Probabilistic Biostatistics with julia

Jeff Mills, University of Cincinnati
Jeff Strawn, University of Cincinnati and
Cincinnati Children's Hospital

- Thanks to:
- The **JuliaCon organizers** for the opportunity to present.
- The **Yung Family Foundation** and **Taft Research Foundation** for research support.
- and credit due to **many other colleagues and coworkers**.

What is “probabilistic” statistics?

- Application of probabilistic logic:
- From If A then B, to If A then B more plausible.
- Definition of probability as ‘degree of plausibility’.
- ***Any*** proposition can be assigned a degree of plausibility based on the available information.
- All degrees of plausibility, and so probabilities, are conditional.
- $0 \leq P(B|A) \leq 1 =$ probability of B if A is true.

Probabilistic Logic → Probabilistic Statistics

- $P(B|A) =$ probability of B if A is true.
- B is often (usually) neither random nor a variable.
- We are interested in the plausibility of propositions faced with incomplete information, and so uncertainty.

- Probabilistic statistics is the application of probability to making inferences with incomplete information (the data).

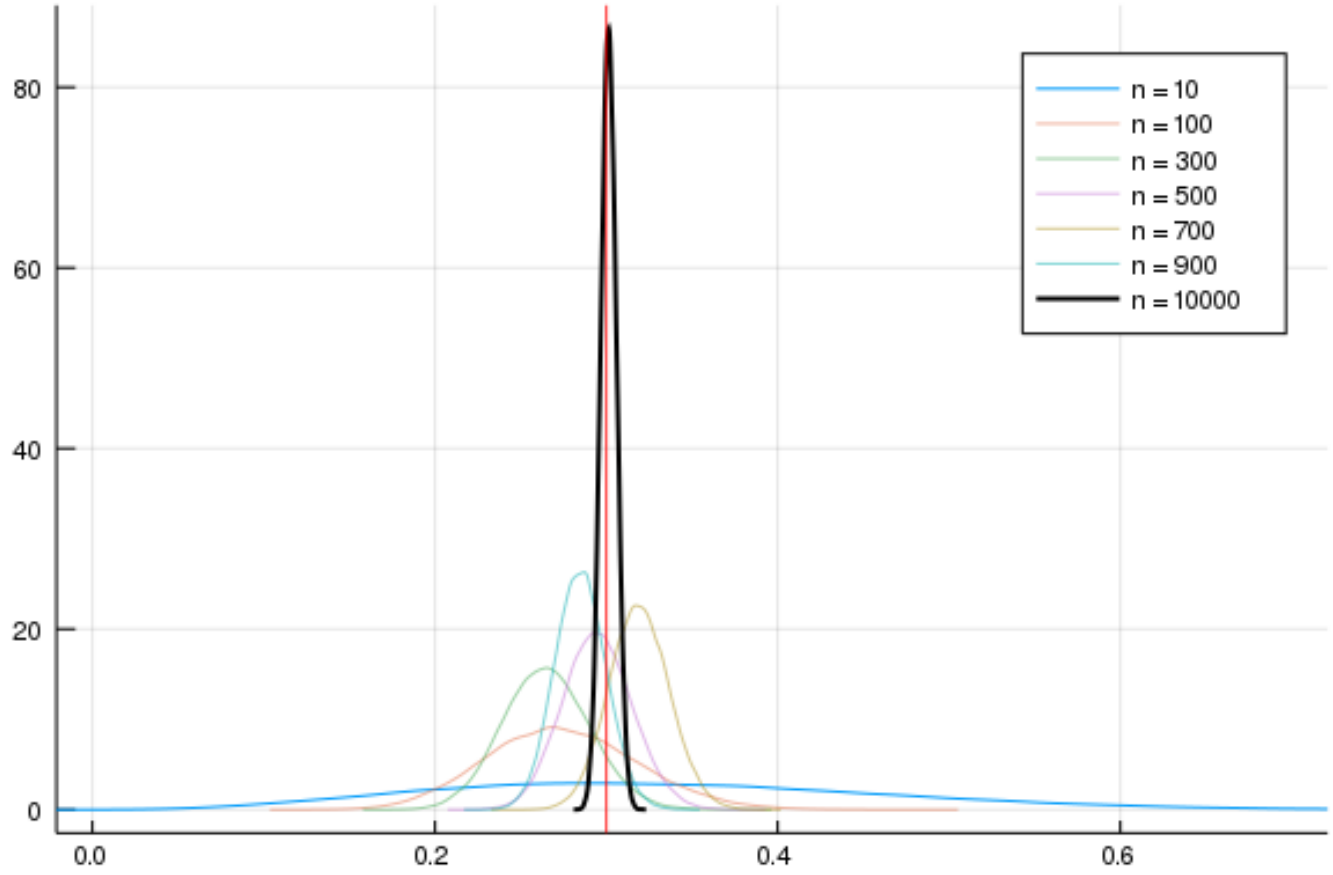
- Jaynes (2004) *Probability: The Logic of Science*

Probabilistic statistics

- Simplest case: s successes in n, with unknown probability of success, θ.
- Common misunderstanding: θ is a constant, not a random variable. Bayesian inference does **not** assume θ is random.
- The posterior distribution represents uncertain knowledge about the unknown constant.
- E.g. `θ = 0.3; Random.seed!(41); x = rand(Bernoulli(θ),10)`
- If we know the RNG algorithm and the seed, with only a few observations we can determine θ exactly.
- With less information (only the sample, x) we can represent our uncertain knowledge of θ probabilistically.

Uncertain knowledge about an unknown ≠ the unknown is random

- The posterior distribution represents uncertain knowledge about the unknown constant, θ.
- Data generating process:
- `x = rand(Bernoulli(θ),n)`
- As $n \rightarrow \infty$, our knowledge about the unknown parameter approaches certainty.



- For $n = 10,\ x = [0, 1, 0, 1, 0, 0, 0, 1, 0, 1]$, unknown $\theta = 0.3$

Probabilistic Computation: A Fairy Tale

- The axioms of probability lead directly to Bayes' Theorem.
- Bayes' Theorem used as a rule for inference => analytical intractability as our problems become more complex.
- Requires multi-dimensional integration => the curse of dimensionality is the main problem.
- Stan Ulum, mathemagician, in hospital, couldn't solve a complex combinatorial problem, so he invented the Monte Carlo method and solved the problem probabilistically.
- He recovered, went back to Los Alamos and ran MC simulations on the computer (ENIAC).
- => MC simulation and Metropolis methods (1940s, 50s)

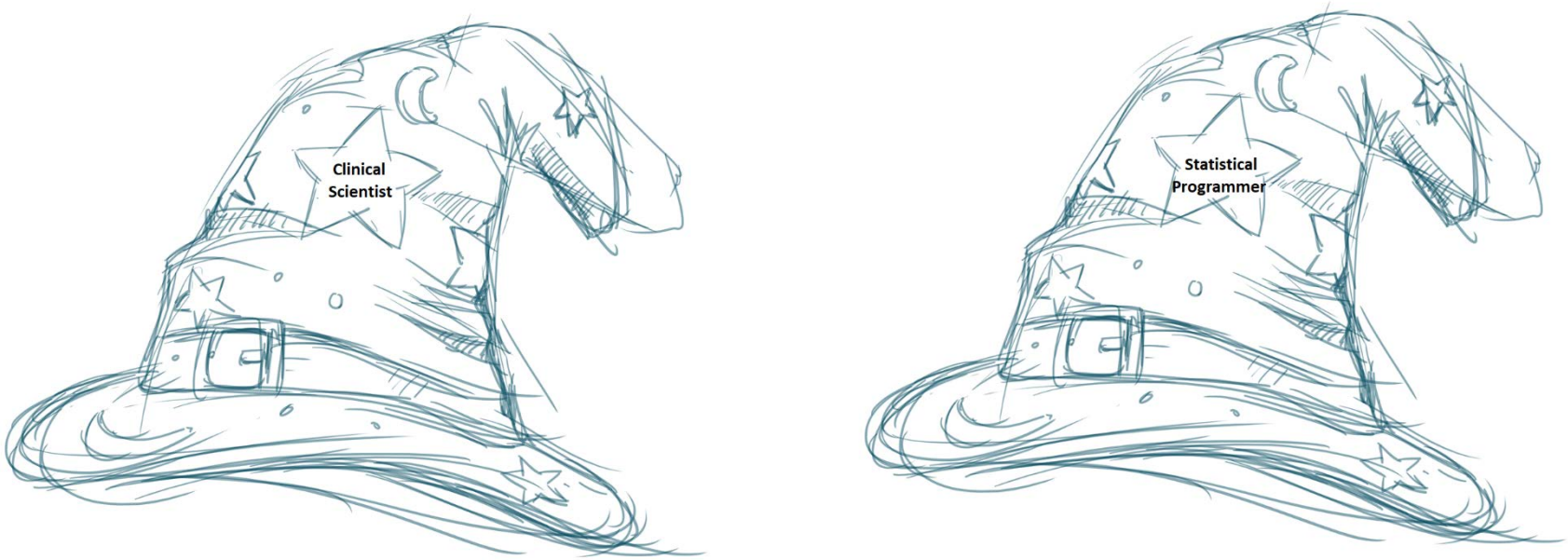
Probabilistic Computation: solving complex problems with uncertain answers numerically

- About 50 years later, Bayesian statisticians realized what this meant for probabilistic reasoning and Bayesian inference.
- MCMC methods use computational power and pseudo-randomization to bypass the curse of dimensionality.
- So, we have an unknown fixed quantity, say θ, we are interested in evaluating or using for prediction.
- The information available concerning θ, the model specification and the data, are insufficient to determine the value of θ with certainty.

Probabilistic Biostatistics

- Analyzing Randomized Controlled Trials (RCTs):
- Simplest case: s successfully treated out of n patients with unknown constant probability of success, θ.
- Interest in underlying probability of success for a new patient, θ.
- For quantitative data, interest in comparing average treatment effect (ATE) for different treatments (or treatment vs. placebo).
- Interest in combining evidence from several RCTs
- Interest in obtaining evidence from clinical practice ('evidential medicine').

The two (or more) field (or hat) problem



Problem: We often must wear multiple expertise ‘hats’ when conducting scientific research. Research teams are expensive, so often the nonexpert has to gain some expertise.

Solution: Julia + computational probabilistic inference

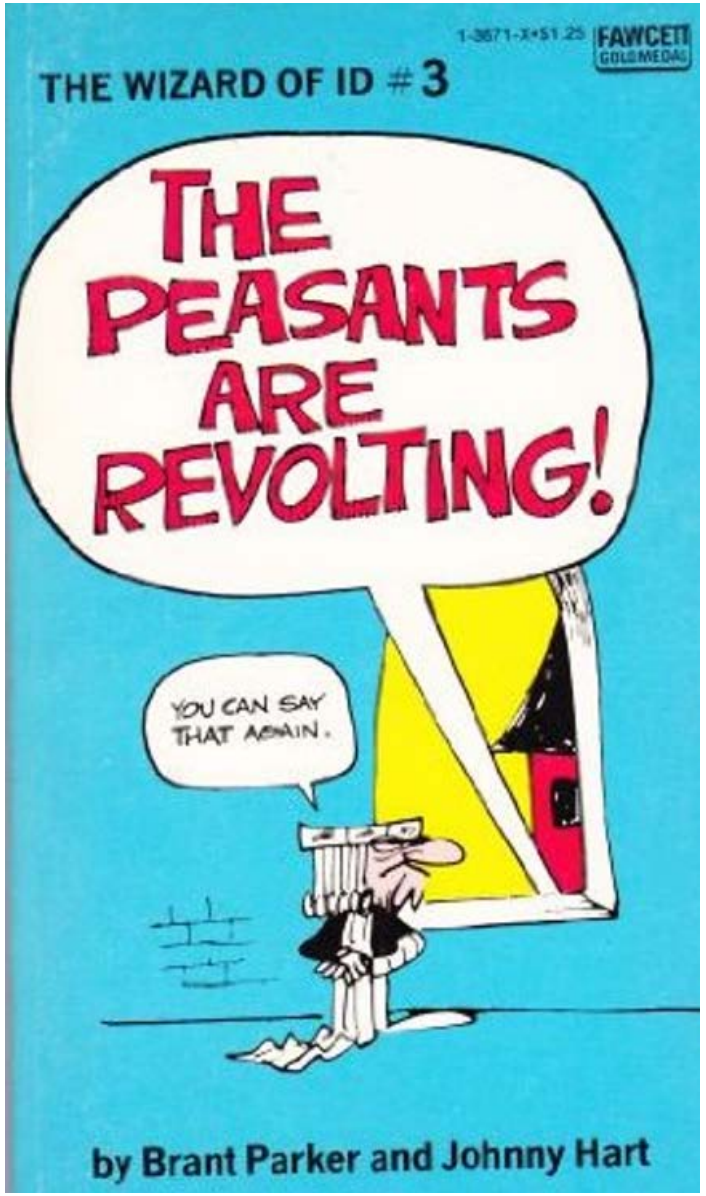
MCMC + julia =  (**K**eeping **I**t **S**ophisticatedly **S**imple)

- The “Bayesian machinery” of Markov chain Monte Carlo (MCMC) methods together with Julia offer a solution to this “two-field problem”.
- They enable exact small sample inference and hypothesis testing for complex models without requiring the restrictive assumptions necessary to obtain analytical tractability (*performance*).
- They facilitate the analysis of complex models with basic statistical concepts: frequency distributions, density plots, means, medians, modes, standard deviations, quantiles, and posterior odds (*user friendliness*).

Overthrowing the dictatorship of *p*-values

- The prevalence of "five percentitus": Despite widespread and persistent criticism of p-values, the "$p \leq 0.05$ is 'statistically significant', $p > 0.05$ is not" is an iron law for publishing in leading journals in many fields.
- The advances in computing power and Bayesian computational methods in the last several decades are only just beginning to have an impact on this.
- Goal: an alternative to p-values

  (80+ years is enough?!).



Overcoming institutional inertia and bias

- Strong **institutional inertia and bias** towards precise testing and use of p-values as `proper' applied science (same methods 50+ years on!).
- At the very least, we need **a bridge from `pure' hypothesis testing to Bayesian inference**.
- **Problem**: Strong institutional bias/inertia of 5% = `significant'.
- **Solution**: Provide posterior odds as well as posterior tail probabilities (`Bayesian *p*-values') and HPDIs.
- **Caveat**: "A picture paints a thousand statistics" - **Visualization** often provides more valuable evidence than a test.
- (See last year's presentation and Bayestesting.jl on github)

Applications of Probabilistic Biostatistics with julia

- RCTs are high cost (e.g., medication expense, time required, and potential exposure of patients to ineffective treatments)
- => greater enthusiasm for improving statistical methods:
- (1) for RCTs
- 2) for ‘naturalistically-collected’ clinical data to inform clinical practice without the need for RCTs.
- Goal: Use and provide statistical tools to clinician-researchers that are intuitive and easy to use, yet sophisticated and powerful enough “under the hood” to answer complex questions.

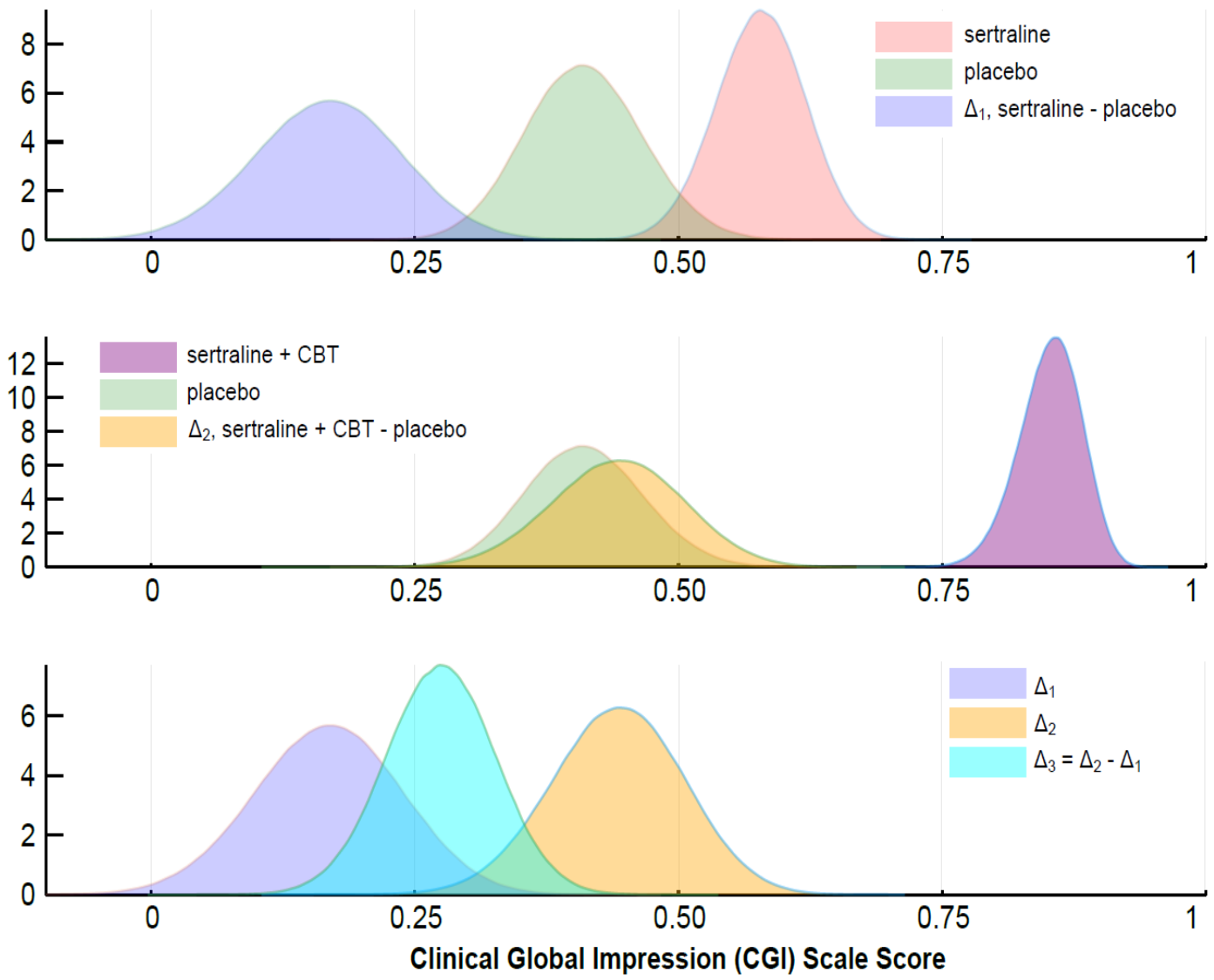
CAMS: Child/Adolescent Multimodal Study

- Evaluate treatments for anxiety in children and adolescents:
- **SSRI (**selective serotonin reuptake inhibitor, sertraline)
- **CBT** (cognitive behavioral therapy)
- **SSRI** + **CBT**
- **Placebo**
- Study participants were randomized 2:2:2:1. Over 400 participants.
- Anxiety measured by the quantitative **PARS** rating scale measure and the categorical **CGI-I** rating scale (CGI ≤ 2 considered successful treatment).

Pseudo-random draws from the Beta posterior given s successes in n patients for each group.

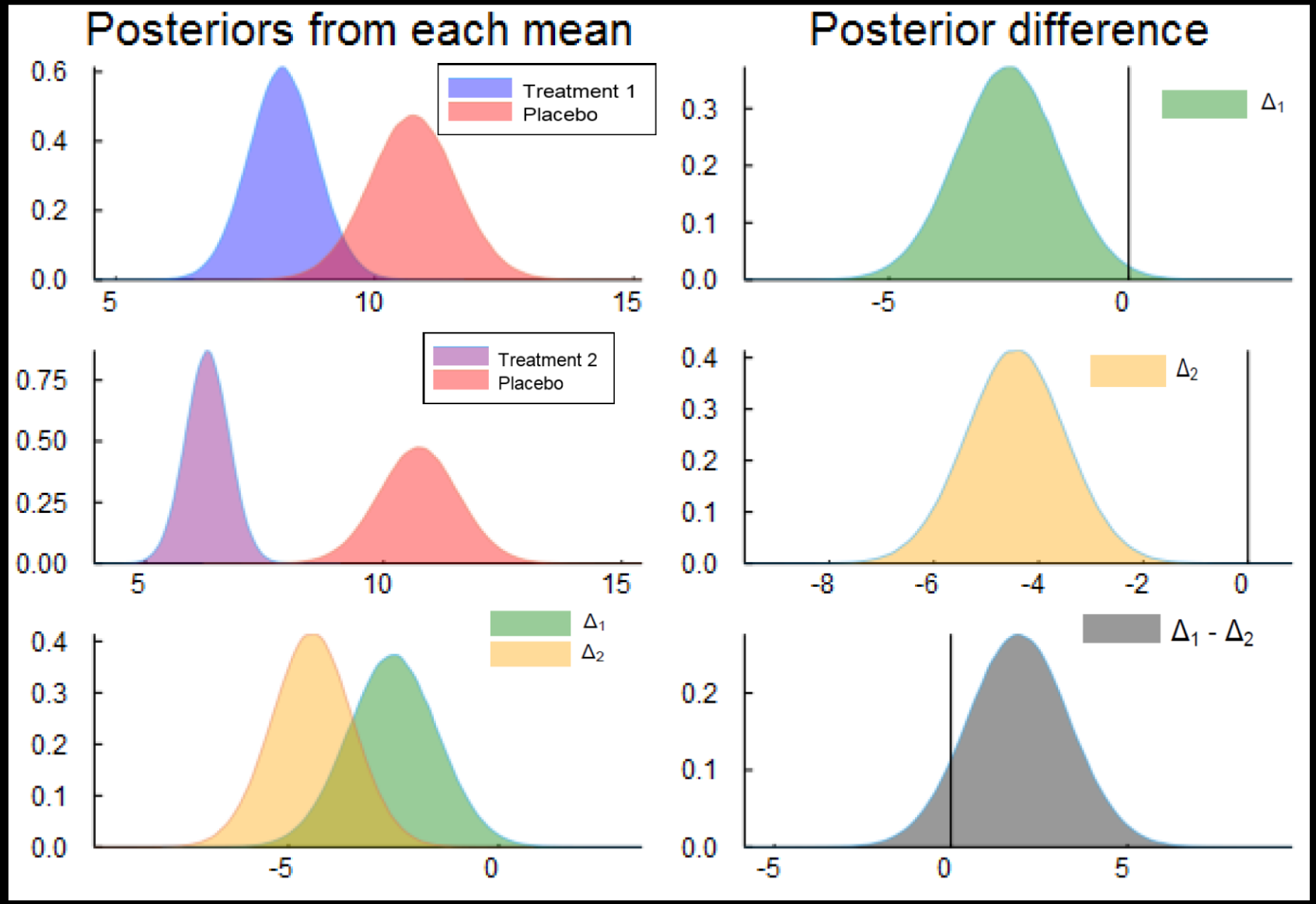
$$\Delta_1 = \theta_{T1} - \theta_{P1}$$
$$\Delta_2 = \theta_{T2} - \theta_{P2}$$
$$\Delta_3 = \Delta_2 - \Delta_1$$

An analytically intractable problem, but just a few lines of code.



Difference in differences (analytical exact distribution unknown)

	Odds	*p*-value
Δ_1	15.73	0.0194
Δ_2	24400.8	<0.0001
Δ_3	2.45	0.1832



Sequential Bayesian updating

For Binomial likelihood with s successes in n trials and a $Beta(1,1)$ (i.e. uniform) prior:

$$\theta|n,s \sim Beta(s+1, n-s+1)$$

To update given additional observations s_2, n_2:

$$\theta|n,s,s_2,n_2 \sim Beta(s+s_2+1, n+n_2-s-s_2+1)$$

No prior stopping rule required

Sequential Bayesian updating

For a quantitative variable,

$$\text{CLT + Maxent} \Rightarrow \mu \sim N(\bar{x}, s^2/n).$$

With Normal-Gamma prior (start uninformative (i.e. large variance),

$$\mu|x \sim TDist\left(\frac{\sqrt{n(\mu-\bar{x})^2}}{s}, n-1\right),$$

$\bar{x}$ = sample mean, s = sample S.D.

Posterior Normal-Gamma parameter values, $\bar{\mu}$ = mean of posterior pseudo-sample, etc., become the new prior when more data arrives.

Sequential Bayesian updating

Alternatively, we can use MCMC. All that is needed is prior × likelihood. Iterate, $m = 1,2, \ldots M$:
mth draw from:

$$\mu_m|\sigma^2, x \sim N\left(\bar{x}, \frac{\sigma^2_{m-1}}{n}\right),$$

$$\sigma^2_m|\mu_m, x \sim IG\left(\frac{n-1}{2}, \frac{SSE}{2}\right), \qquad SSE = \sum_{i=1}^{n}(x_i - \mu_m)^2$$

Since sample mean, SD and n are sufficient statistics, updating formulas for Normal-Gamma allow posterior to be updated one observation at a time.

Bayesian updating without the math

The update-mean function in BayesTesting.jl:

Can be used repeatedly.

We know the posterior is a Student-t and m2, s2, n2 are sufficient statistics

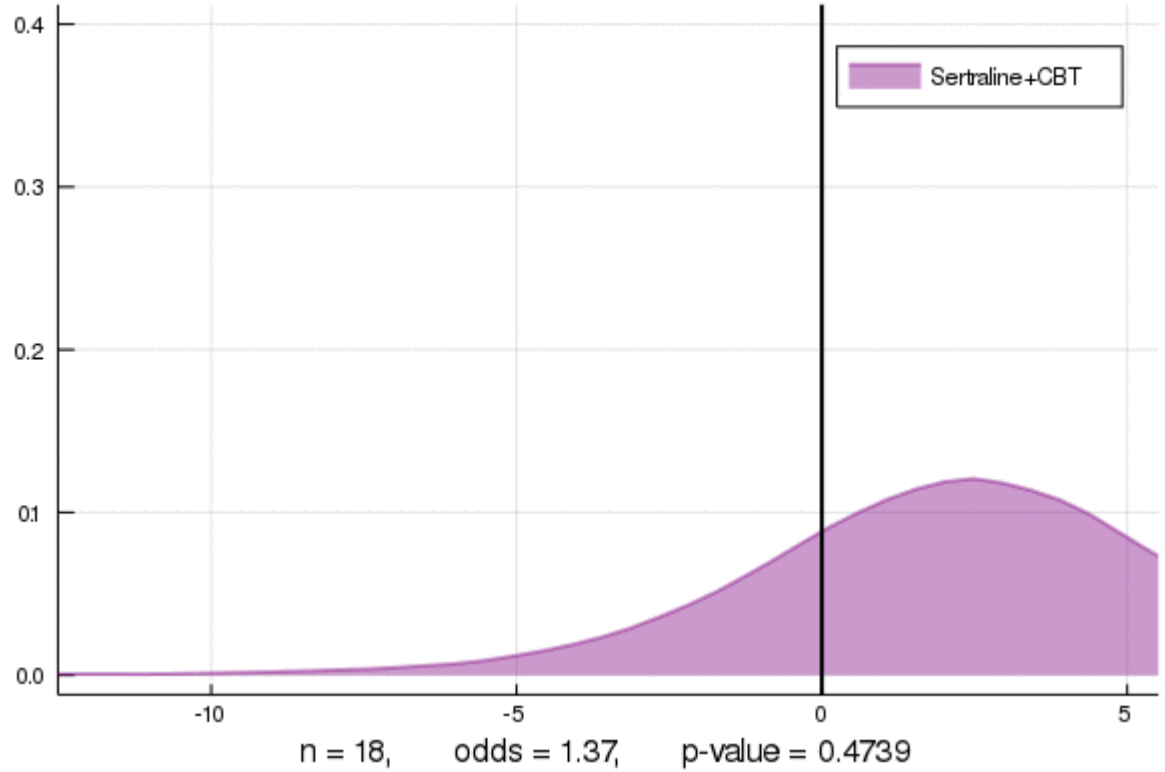
```
function update_mean(m1,m0,s1,s0,n1,n0)
  s21 = s1^2
  s20 = s0^2
  n2 = n1 + n0
  m2 = (n1/n2)*m1 + (n0/n2)*m0
  v2 = n2 - 1.
  vs22 = (n1-1.)*s21 + (n0-1.)*s20 +
                ((n1*n0)/n2)*(m1 - m0)^2.
  s2 = sqrt(vs22/v2)
  return m2, s2, n2
end
```

Sequential updating for CAMS data: when to stop?

Sertraline+CBT – Placebo

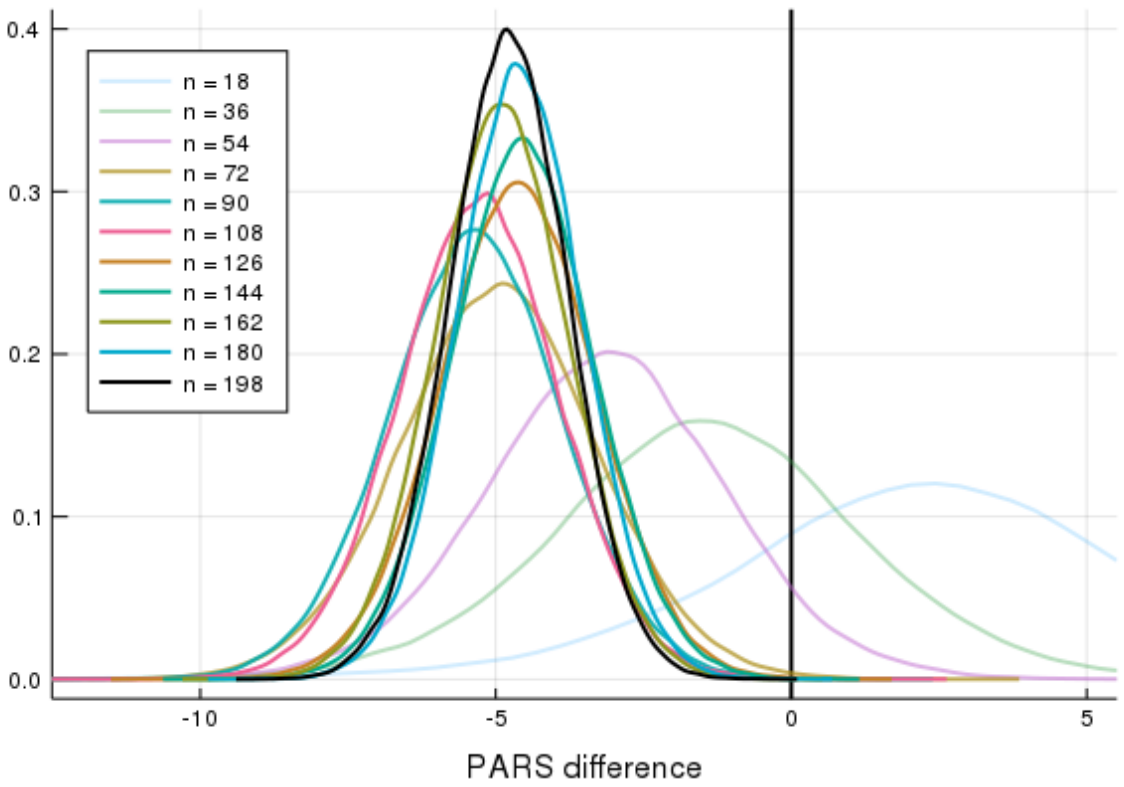
Updating the posterior density as data become available
(every 6 obs., 4 treatment, 2 placebo).
PDR = posterior odds against 0

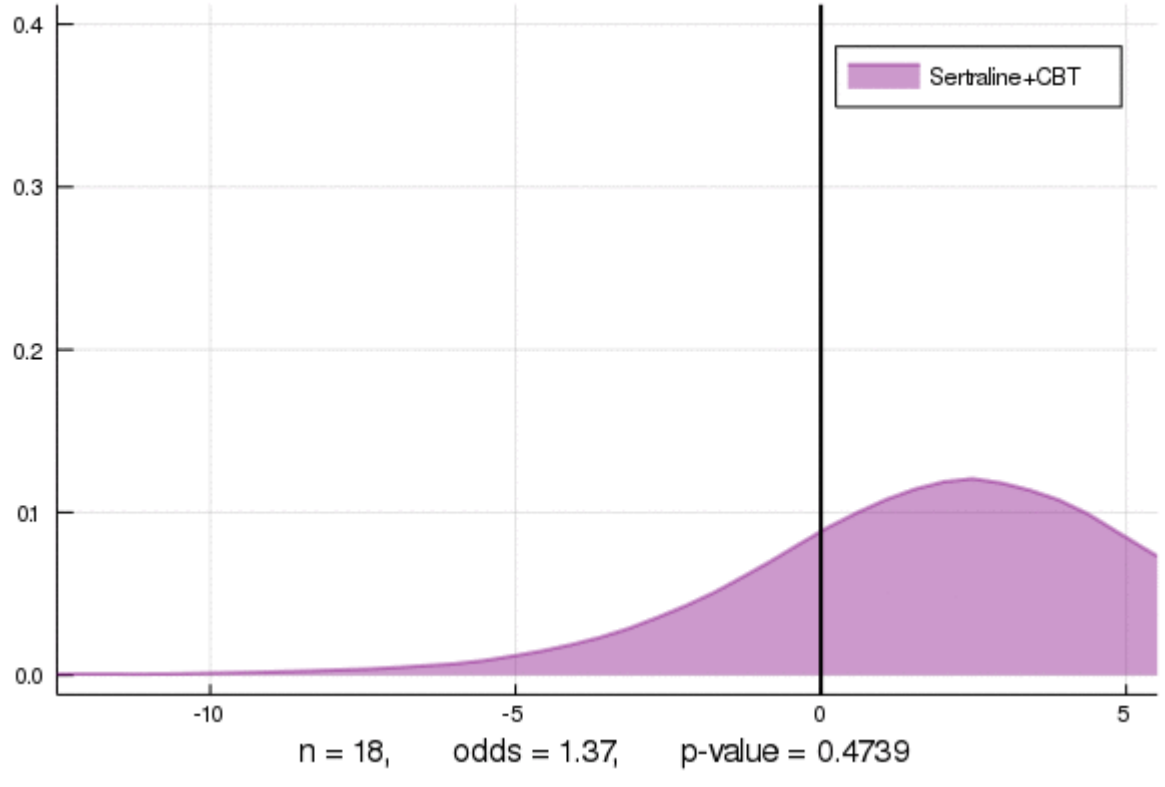
n	Posterior Odds	$p(\theta \geq 0 \mid s, n)$
18	1.4	0.763
63	7.1	0.027
72	64.4	0.002
141	718.3	0.0001
198	30629.5	<0.0001



Difference in ATE for PARS

PARS for Difference in ATE (Sertraline+CBT – Placebo)





Bayesian Meta-analyses: the BHM

- Bayesian Hierarchical Models (BHM) have a number of advantages when conducting a meta-analysis.

- To combine the evidence from several different studies (results from different RCTs). Only summary statistics may be available.

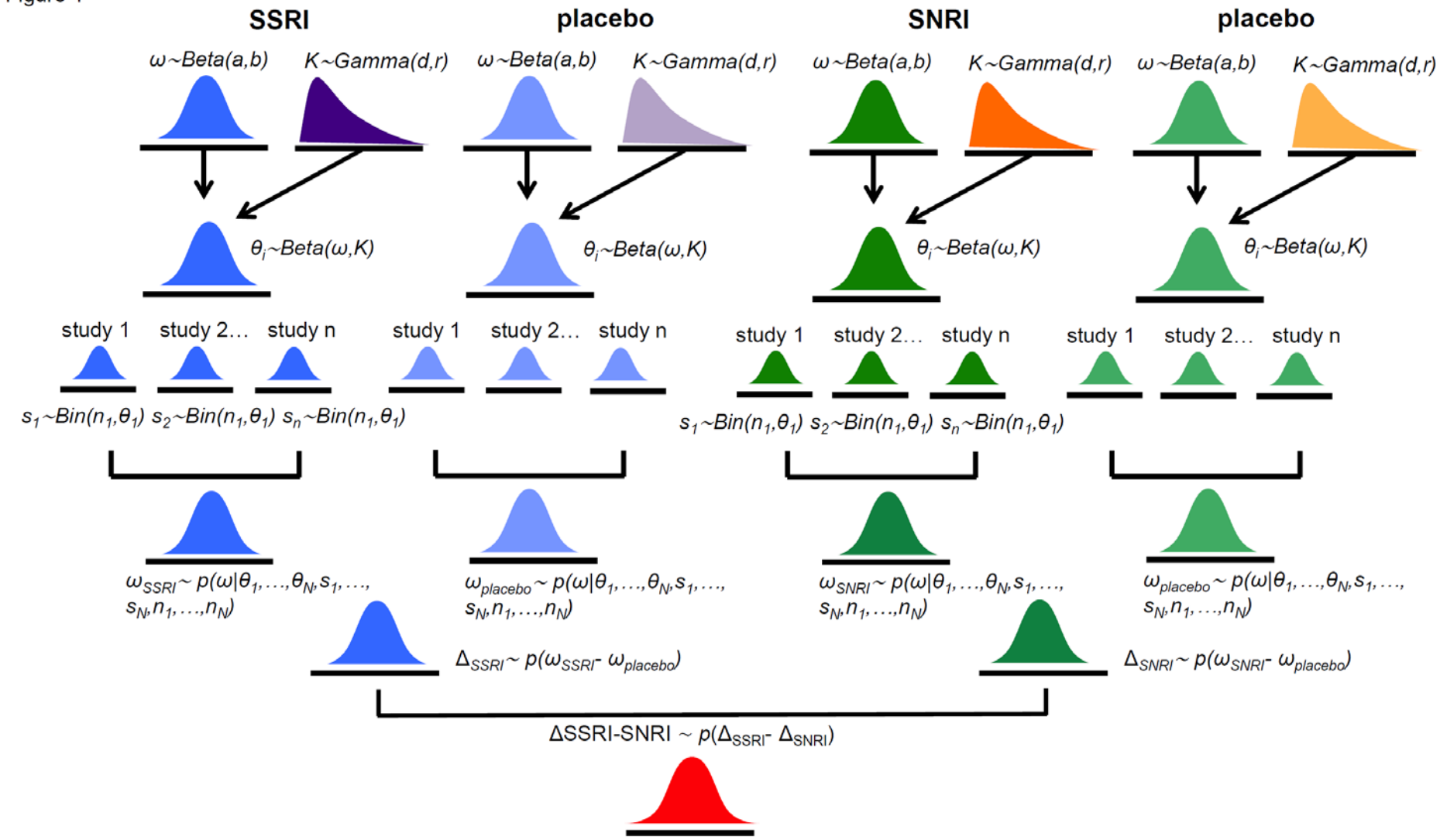
- Bayesian updating (w/o hierarchical priors) gives all data equal weighting, assuming the trials are homogeneous.

BHMs explicitly model across trial heterogeneity

- By specifying a hierarchy of prior distributions, we can allow for variations in trial outcomes due to unobserved differences in trial conduct:
- Different types of patients, different investigators, different locations and number of locations used, different treatments, dosing, trial length, etc.
- The BHM allows estimation of the average outcomes from each individual trial *and* the average outcome across all trials, and also estimates the degree of heterogeneity.

A Bayesian Hierarchical Model (BHM)

Figure 1



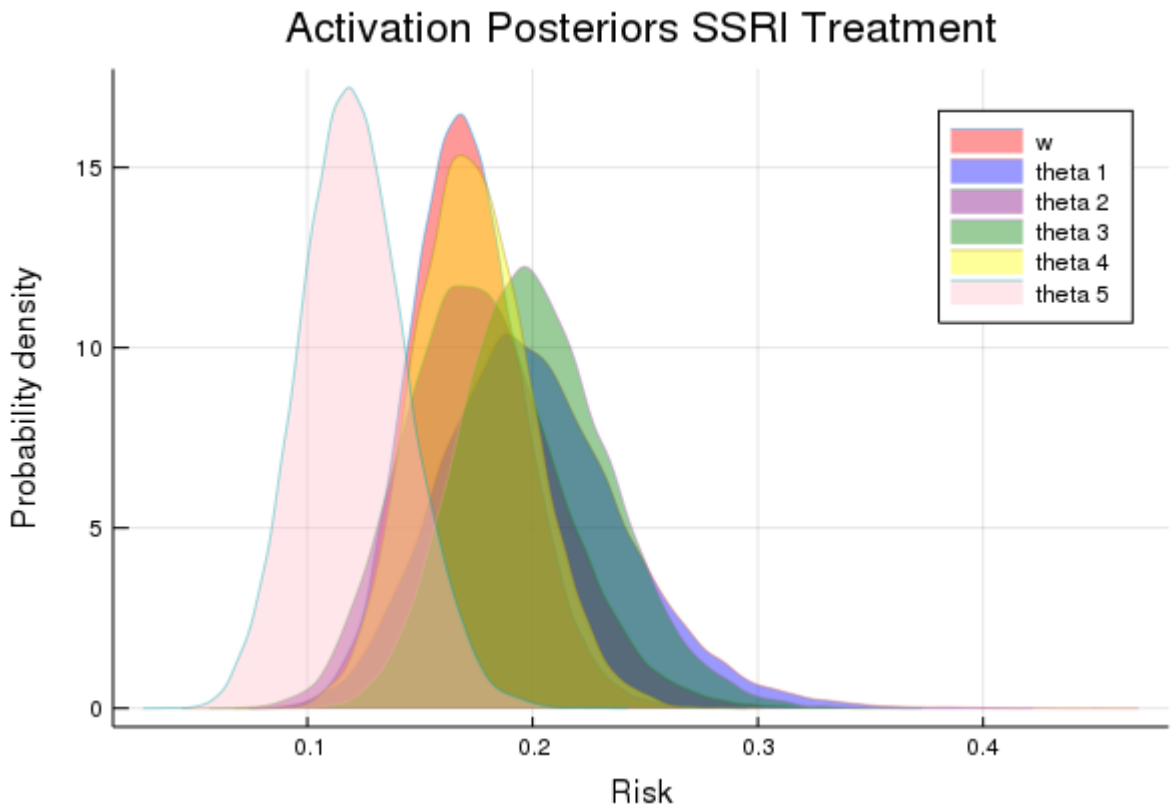
Turing.jl: Two specifications for models with a binomial (or Bernoulli) likelihood

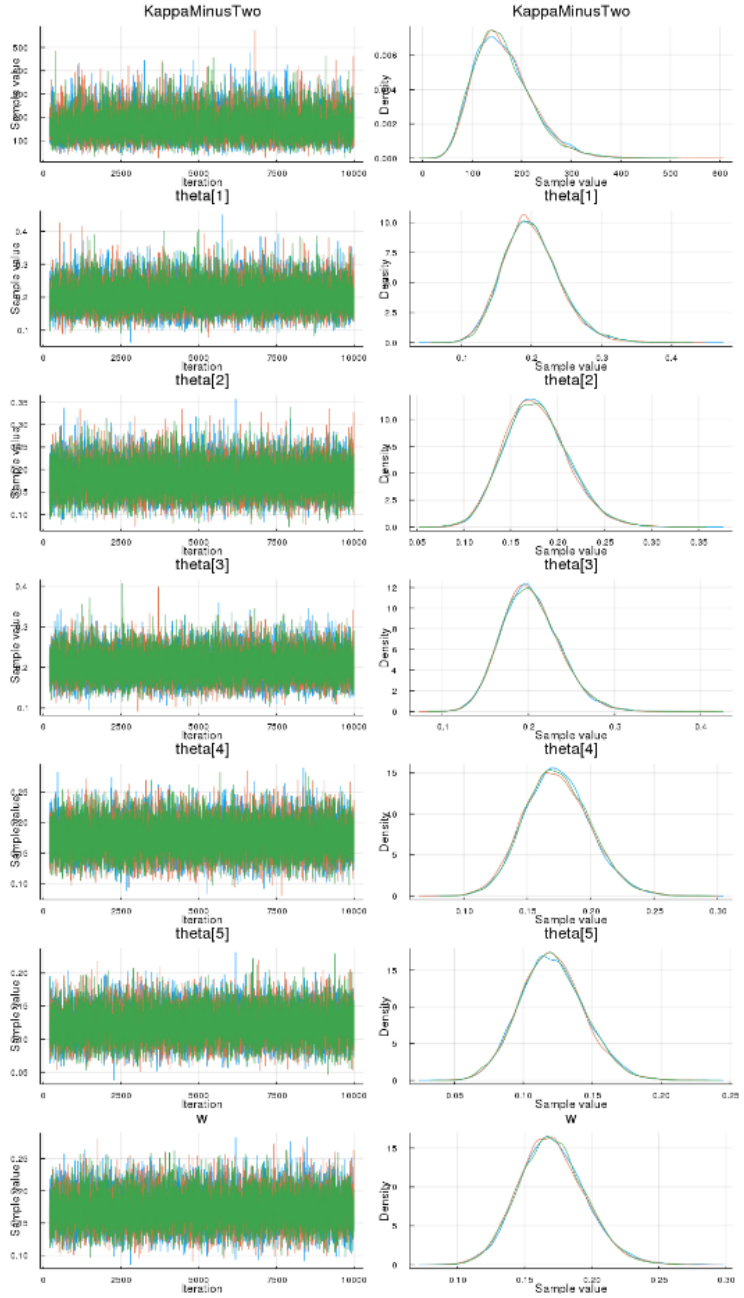
```
@model logit_bin_model(s1, n1, s2, n2) = begin
    # hierarchical prior
    m ~ Normal(0, 1)
    s ~ Exponential(1)
    a1 ~ Normal(m, s)
    a2 ~ Normal(m, s)

    # prior for theta
    theta1 = logistic(a1)
    theta2 = logistic(a2)

    # likelihood
    s1 ~ Binomial(n1,theta1)
    s2 ~ Binomial(n2,theta2)
end;
```

```
@model binomial_trials(s,n) = begin
    g = length(n)  # number of groups

    # hierarchical prior
    w ~ Beta(2,3)
    K ~ Gamma(10,1/0.05)
    a = w*K + 1.0
    b = (1.0 - w)*K+1.0

    # priors for each success rate
    theta = Array{Real}(undef, g)
    for k in 1:g
        theta[k] ~ Beta(a,b)
    end

    # likelihood
    for i in 1:g
        s[i] ~ Binomial(n[i],theta[i])
    end
end;
```

Turing.jl HMC estimation

- Two chains - HMC No U-Turn sampler
- BHM meta-analysis combining evidence from 5 RCTs
- Evaluating side effect incidence from SSRIs and SNRIs
- $\boldsymbol{w}$ $=$ the probability of occurrence side effect





[NUTS] time = 33.1 seconds; 3 chains; samples per chain = 10,000
Parameters: w, theta[1:5], KappaMinusTwo (K-2)

Empirical Posterior Estimates

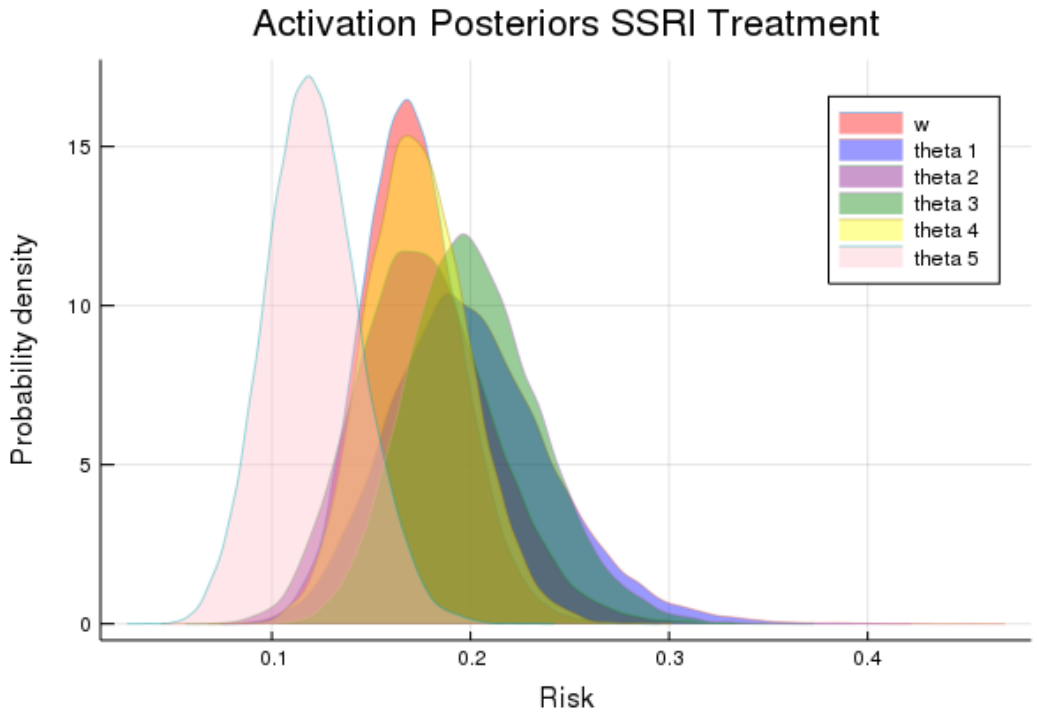
parameters

	Mean	SD	Naive SE	MCSE	ESS
KappaMinusTwo	162.2216	58.2411	0.3363	0.6367	8368.686
theta[1]	0.2011	0.0435	0.0003	0.0004	10000.000
theta[2]	0.1777	0.0343	0.0002	0.0003	10000.000
theta[3]	0.2027	0.0338	0.0002	0.0003	10000.000
theta[4]	0.1737	0.0259	0.0001	0.0002	10000.000
theta[5]	0.1209	0.0235	0.0001	0.0002	10000.000
w	0.1706	0.0270	0.0002	0.0003	10000.000

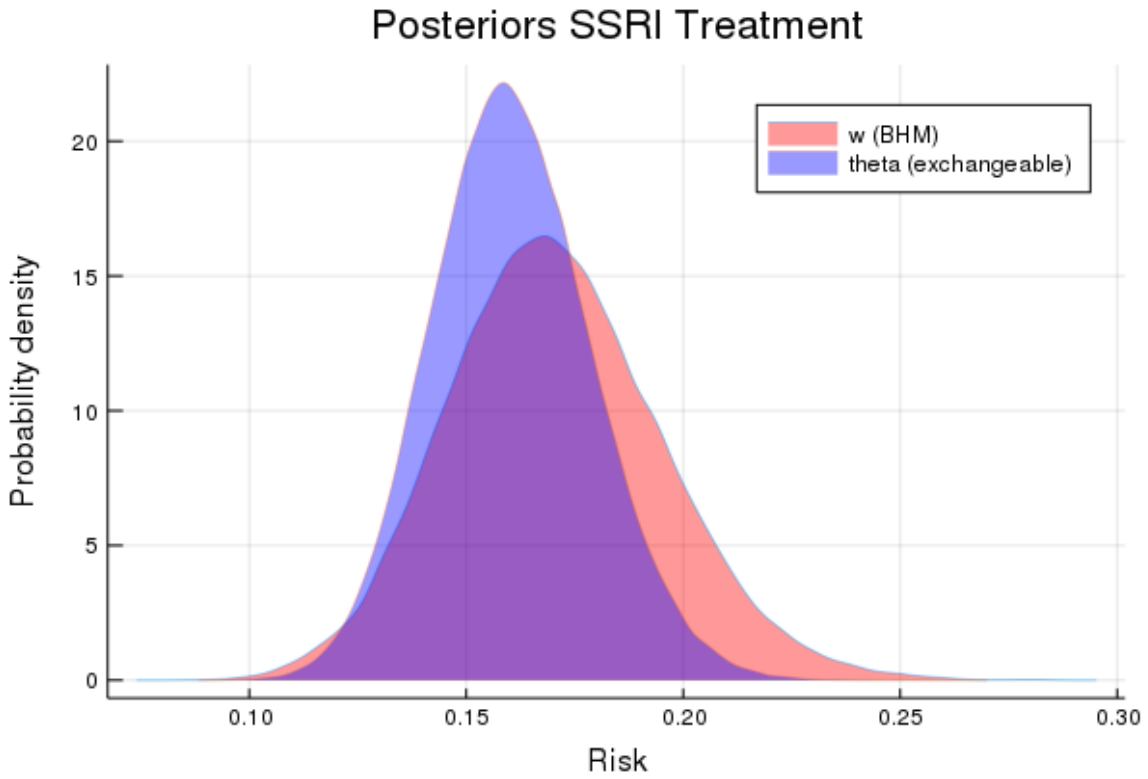
Quantiles

	2.5%	50.0%	97.5%
KappaMinusTwo	0.6550	155.0876	574.4035
theta[1]	0.0625	0.1976	0.9160
theta[2]	0.0719	0.1756	0.6648
theta[3]	0.0915	0.2003	0.4474
theta[4]	0.0158	0.1728	0.2900
theta[5]	0.0004	0.1201	0.2619
w	0.0853	0.1693	0.8509

Meta-analysis for Activation Side Effect (AE)



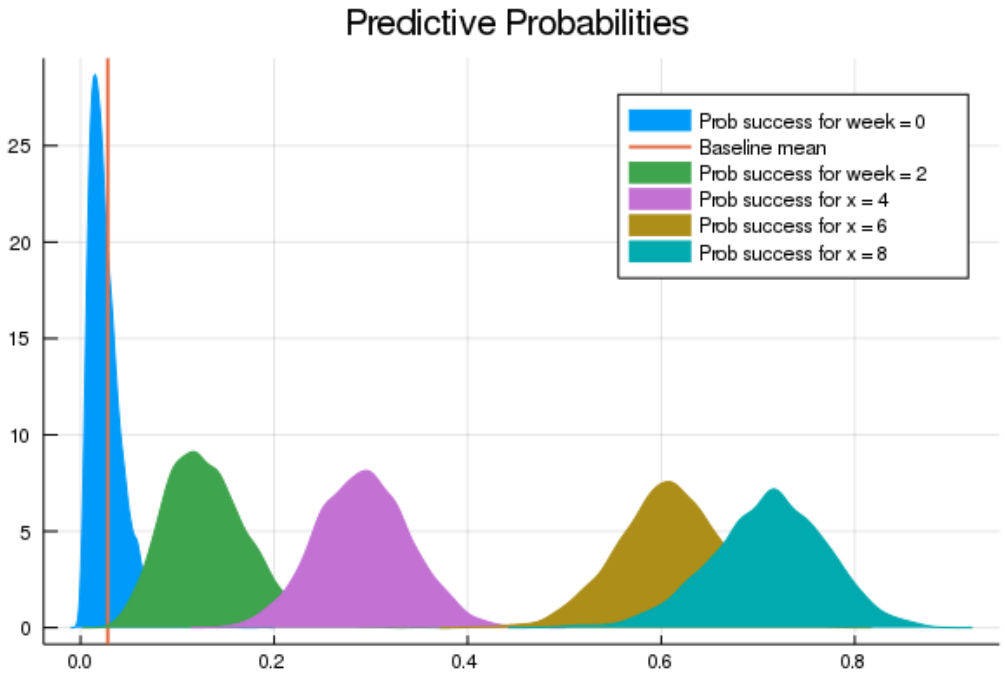
BHM allows for heterogeneity across studies

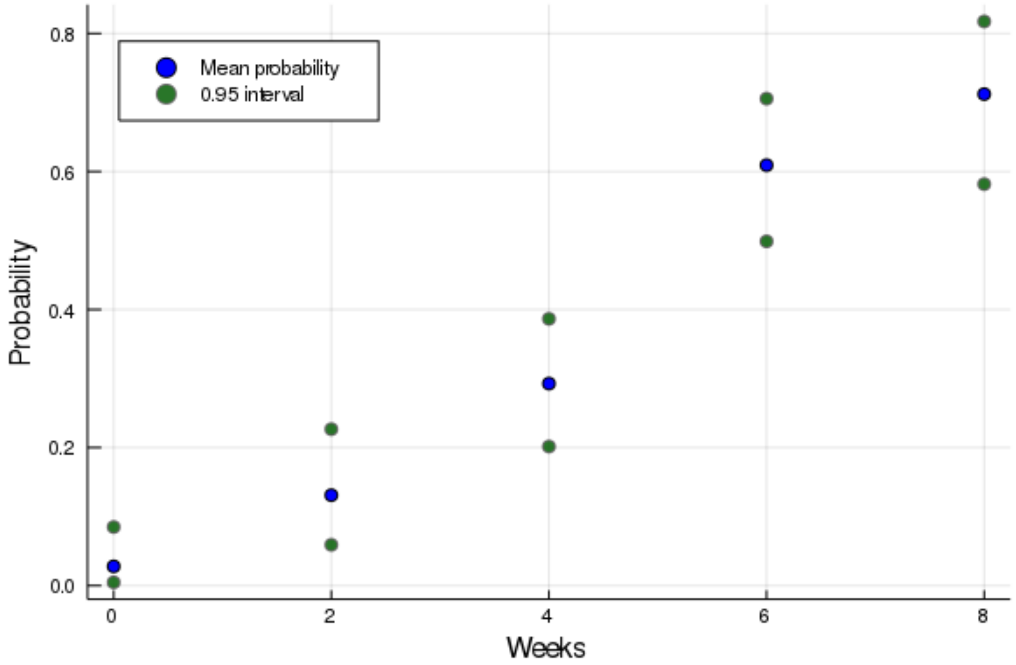


Bayesian updating assumes homogeneity across studies

The trajectory of probability of CGI ≤ 2 over weeks 0, 2, 4, 6, 8 controlling for:
Demographic (age, sex)
Physiological (weight, vital signs)
Genetic (*SLC6A4* , CYP2C19, *HTR2A*)

```
@model logistic_multi_regression(x, y) = begin
    a ~ Normal(0, 10)
    n, k = size(x)
    b = Array{Real}(undef, k)
    for i in 1:k
        b[i] ~ Normal(0, 10)
    end
    mu = a .+ x*b
    for i = 1:n
        v = logistic(mu[i])
        y[i] ~ Bernoulli(v)
    end
end  > logistic_multi_regression
```





Key Advantage of Julia:

Conservation of time and energy of the applied researcher (anyone who is not a computer scientist). For a computer scientist or programmer, the code is often the ultimate goal.

For an applied researcher, the code is just a means to an end.

Examples using Julia from our research

- Reevaluating the evidence from previously conducted RCTs using Bayesian updating.
- Analysis of abandoned trials – Bayesian posteriors from summary statistics.
- Joint evaluation of tolerability and efficacy in RCTs – examine the **joint** posterior distribution.
- Bayesian hierarchical modeling for meta-analyses, e.g. evaluating adverse events (“side effects”) in trial participants.

Thanks!

Code to generate the figures and results is available in the github repositories:

https://github.com/tszanalytics/Juliacon2019

https://github.com/tszanalytics/BayesTesting.jl