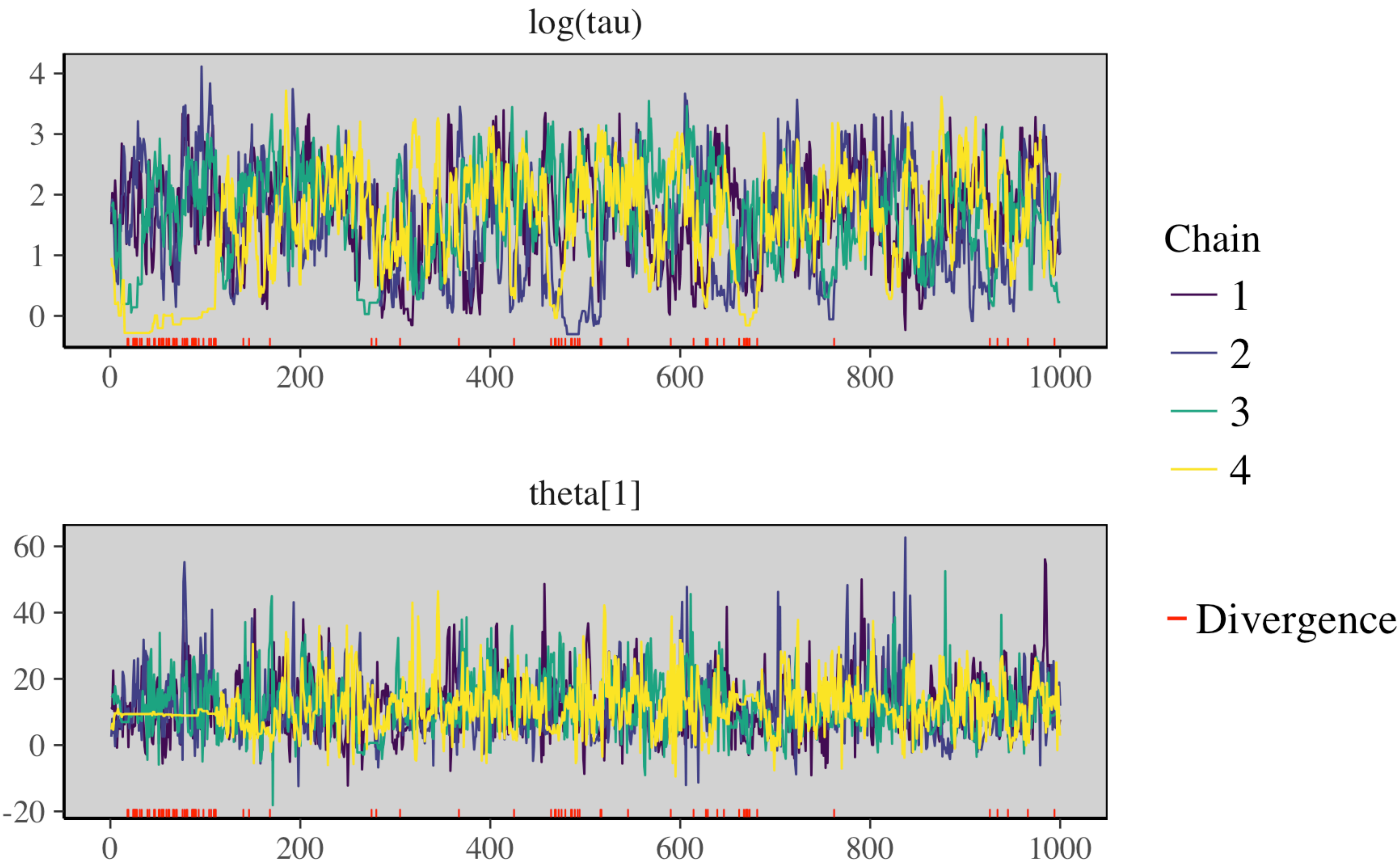


(a)

FEAR!

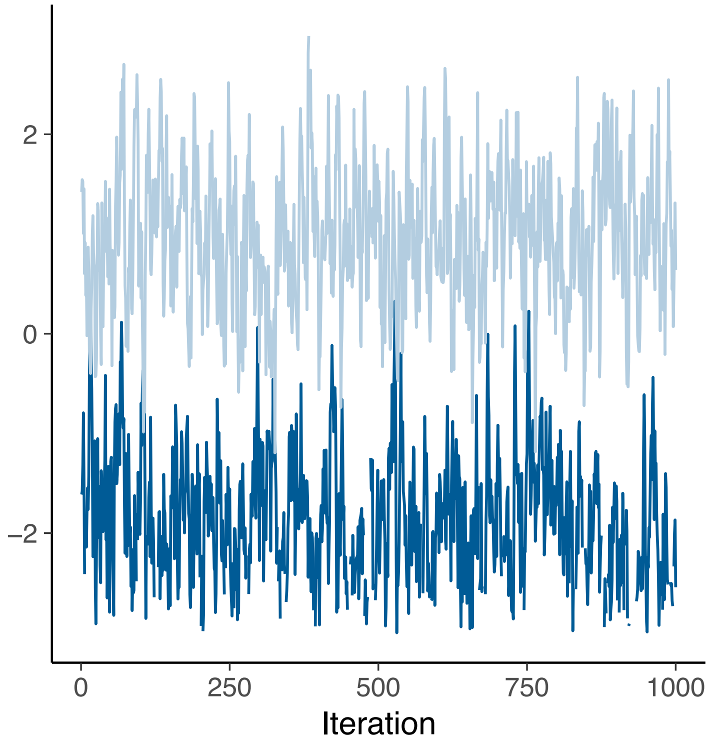
THE GODS SPIRITED HELEN TO EGYPT, SENDING TO TROY AN EIDOLON INSTEAD

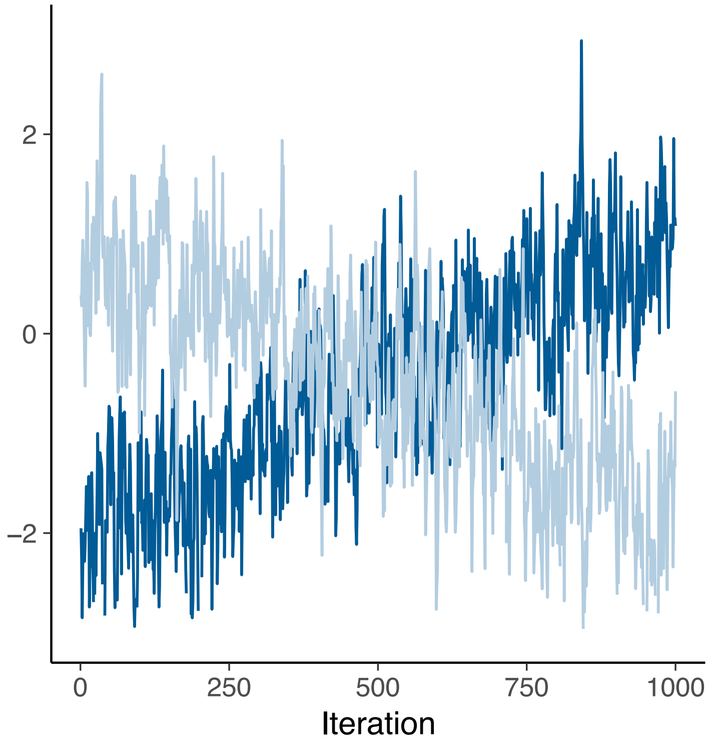
HOW CAN WE TELL IF MCMC WORKS?



HOW ABOUT SOMETHING MORE USEFUL?

- First things first, we need more than one chain:
  - Multimodality?
  - Bad adaptation?
  - Unlucky starting point?



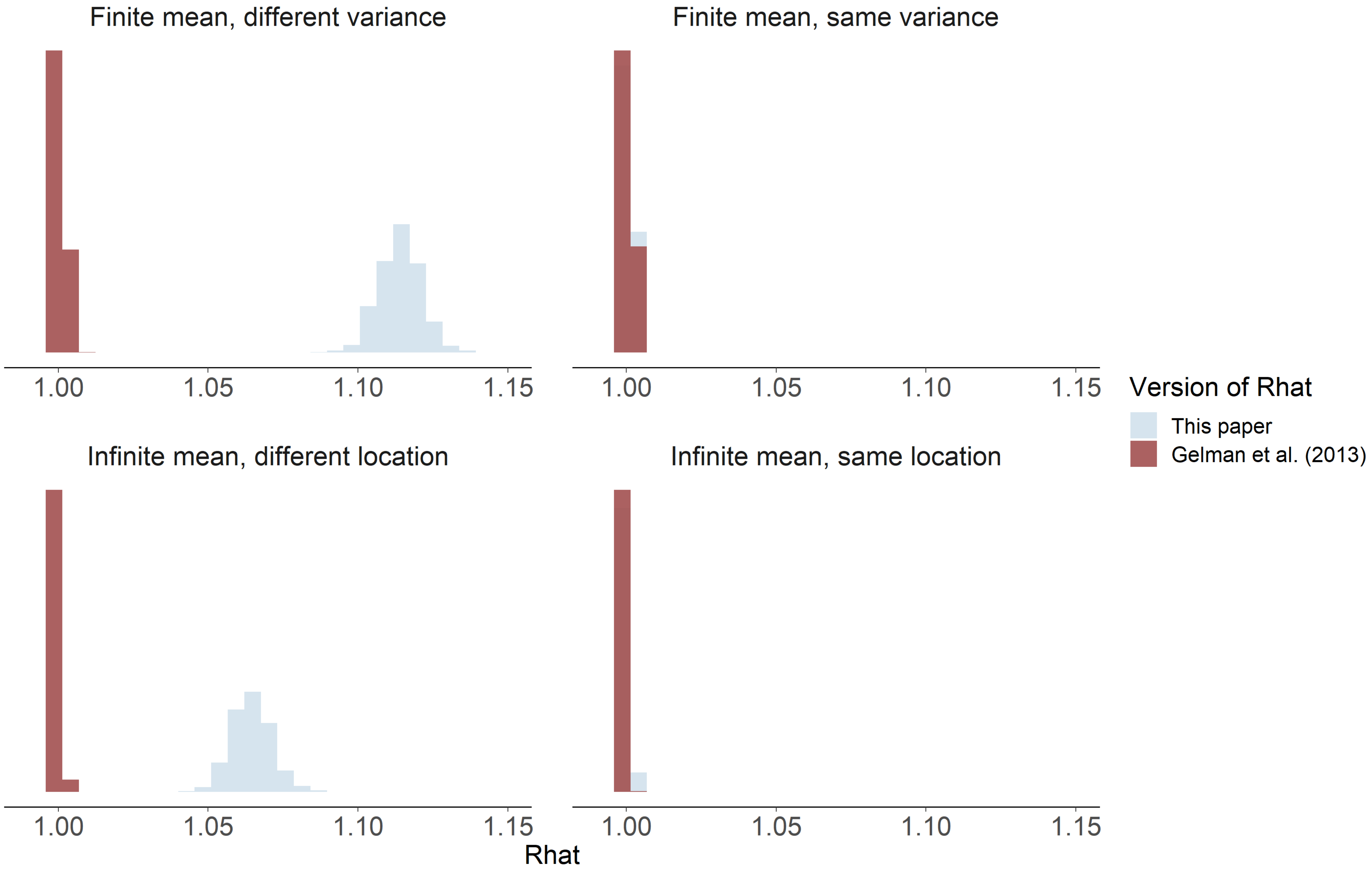


R-HAT: A GENERIC CONVERGENCE HEURISTIC

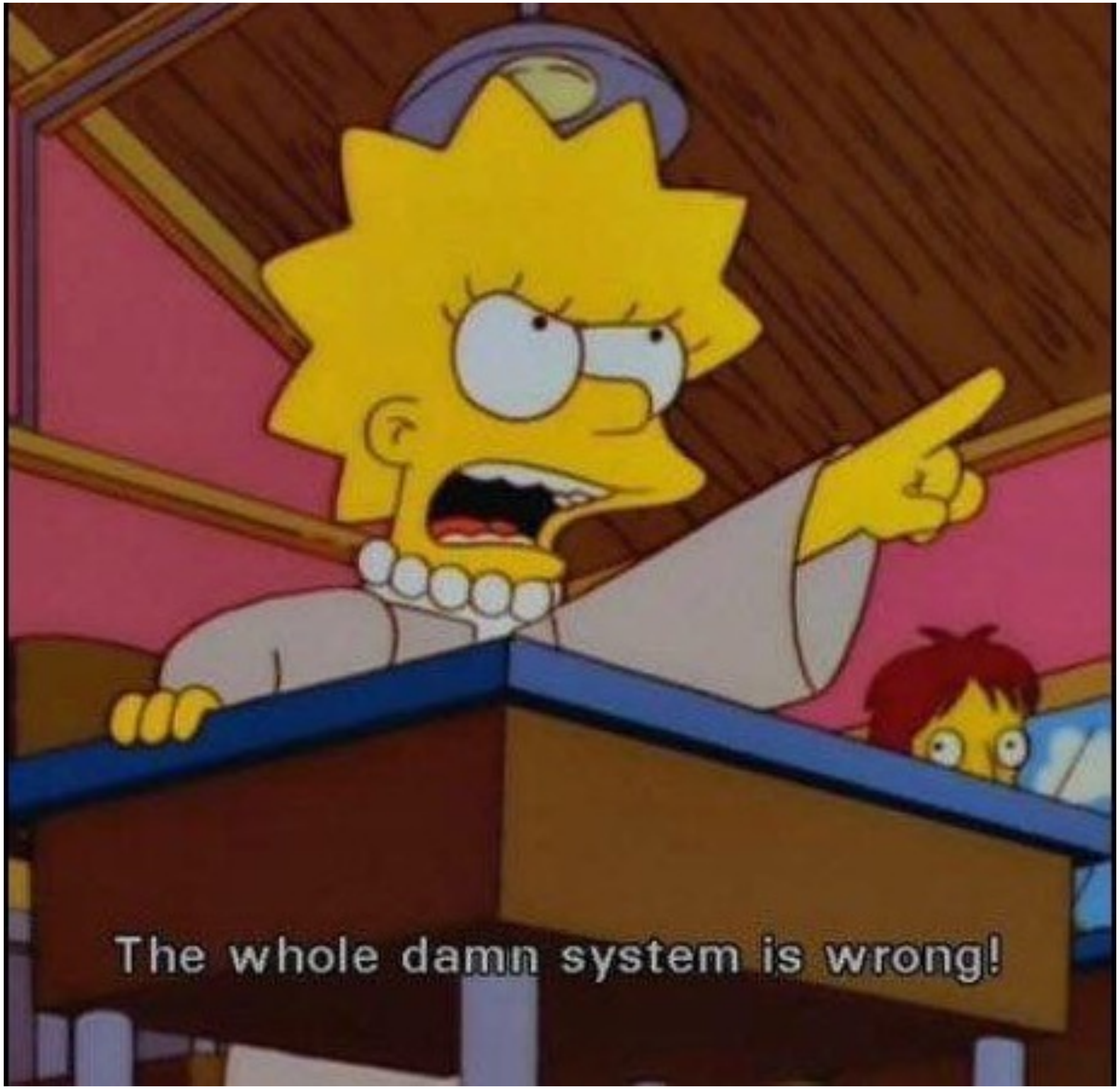
- This is an old idea due originally to Andrew Gelman

1. Run the same MCMC algorithm from M diffuse starting points or N samples
2. Compute the **between** chain variance B (the variance of the within chain mean compared to the overall mean)
3. Compute the **within** chain variance W (the sum of the variances within each chains)
4. Compute $\hat{R} = \sqrt{1 - \frac{1}{N} + \frac{1}{N}\frac{B}{W}}$

ONE SMALL PROBLEM…



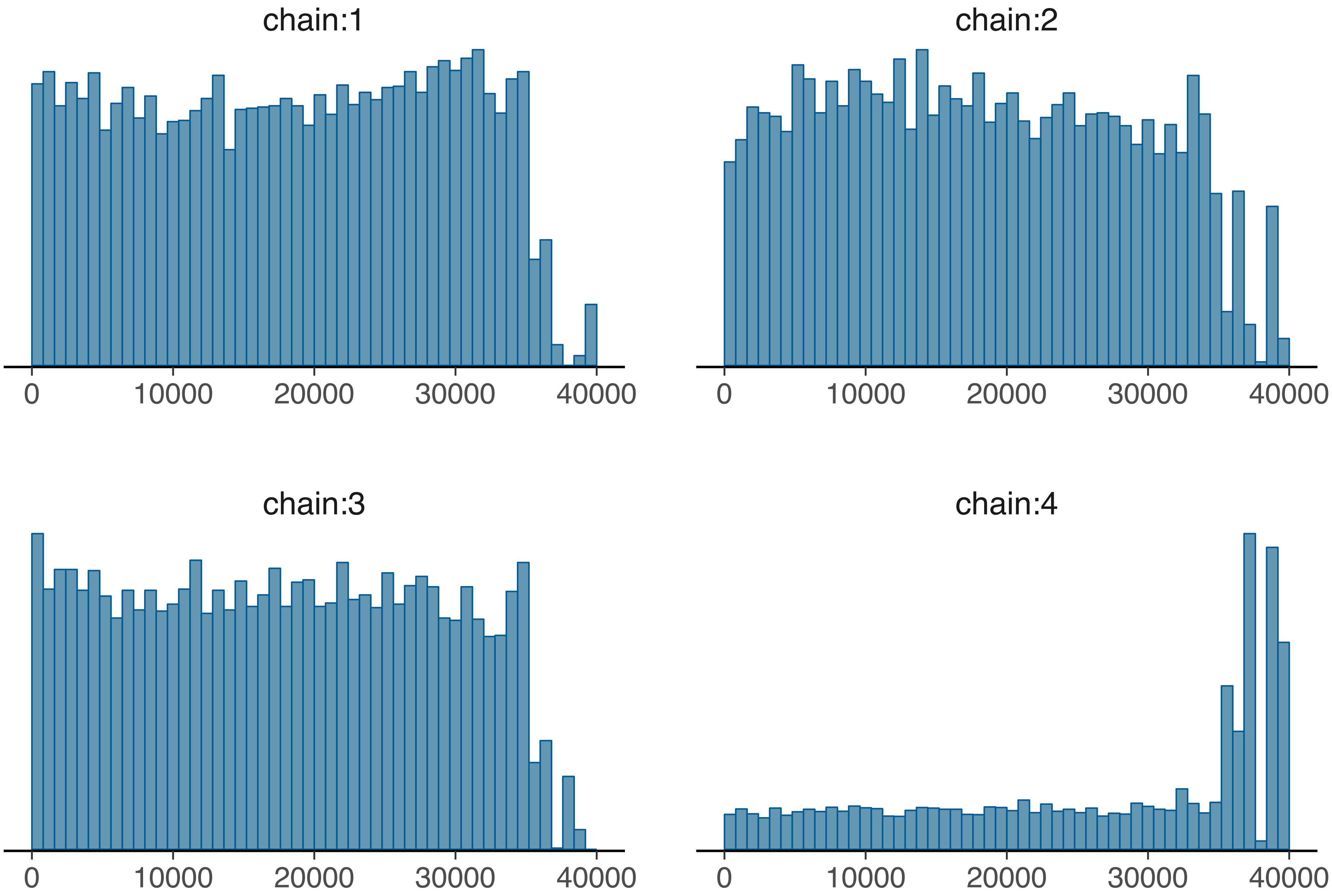
OH DEAR…



MAKING BETTER NUMBERS TO MAKE MORE FEAR

- We can fix these problems by doing two things:
  - Transforming the sample to make sure it has finite variance
  - “Folding” the sample around its median to check the second order properties.
- But the end point is: don’t just **trust** diagnostics. Actually check if they work!
- Numerical summaries have much less information than a well-chosen picture, but sometimes you can’t look at thousands of figures.
- So use it as a flag to see what you should fear and then use pictures to understand your fear.

RANKS: BETTER THAN A TRACEPLOT



STESICHORUS WAS STRUCK BLIND

THERE IS NO OTHER TROY FOR ME TO BURN

- It is very easy to spend a lot of time on statistical and theoretical minutiae.
- In the end, everything we are trying to do is extremely hard.
- There is a limit to what our theory can achieve.
- But there is no limit to the way that we can spin out the clever insights theory offers.

When Menelaus washed up on the shore in Egypt and was found by Helen, he refused to believe it was her until word reached him that the Helen he had hidden in a cave for her safety had evaporated into thin air.

“

There is no truth in that story,
You didn't ride in the well-rowed galleys,
You didn't reach the walls of Troy.

-Stesichorus (tr. Anne Carson)

SOME REFERENCES

- Kennedy, Lauren, Daniel Simpson, and Andrew Gelman. "The experiment is just as important as the likelihood in understanding the prior: A cautionary note on robust cognitive modelling." arXiv preprint arXiv:1905.10341 (2019).
- Jonah Gabry, Daniel Simpson (Joint first author), Aki Vehtari, Michael Betancourt, and Andrew Gelman (2018). Visualization in Bayesian workflow (with Discussion). Journal of the Royal Statistical Society Series A. Volume 182(2), pp. 389–402.
- Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, Andrew Gelman (2018). Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv preprint: https://arxiv.org/abs/1804.06788
- Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. "Pareto smoothed importance sampling." arXiv preprint arXiv:1507.02646 (2019).
- Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. "Rank-normalization, folding, and localization: An improved for assessing convergence of MCMC." arXiv preprint arXiv:1903.08008 (2019).