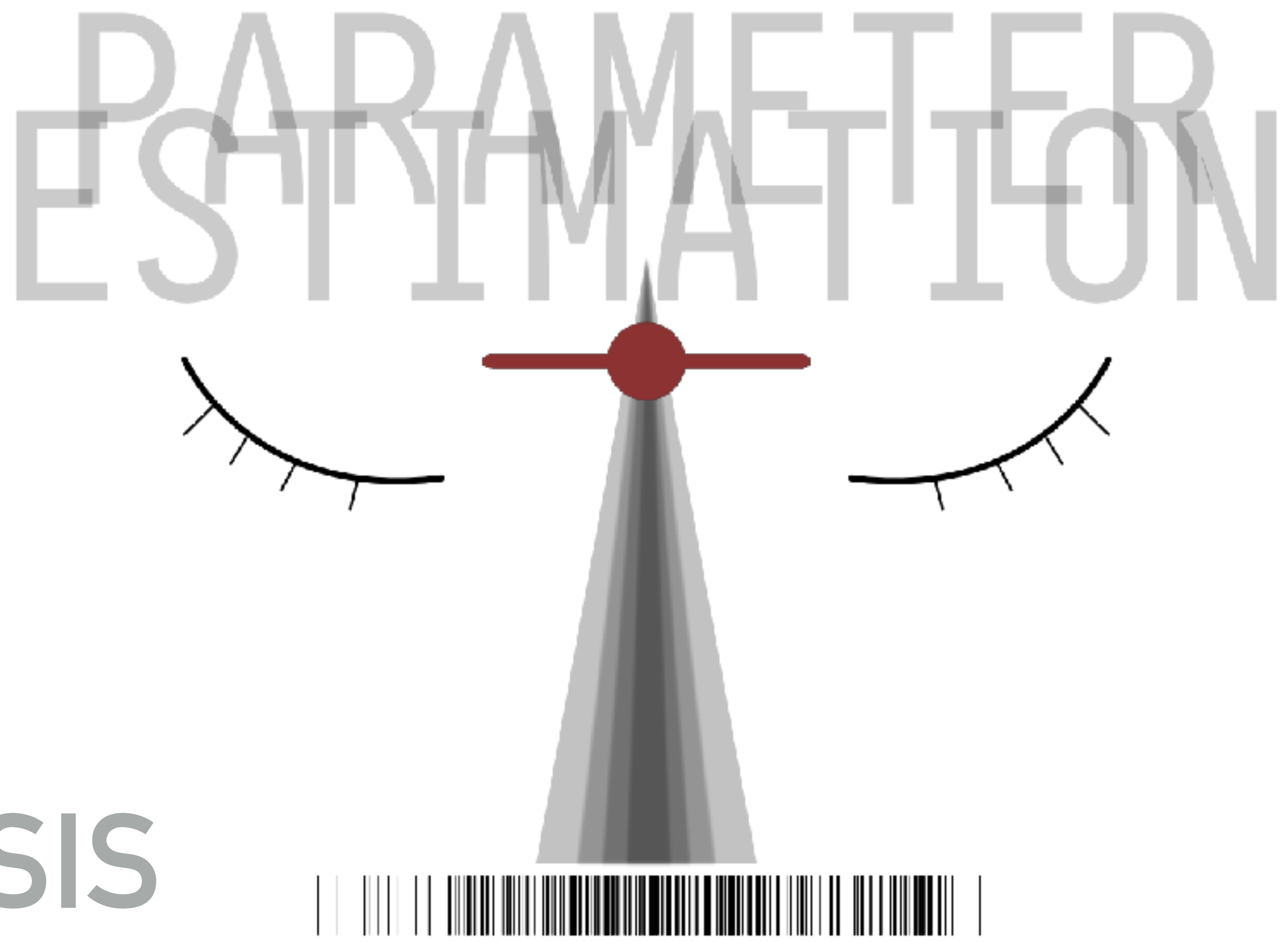
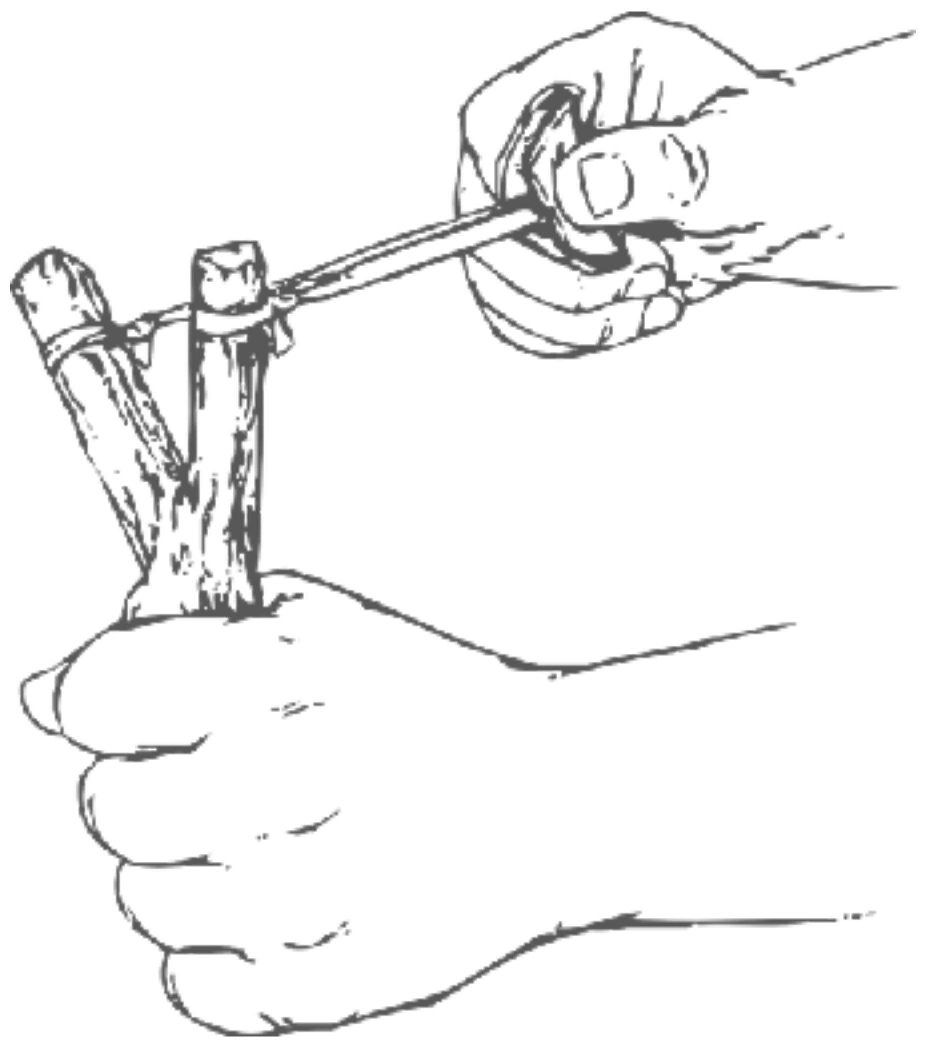


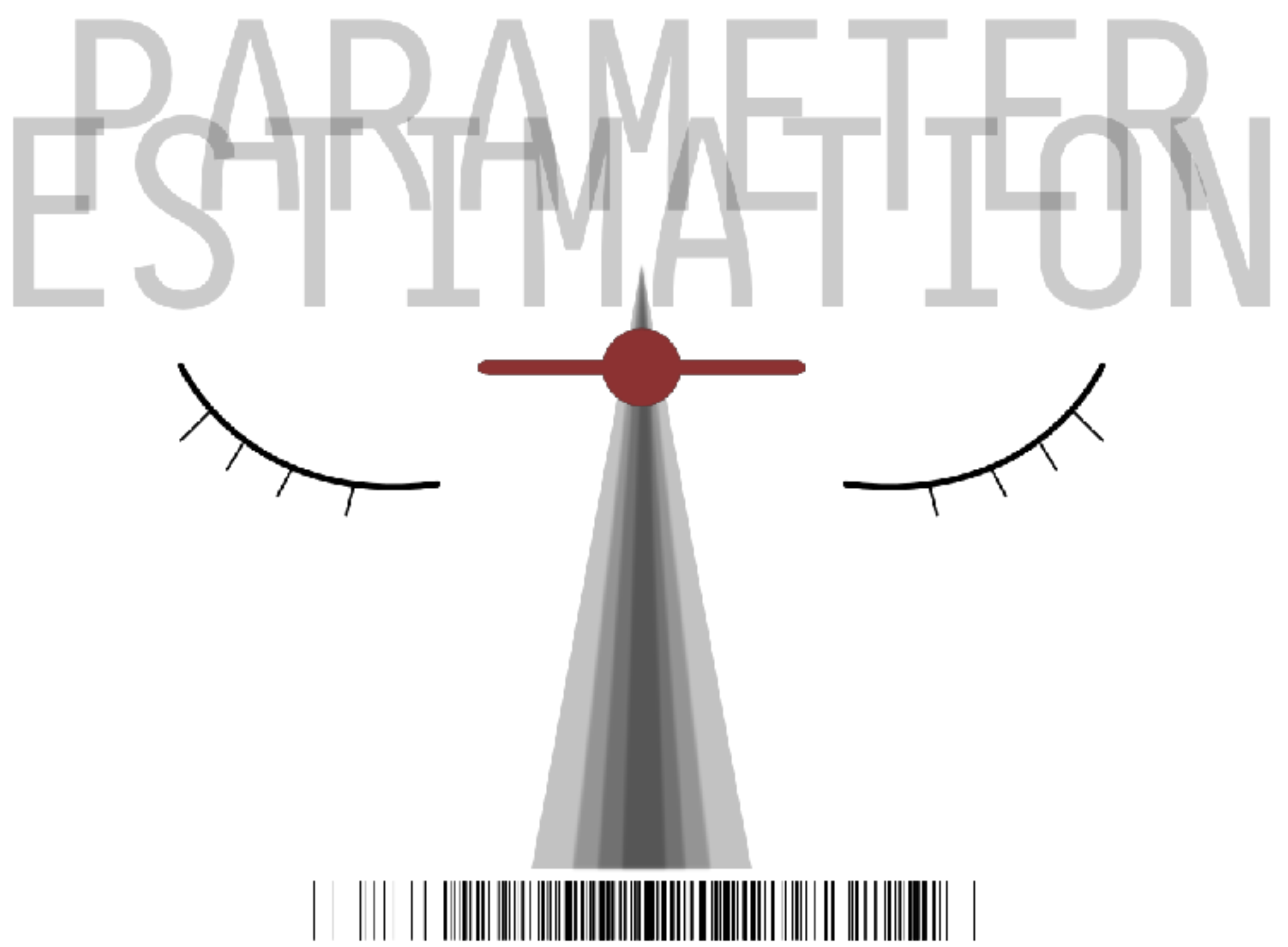
INTRODUCTION TO DATA ANALYSIS

PARAMETER ESTIMATION

LEARNING GOALS



- understand Bayes rule for parameter estimation
  - (conjugate) priors, likelihood
- point-valued & interval-based estimators
  - frequentist: MLE, confidence intervals
  - Bayes: mean of posterior, credible intervals
- implement probabilistic models in `greta`
- compute with posterior samples



ESTIMATES

- point-valued: single “best” values
- interval-range: “good” values (around “best” value)

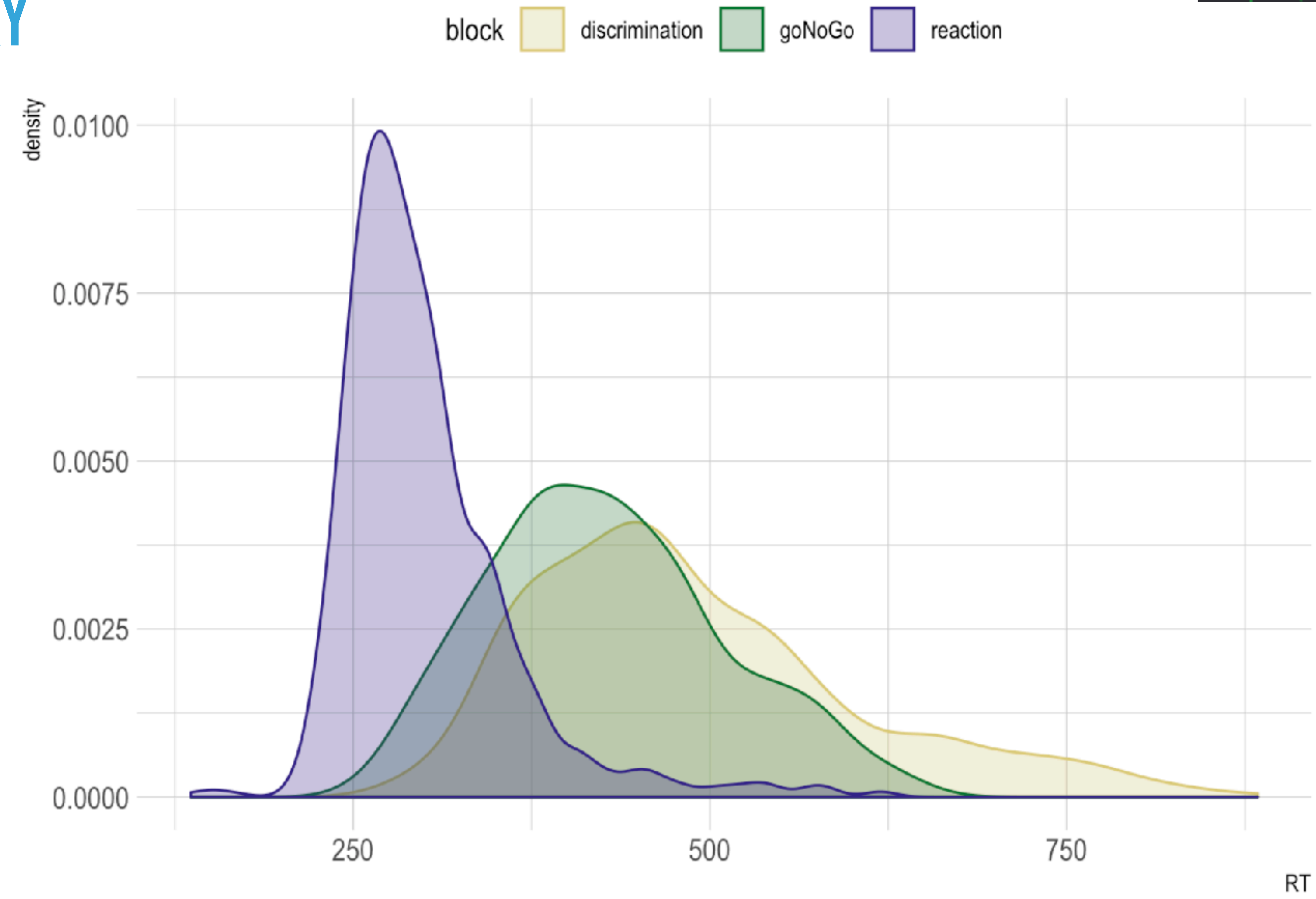
estimate	Bayesian	frequentist
best value	mean of posterior posterior	maximum likelihood estimate
interval range	credible interval (HDI)	confidence interval

model-based hypothesis testing

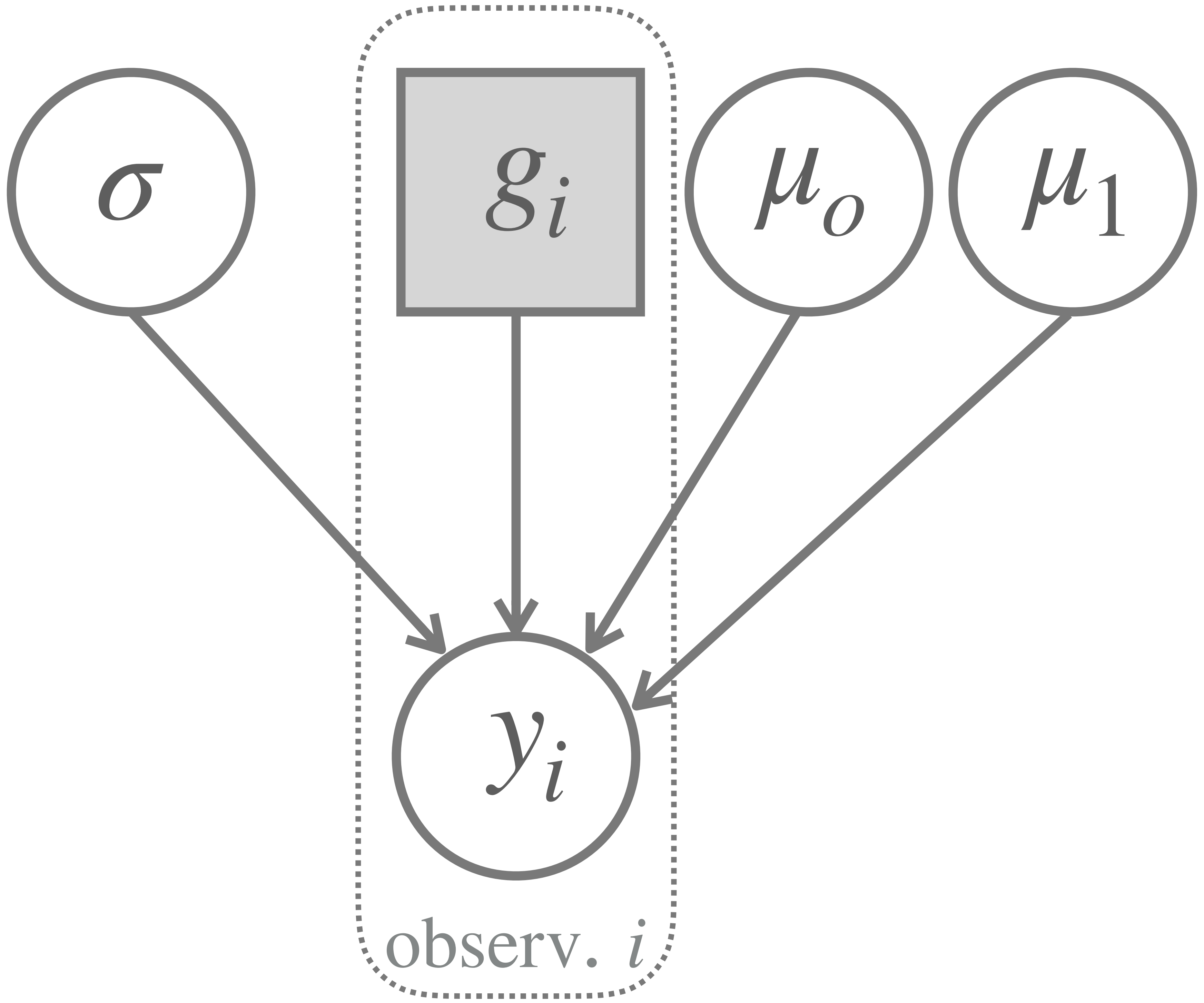
MENTAL CHRONOMETRY

- N=50 participants recruited via Prolific
- three blocks / conditions
  - reaction press button when a shape appears
  - go/no-go press button for shape 1; don't press for shape 2
  - discrimination press one button for shape 1, another for shape 2

MENTAL CHRONOMETRY

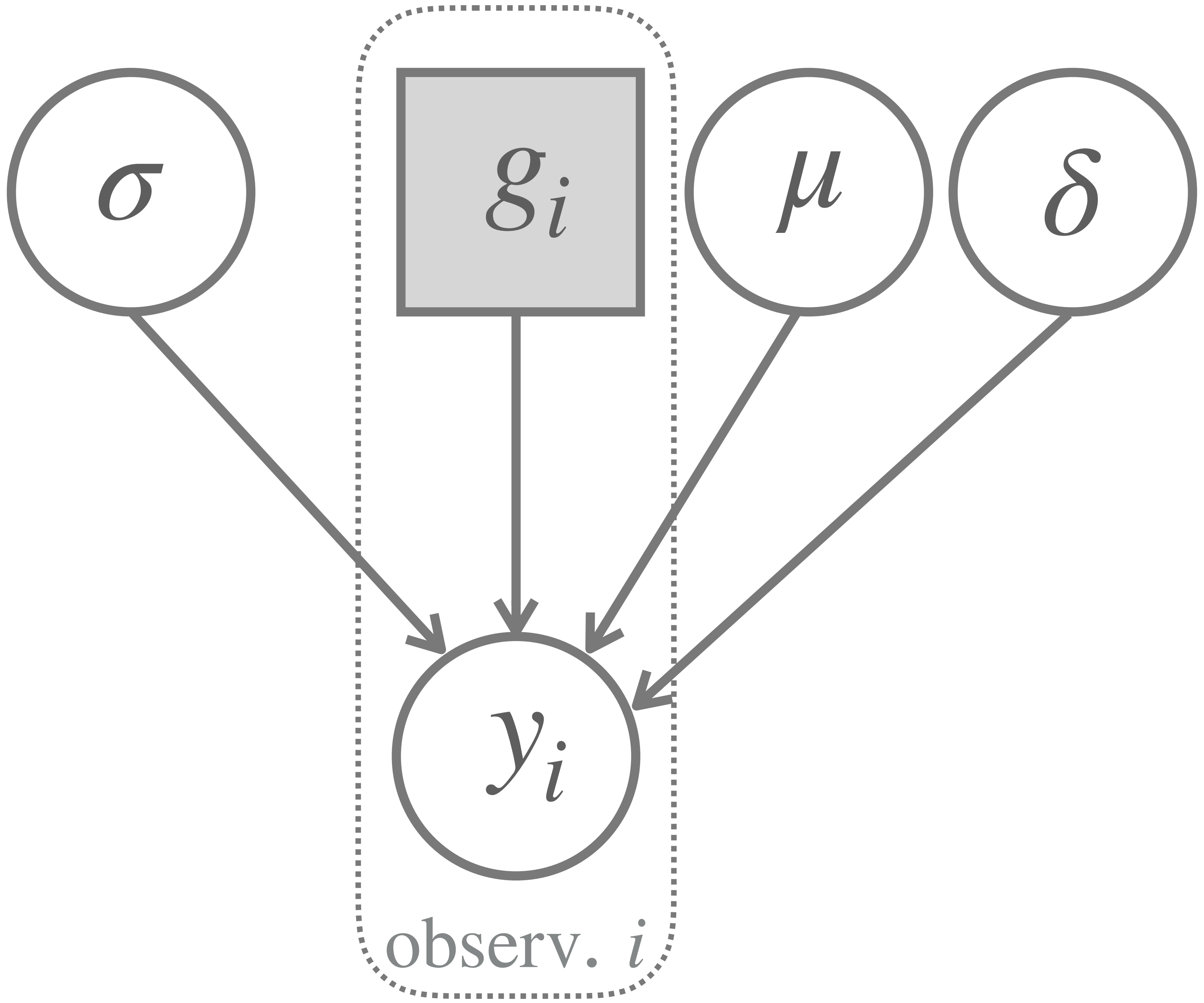


T-TEST MODEL [TWO UNCOUPLED MEANS]



$$\sigma \sim \text{Trunc-Norm}(\ldots, \text{lower} = 0)$$
$$\mu_0 \sim \text{Normal}(\ldots)$$
$$\mu_1 \sim \text{Normal}(\ldots)$$
$$y_i \sim \text{Normal}(\mu_{g_i}, \sigma)$$

T-TEST MODEL [WITH DIFFERENCE BETWEEN MEANS]



$$\sigma \sim \text{Trunc-Norm}(\ldots, \text{lower} = 0)$$

$$\mu \sim \text{Normal}(\ldots)$$

$$\delta \sim \text{Normal}(0, \ldots)$$

$$y_i \sim \begin{cases} \text{Normal}(\mu, \sigma) & \text{if } g_i = 0 \\ \text{Normal}(\mu + \delta, \sigma) & \text{if } g_i = 1 \end{cases}$$

HYPOTHESES & PARAMETER VALUES

- point-valued null hypothesis: $\delta = 0$
- observe data D
- three ways of testing [recall three pillars of DA]:
  - estimation: is 0 among the parameters estimated from D?
  - prediction: is D among the data predicted by a model with $\delta = 0$?
  - comparison: take two models: one with $\delta = 0$, one where δ takes on different values, too; which one explains D better?

Bayes rule for parameter estimation

BAYES RULE FOR PARAMETER ESTIMATION

$$\underbrace{P(\theta \mid D)}_{\text{posterior}} = \frac{\underbrace{P(D \mid \theta)}_{\text{likelihood}} \; \underbrace{P(\theta)}_{\text{prior}}}{\underbrace{P(D)}_{\text{marginal likelihood}}}$$

$$\underbrace{P(D)}_{\text{marginal likelihood}} = \int P(D \mid \theta) \; P(\theta) \; \mathrm{d}\theta$$

REMARKS ON NOTATION

- if there is only one model M, we leave out the model index, writing $P(\theta)$ instead of $P_M(\theta)$
- we write $P(\theta \mid D)$ instead of $P(\Theta = \theta \mid \mathscr{D} = D)$
- short-hand with non-normalized probabilities (implicit normalizing constant):

$$\underbrace{P(\theta \mid D)}_{posterior} \propto \underbrace{P(\theta)}_{prior} \; \underbrace{P(D \mid \theta)}_{likelihood}$$

EXAMPLE

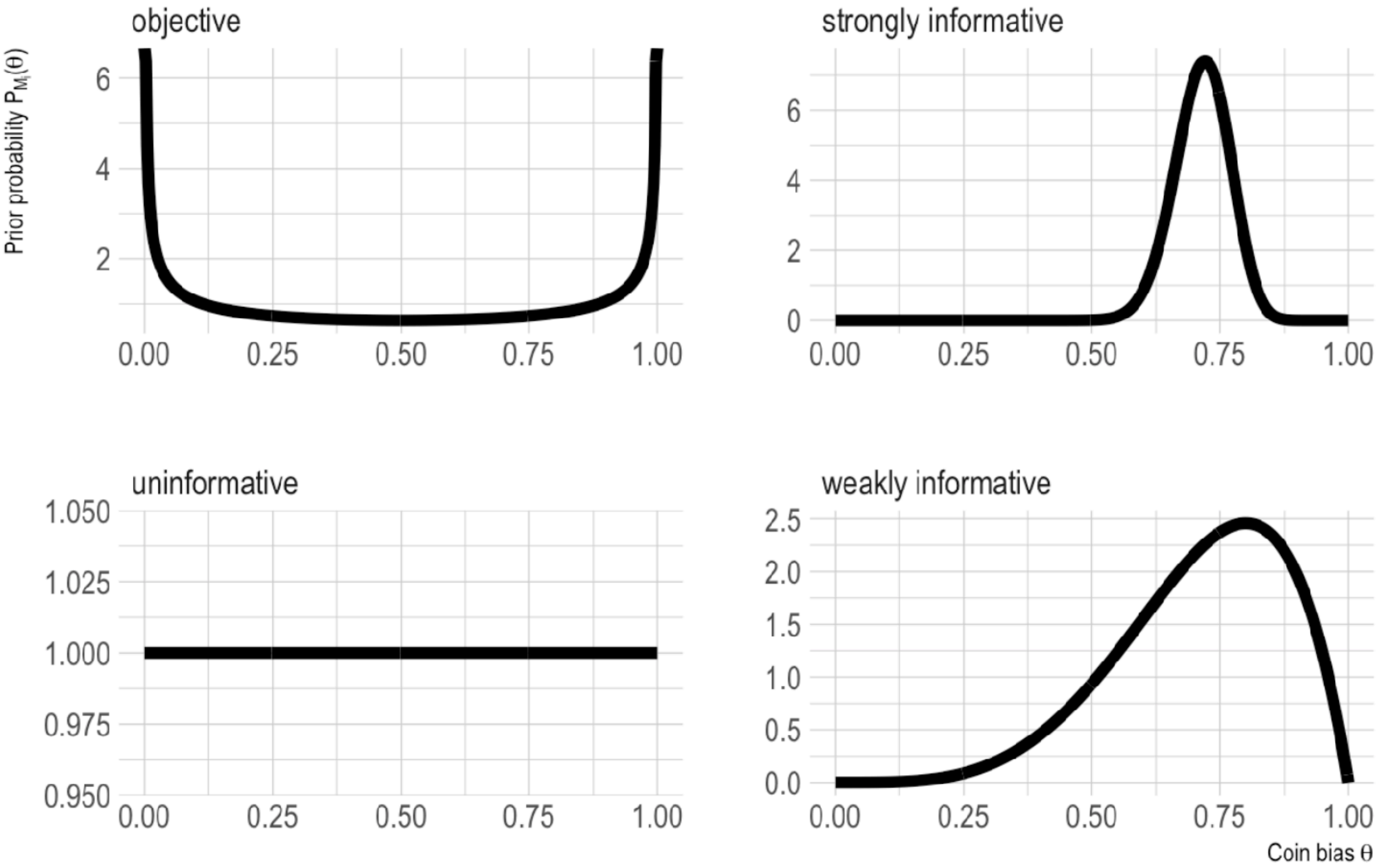
- model:

$$k \sim \text{Binomial}(N, \theta)$$
$$\theta \sim \text{Beta}(\alpha, \beta)$$

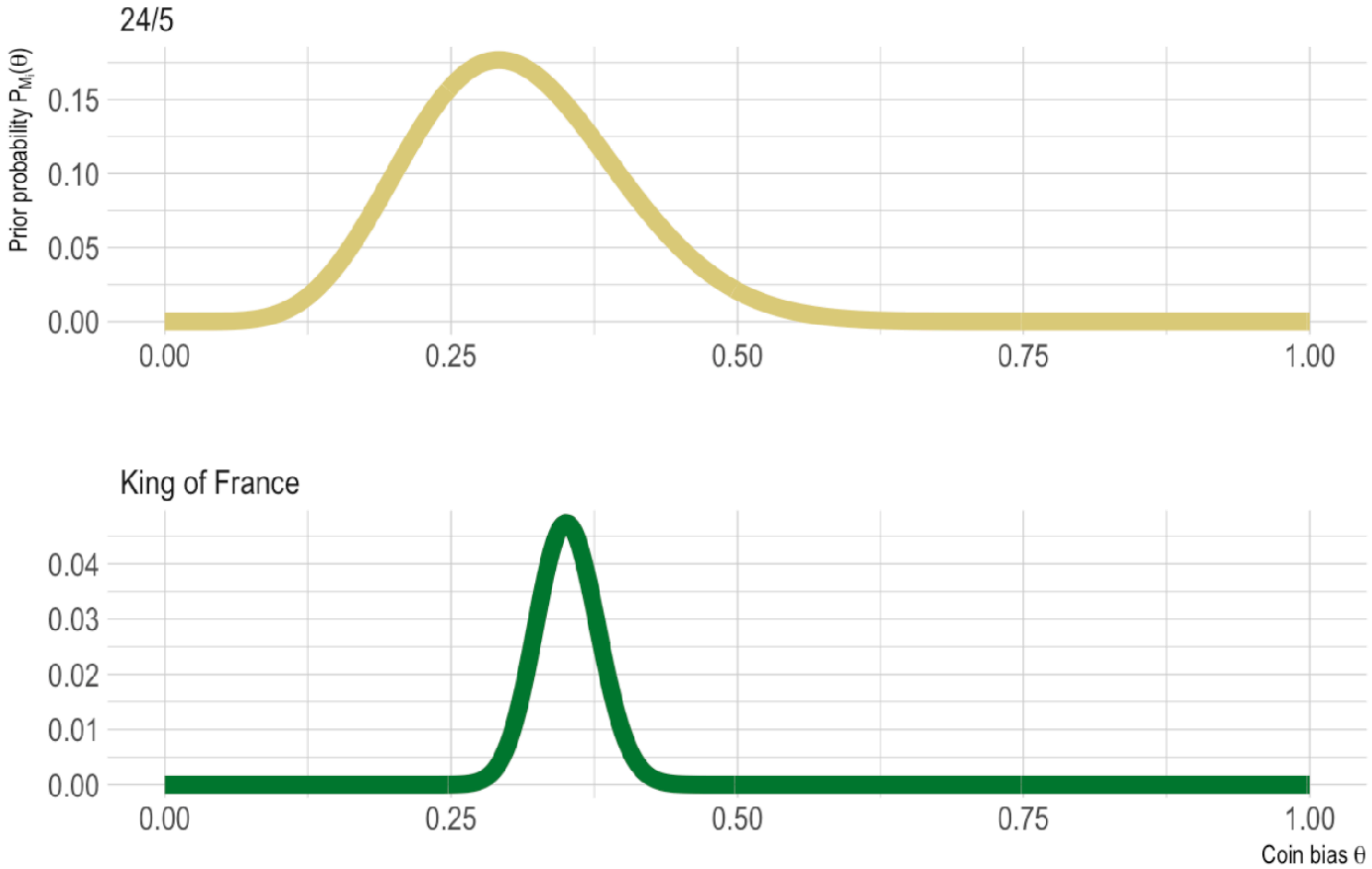
- data:
  - "24/7" $k = 7$ $N = 24$
  - "KoF" $k = 109$ $N = 311$

  [number of "true" responses to all sentences with a false presupposition]

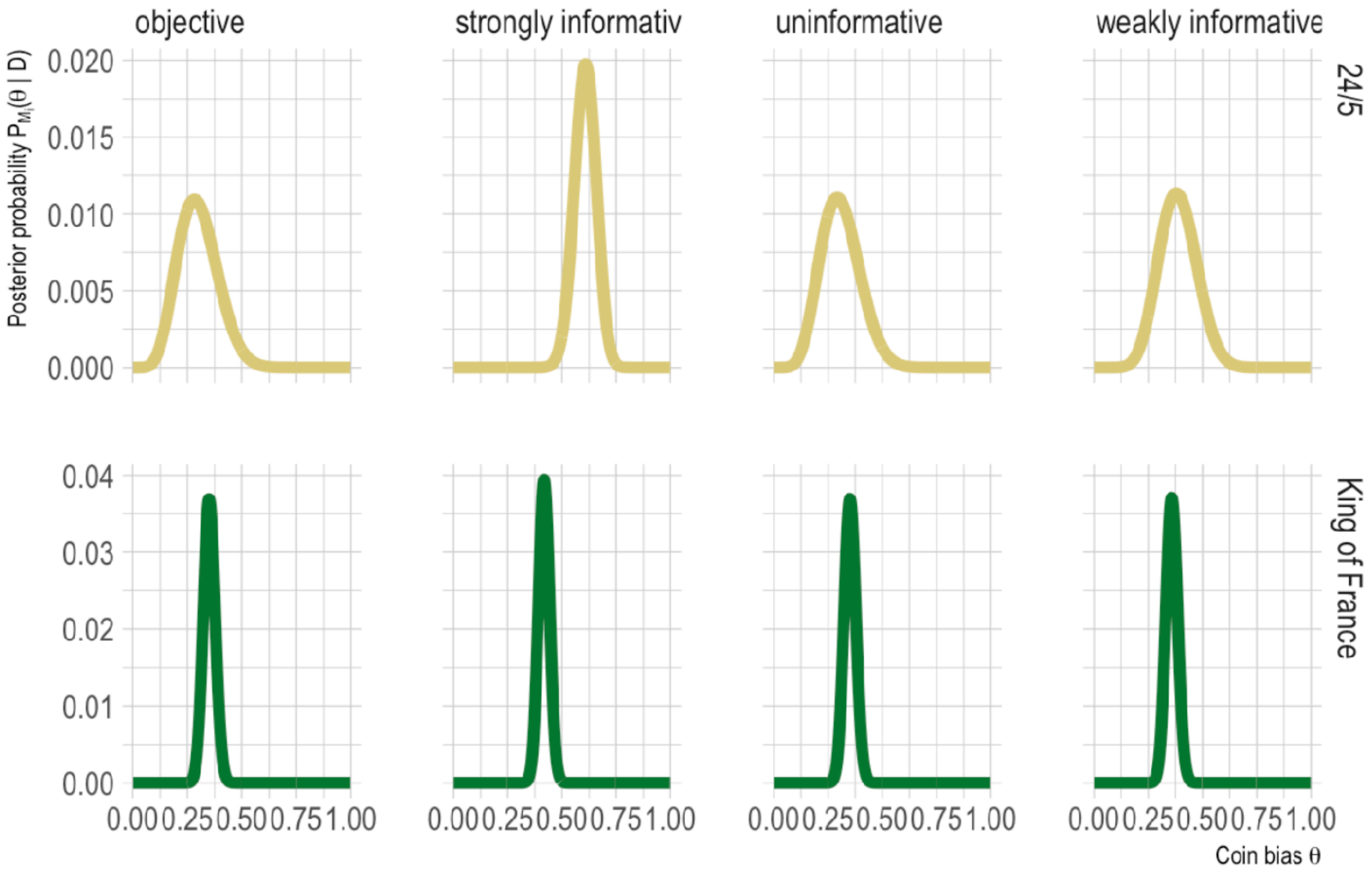
PRIOR


objective
strongly informative
uninformative
weakly informative
Prior probability $P_{M_i}(\theta)$
6
4
2
0
0.00
0.25
0.50
0.75
1.00
1.050
1.025
1.000
0.975
0.950
2.5
2.0
1.5
1.0
0.5
0.0
Coin bias θ

LIKELIHOOD


24/5
Prior probability $P_{M_i}(\theta)$
0.15
0.10
0.05
0.00
0.00
0.25
0.50
0.75
1.00
King of France
0.04
0.03
0.02
0.01
0.00
0.00
0.25
0.50
0.75
1.00
Coin bias θ

POSTERIOR

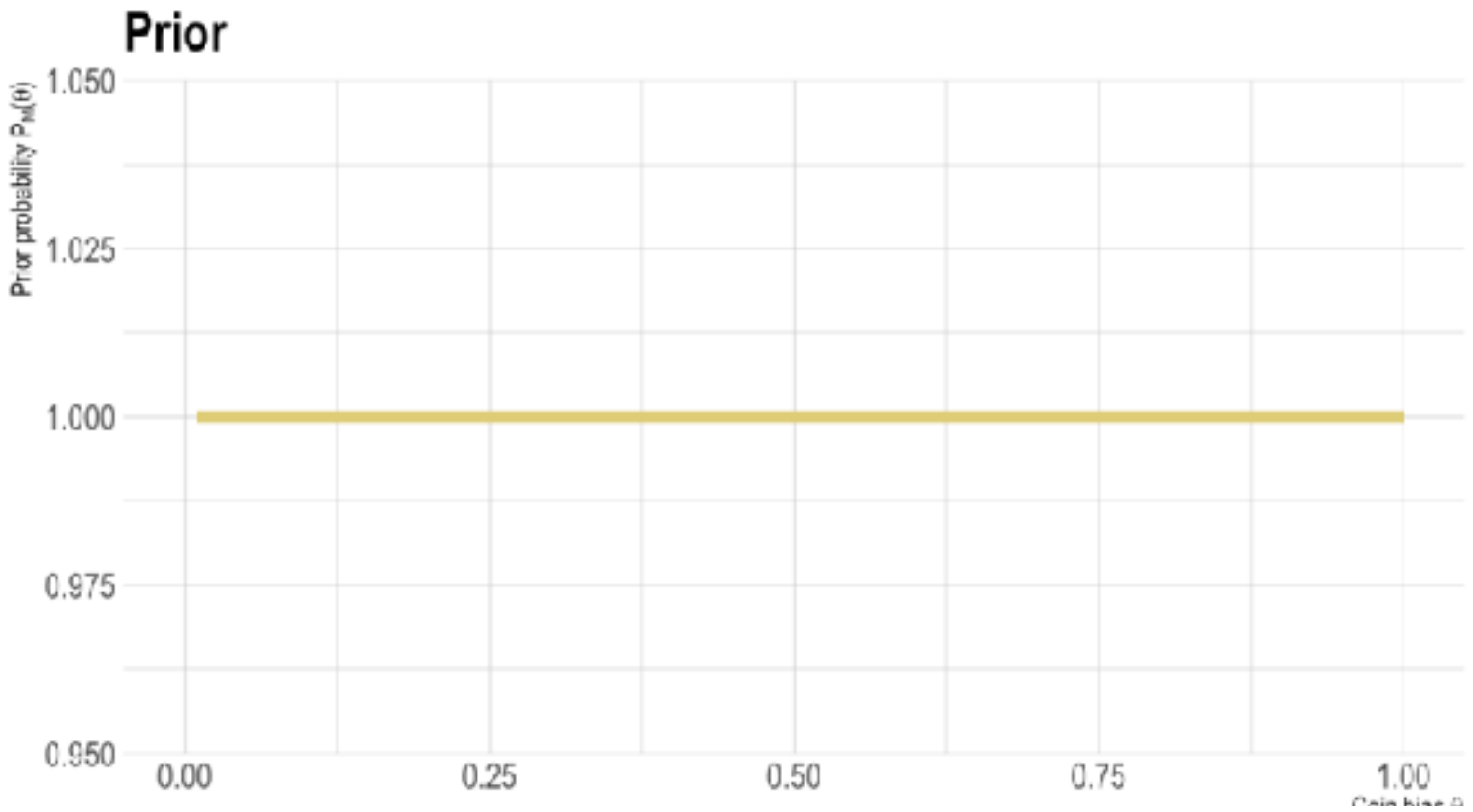


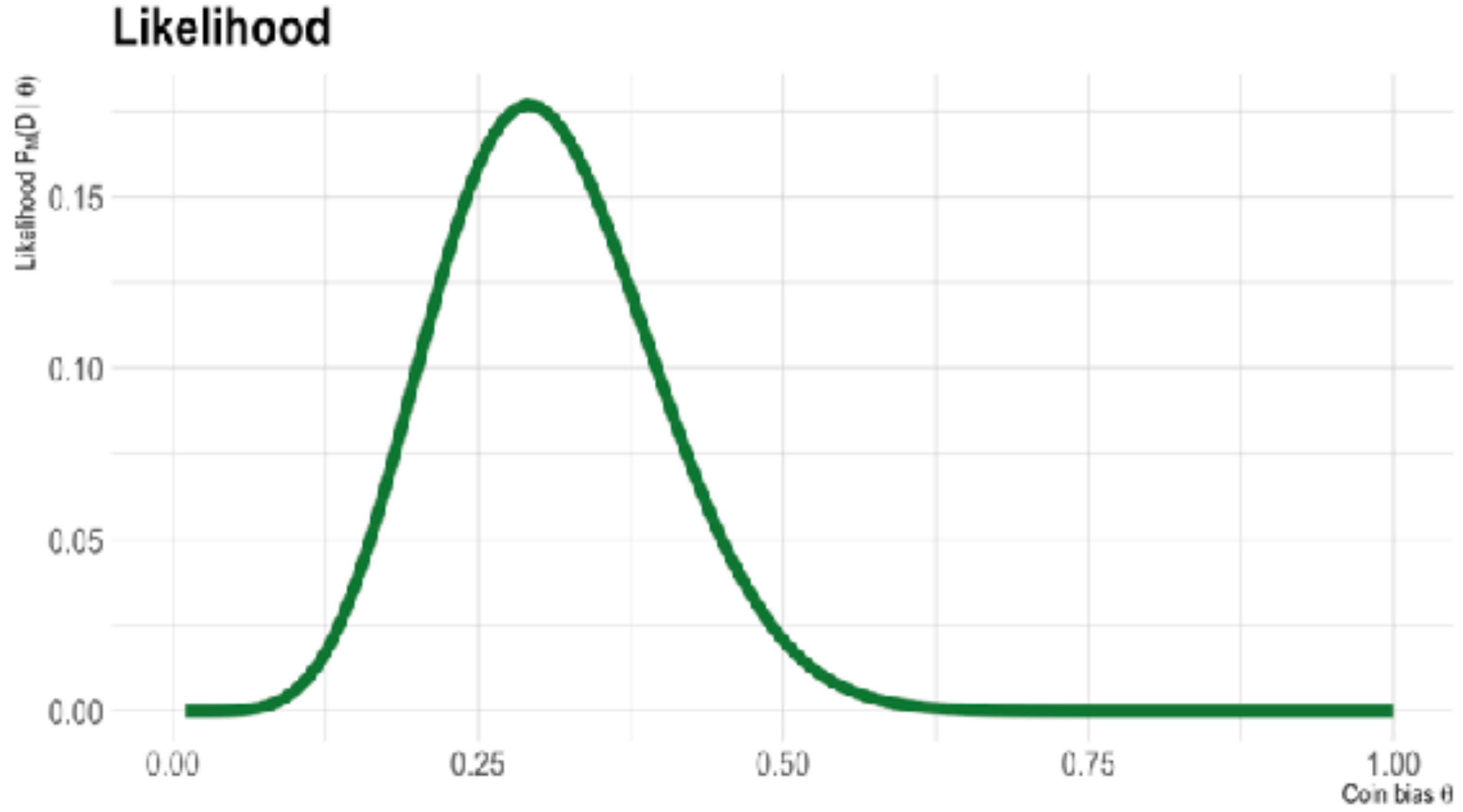
objective
strongly informativ
uninformative
weakly informative
24/5
King of France
Posterior probability $P_{M_i}(\theta \mid D)$
0.020
0.015
0.010
0.005
0.000
0.04
0.03
0.02
0.01
0.00
0.00 0.25 0.50 0.75 1.00
Coin bias θ

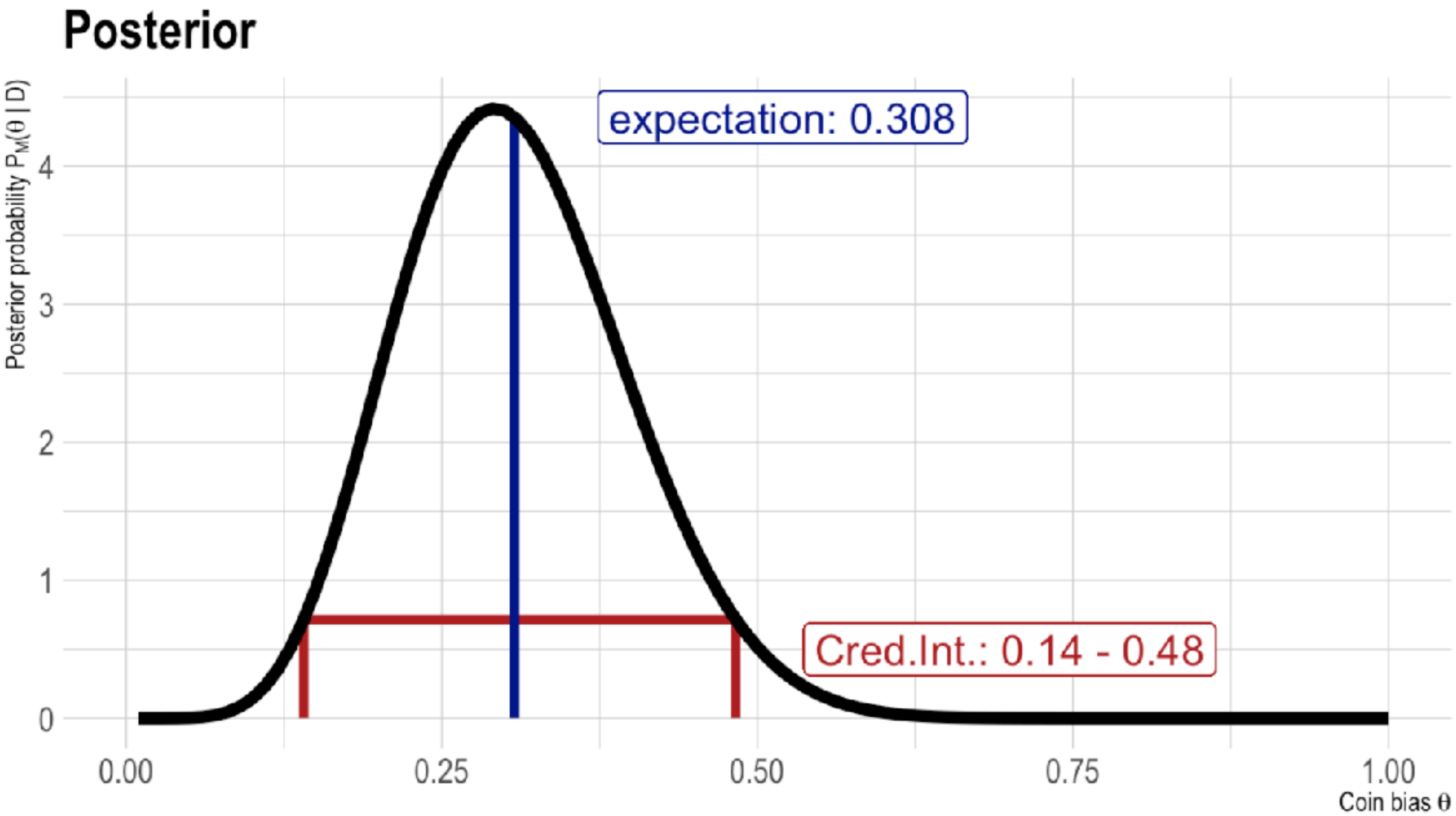
Bayesian point- & interval-estimates

EXAMPLE

- model: $k \sim \text{Binomial}(N, \theta), \theta \sim \text{Beta}(1,1)$
- data: $k = 7,\ N = 24$







POSTERIOR MEAN & MAP

- posterior mean:

$$\mathbb{E}_{P(\theta|D)} = \int \theta \; P(\theta \mid D) \; \mathrm{d}\theta$$

- maximum a posteriori:

$$\text{MAP}(P(\theta \mid D)) = \arg\max_{\theta} P(\theta \mid D)$$

- posterior mean is proper Bayesian measure, because it is holistic = influenced by whole distribution
- MAP is local, not influenced by whole distribution
- estimation of posterior mean is (usually) less error-prone than estimation of MAP

CREDIBLE INTERVAL

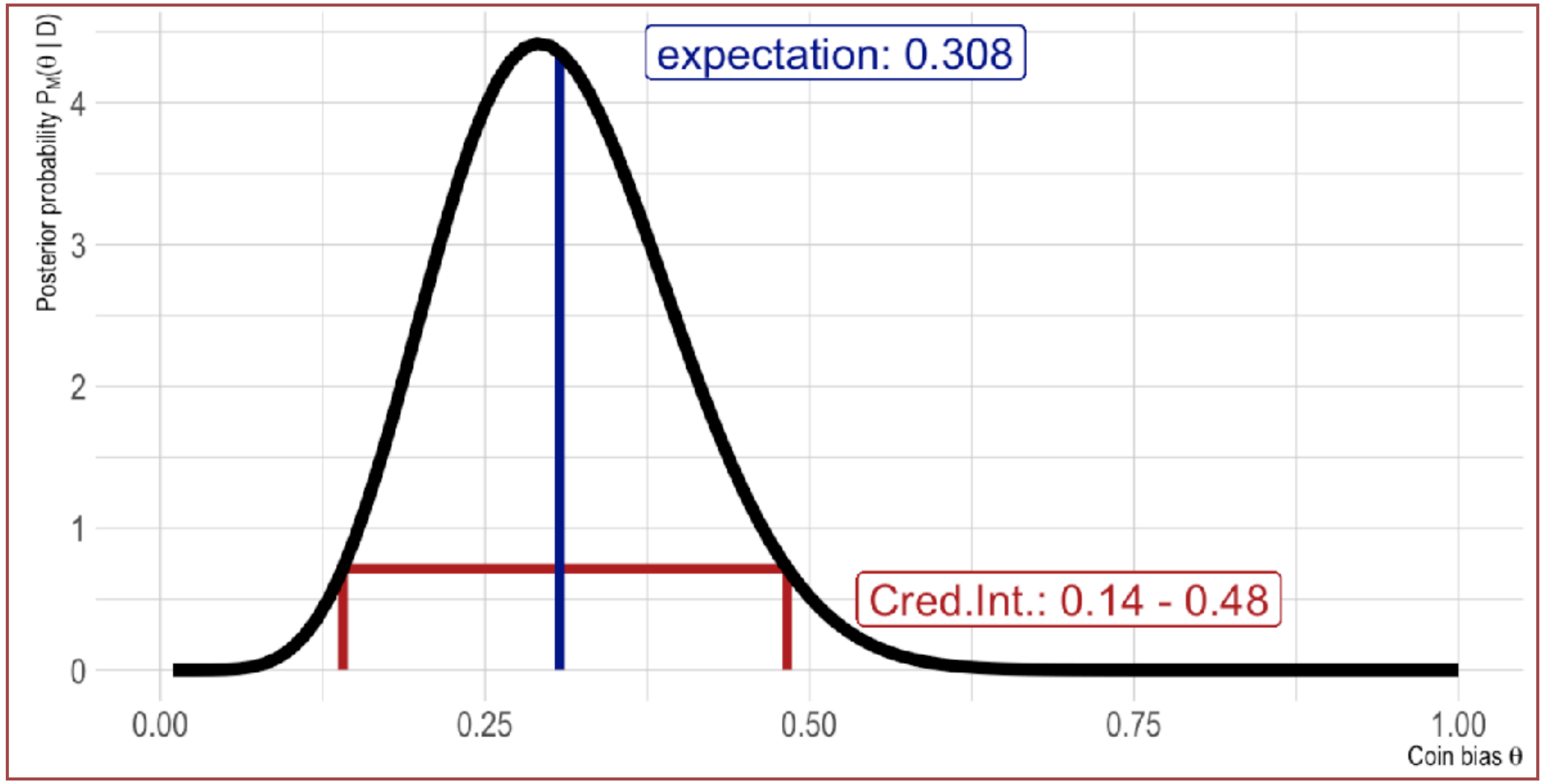
- interval $[l;u]$ is a $\gamma\%$ credible interval for a random variable X if

  (I) $P(l \leq X \leq u) = \frac{\gamma}{100}$, and

  (II) for every $x \in [l;u]$ and $x' \notin [l;u]$ we have $P(X = x) > P(X = x')$

- "range of values too probable to properly ignore"

  [see David Lewis on "Elusive Knowledge"]



posteriors from conjugacy

BAYES RULE FOR PARAMETER ESTIMATION

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{\int P(D \mid \theta)\, P(\theta)\, \mathrm{d}\theta}$$

✓fast & easy

✓fast & easy

❌☠ possibly intractable ☠❌

CONJUGACY

- prior $P(\theta)$ is a conjugate prior for likelihood $P(D \mid \theta)$ iff prior $P(\theta)$ and posterior $P(\theta \mid D)$ are of the same kind of probability distribution (possibly with different parameter values)
- e.g., prior and posterior are both normal distributions, but have different means and standard deviations

CONJUGACY OF BETA & BINOMIAL

- claim: beta & binomial are conjugate
- proof:

$$P(\theta \mid k, N) \propto \text{Binomial}(k; N, \theta) \ \text{Beta}(\theta \mid a, b)$$

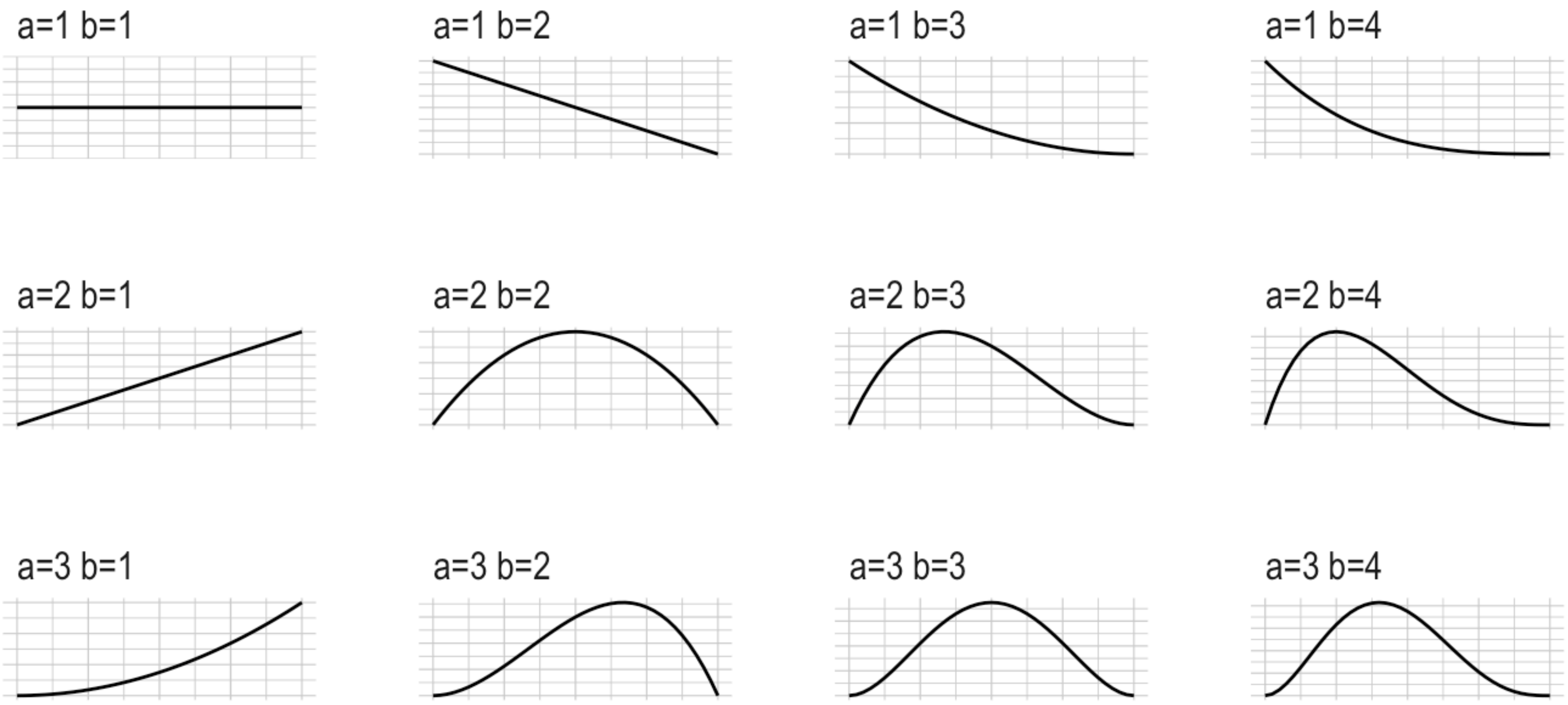
$$P(\theta \mid k, N) \propto \theta^{k} (1-\theta)^{N-k} \theta^{a-1} (1-\theta)^{b-1}$$

$$P(\theta \mid k, N) \propto \theta^{k+a-1} (1-\theta)^{N-k+b-1}$$

$$P(\theta \mid k, N) = \text{Beta}(\theta \mid k+a, N-k+b)$$

sequential updating

SEQUENTIAL UPDATING IN THE BETA-BINOMIAL MODEL



SEQUENTIAL UPDATING IN GENERAL

- claim: if D_1 and D_2 are disjoint and $D_1 \cup D_2 = D$, $P(\theta \mid D) \propto P(\theta \mid D_1)\ P(D_2 \mid \theta)$
- proof: $P(\theta \mid D) = \dfrac{P(\theta)\ P(D \mid \theta)}{\int P(\theta')\ P(D \mid \theta') \text{d}\theta'}$

$$= \frac{P(\theta)\ P(D_1 \mid \theta)\ P(D_2 \mid \theta)}{\int P(\theta')\ P(D_1 \mid \theta')\ P(D_2 \mid \theta') \text{d}\theta'}$$

[from multiplicativity of likelihood]

$$= \frac{P(\theta)\ P(D_1 \mid \theta)\ P(D_2 \mid \theta)}{\frac{k}{k} \int P(\theta')\ P(D_1 \mid \theta')\ P(D_2 \mid \theta') \text{d}\theta'}$$

[for random positive k]

$$= \frac{\frac{P(\theta)\ P(D_1 \mid \theta)}{k}\ P(D_2 \mid \theta)}{\int \frac{P(\theta')\ P(D_1 \mid \theta')}{k}\ P(D_2 \mid \theta') \text{d}\theta'}$$

[rules of integration; basic calculus]

$$= \frac{P(\theta \mid D_1)\ P(D_2 \mid \theta)}{\int P(\theta' \mid D_1)\ P(D_2 \mid \theta') \text{d}\theta'}$$

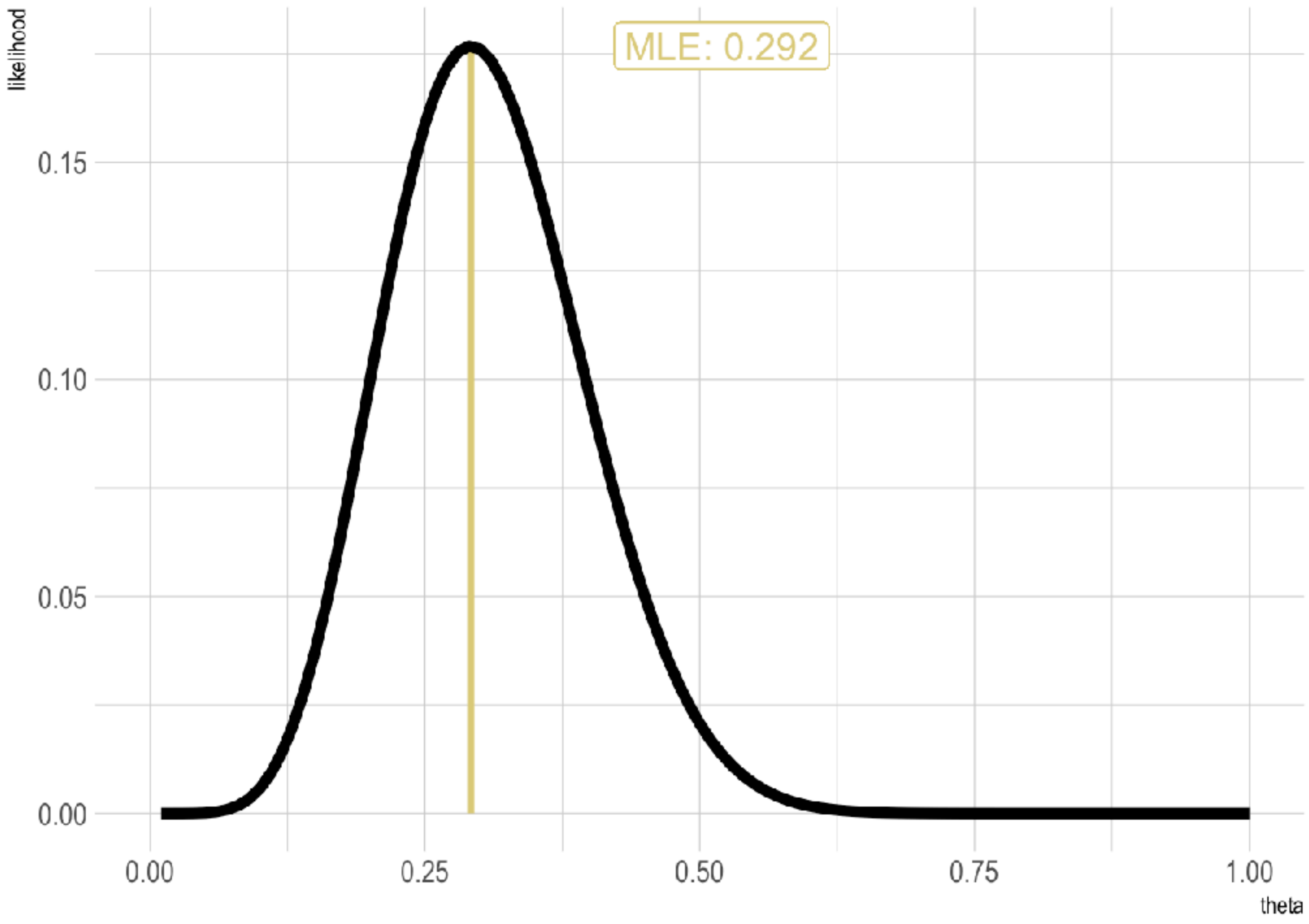
[Bayes rule with $k = \int P(\theta) P(D_1 \mid \theta) \text{d}\theta$]

frequentist estimation

MAXIMUM LIKELIHOOD ESTIMATE

- maximum likelihood estimate:

$$\hat{\theta} = \arg\max_{\theta} P(d \mid \theta)$$



CONFIDENCE INTERVAL [MATHEMATICALLY]

- let $\mathscr{D}$ be the random variable describing the probability of data
- X_l and X_u are random variables derived from $\mathscr{D}$ via functions g_l and g_u so that $g_{l,u} : D \mapsto \mathbb{R}$
- a $\gamma\,\%$ confidence interval for observed data D_{obs} is the interval:

$$\left[g_l(D_{\text{obs}}), g_u(D_{\text{obs}})\right]$$

- where functions $g_{l,u}$ are constructed so that:

$$P(X_l \le \theta_{\text{true}} \le X_u) = \frac{\gamma}{100}$$

- and where θ_{true} is the true value

CONFIDENCE INTERVAL [ALGORITHMICALLY]

- fix number of coin flips N (not really necessary, but easier)
- suppose the true coin bias is θ_{true} (but we don't know it)
- we have a magic function $MF: k \mapsto [u_k; l_k]$
- we now sample repeatedly $k \sim \text{Binomial}(N, \theta_{\text{true}})$
- for each sample k, compute $MF(k) = [u_k; l_k]$
- MF gives us a $\gamma\,\%$ confidence interval if θ_{true} is inside of $MF(k) = [u_k; l_k]$ in $\gamma\,\%$ of the sampled ks

comparison

BAYESIAN VS FREQUENTIST ESTIMATES

- for Bayesianism the full posterior is the primary object of concern; point- and interval-estimates are essentially just summary statistics for the full posterior
- for frequentists the point- and interval-estimates are the primary object of concern
- MLEs are much easier to compute but might not exist
- posteriors can be very hard to compute (long run time)

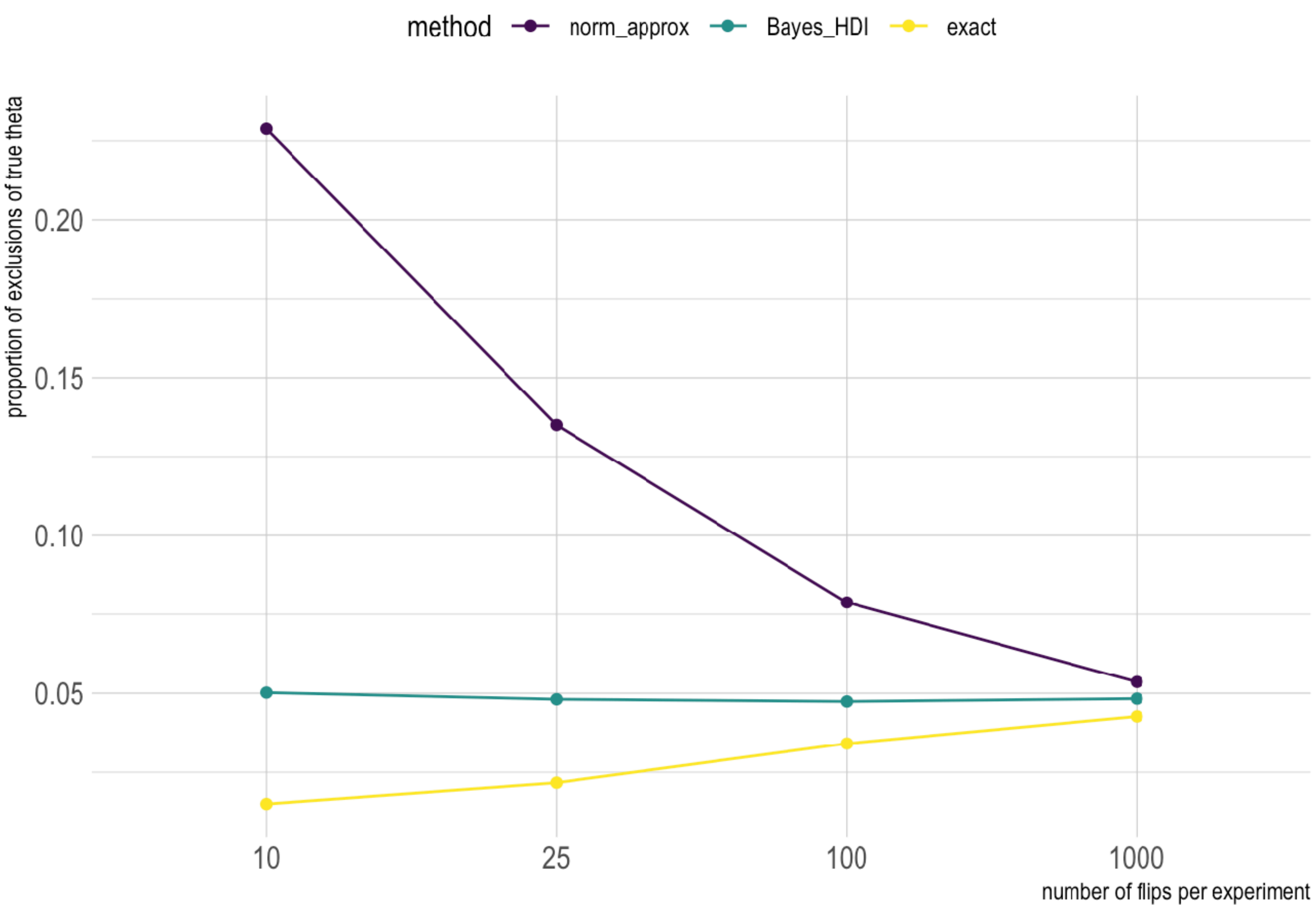
A PUZZLE ABOUT POINT-ESTIMATES

- flip a coin of unknown bias once
- suppose you see heads
- what's your best estimate of the bias?

SIMULATION-BASED COMPARISON OF INTERVAL-ESTIMATES

- fix $N \in \{10, 25, 100, 1000\}$
- repeatedly do:
  - sample $\theta_{\text{true}} \sim \text{Beta}(1,1)$
  - sample $k \sim \text{Binomial}(\theta_{\text{true}}, N)$
  - compute intervals for k and N
    - HDI, exact CI, approximate CI
- look at percentage that θ_{true} is included in each interval construction

RESULTS


method
norm_approx
Bayes_HDI
exact
proportion of exclusions of true theta
0.20
0.15
0.10
0.05
10
25
100
1000
number of flips per experiment