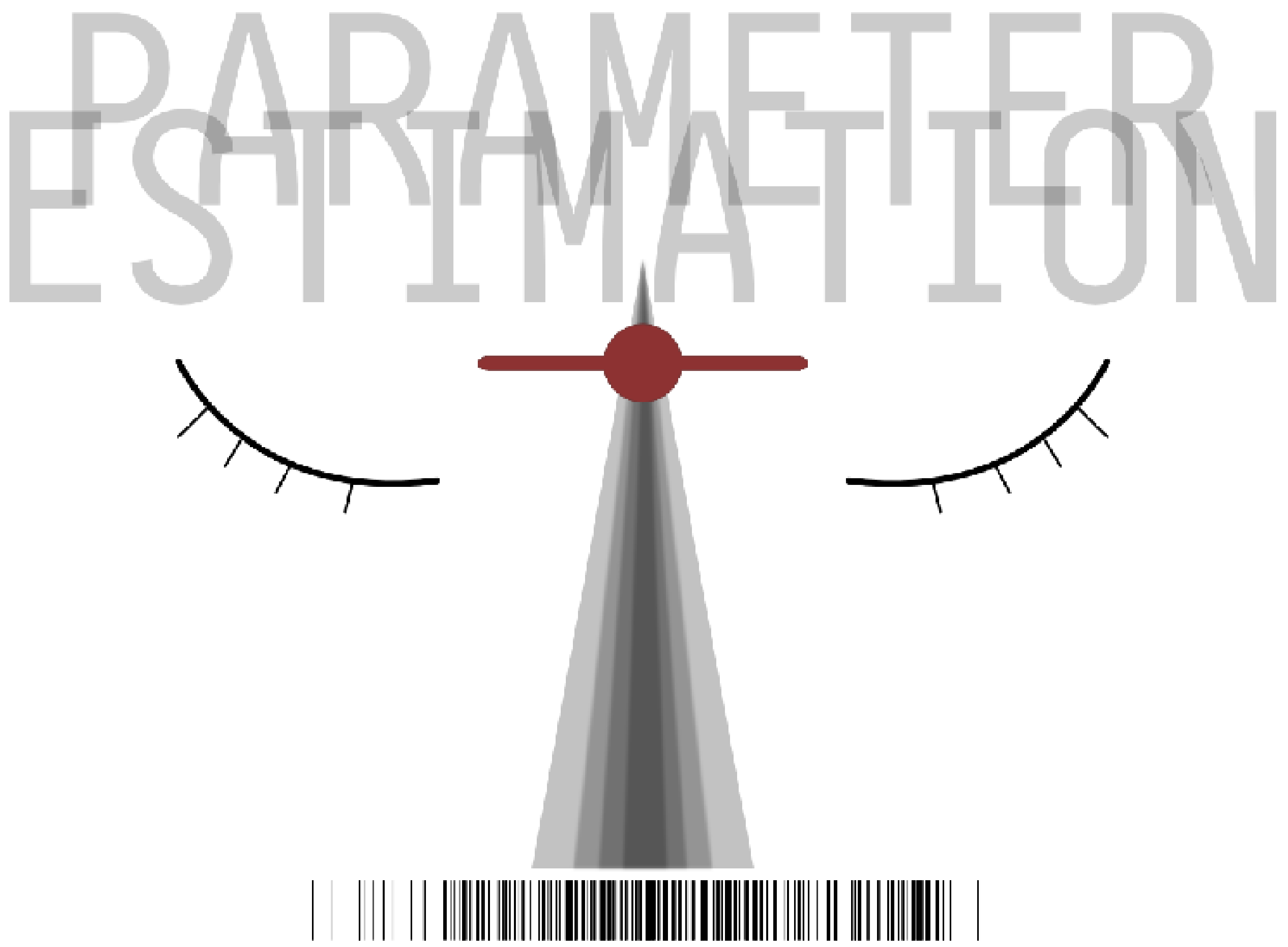
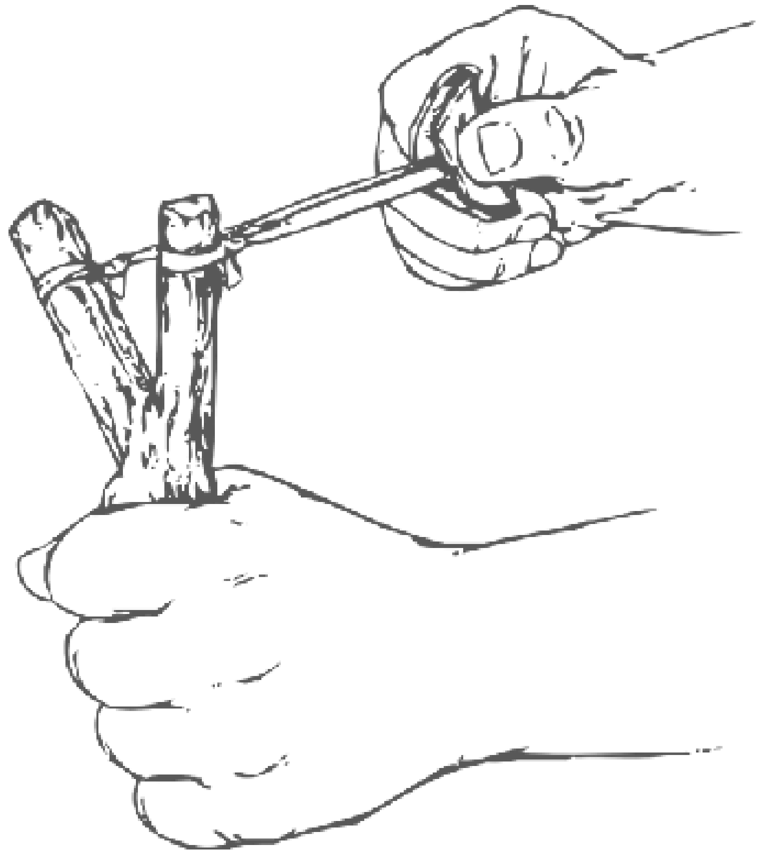
DATA ANALYSIS

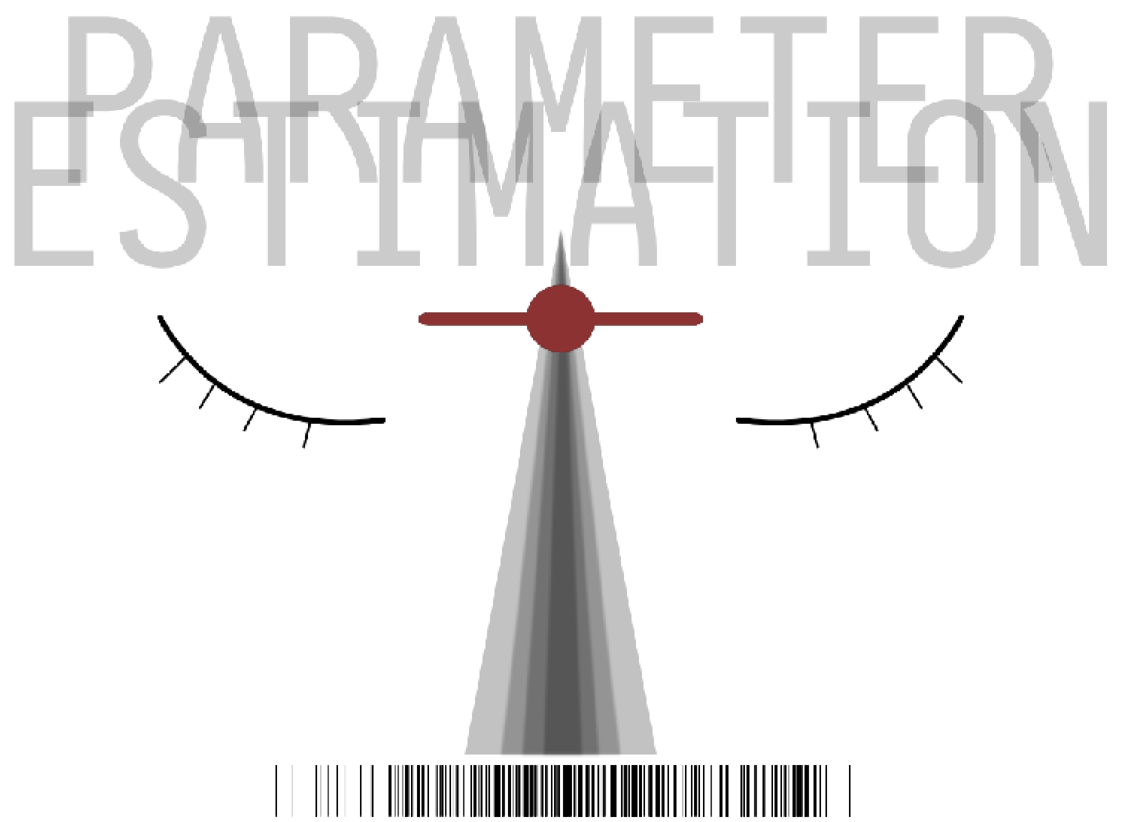



# PARAMETER ESTIMATION

# 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



# WHAT'S A MODEL PARAMETER

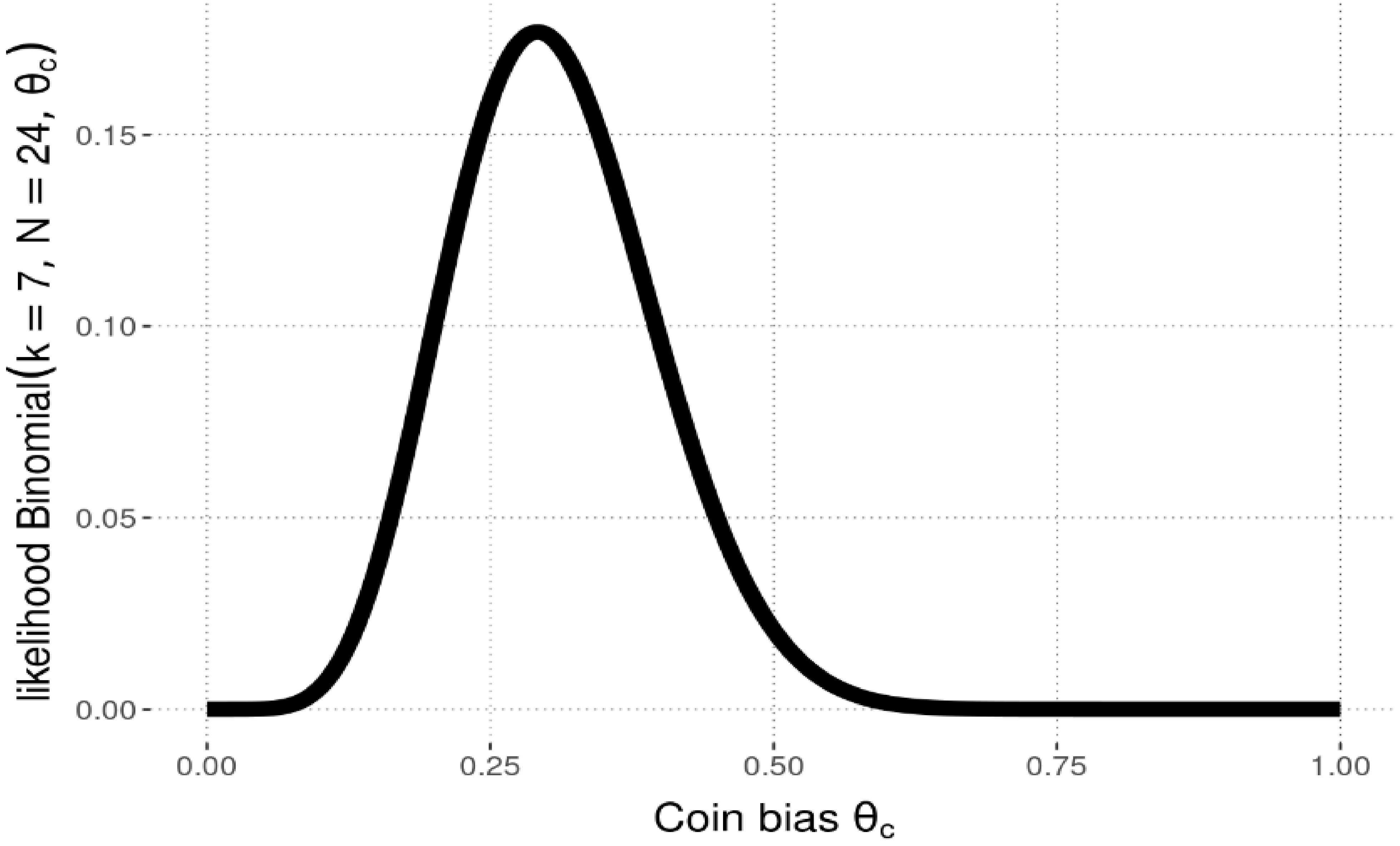
- A model parameter is a value that the likelihood depends on
- In the graphical notation we introduced in previous lecture, parameters usually (but not necessarily) show up as white nodes, because they are unknowns.

# WHAT'S A MODEL PARAMETER

For example, the single parameter $\theta_c$ in the Binomial Model shapes or fine-tunes the likelihood function.

Remember that the likelihood function for the Binomial Model is:

$$P_M(k \mid \theta_c, N) = \text{Binomial}(k, N, \theta_c) = \binom{N}{k} \theta_c^k (1 - \theta_c)^{N-k}$$



# EXERCISES

a. Use R to calculate how likely it is to get $k = 22$ heads when tossing a coin with bias $\theta_c = 0.5$ a total of $N = 100$ times.

b. Which parameter value, $\theta_c = 0.4$ or $\theta_c = 0.6$, makes the data from the previous part of this exercise ($N = 100$ and $k = 22$) more likely? - Give a reason for your intuitive guess and use R to check your intuition.

# 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 |

# 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 $P(\Theta = \theta \mid \mathcal{D} = D)$
- short-hand with non-normalized probabilities (implicit normalizing constant):

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

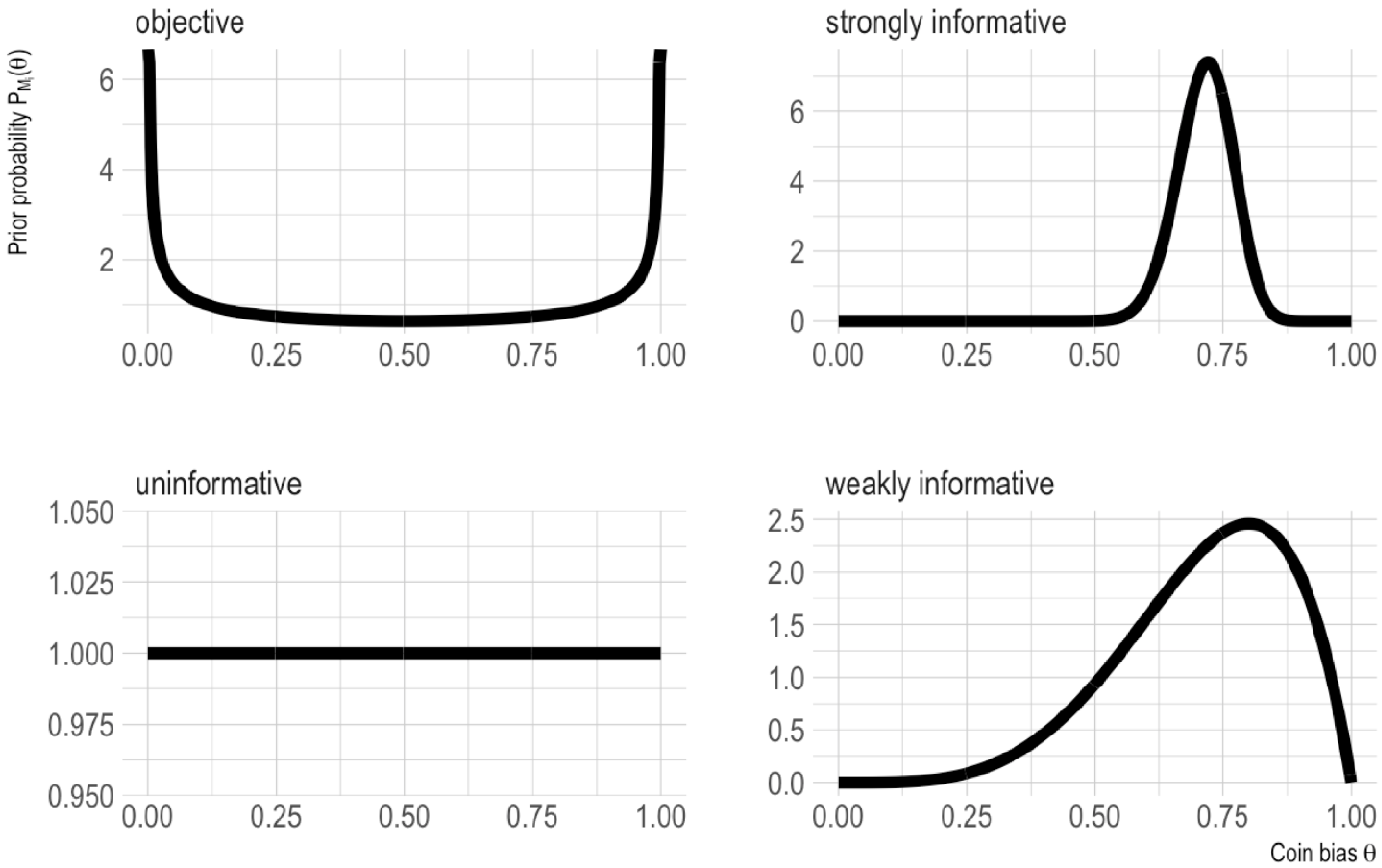
# EXAMPLE

- model:

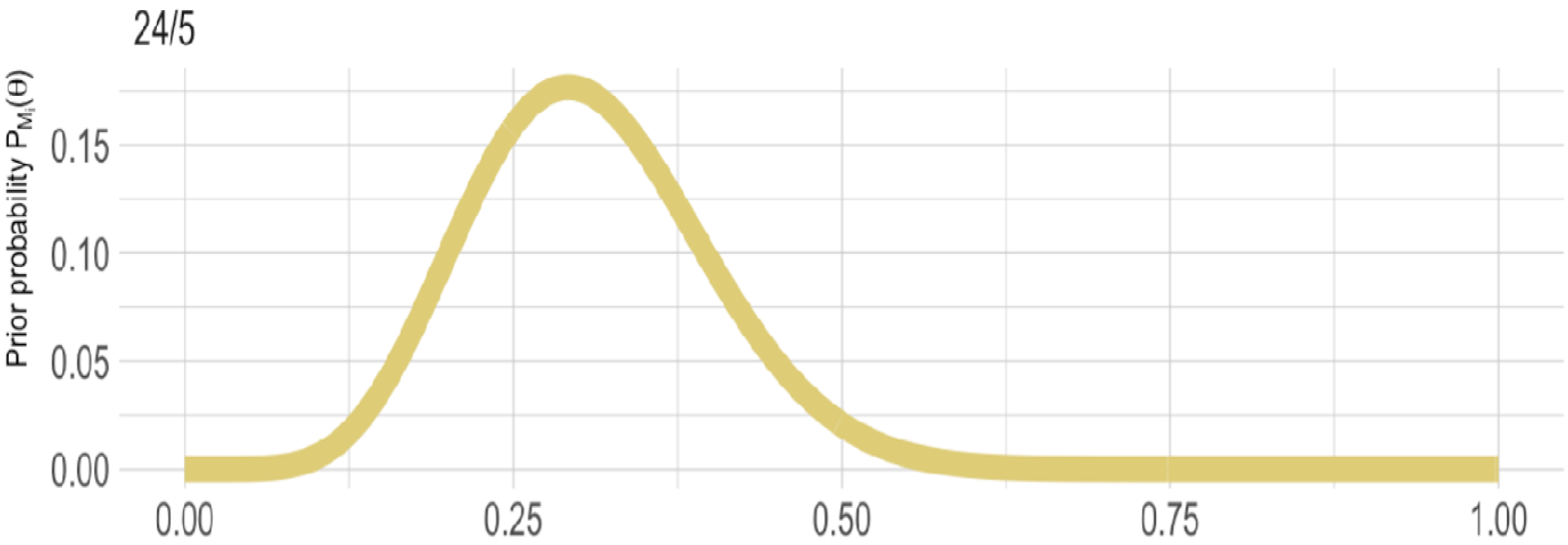
$$k \sim \mathrm{Binomial}(N, \theta)$$
$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

- data:
  - "24/7" $k = 7$ $N = 24$

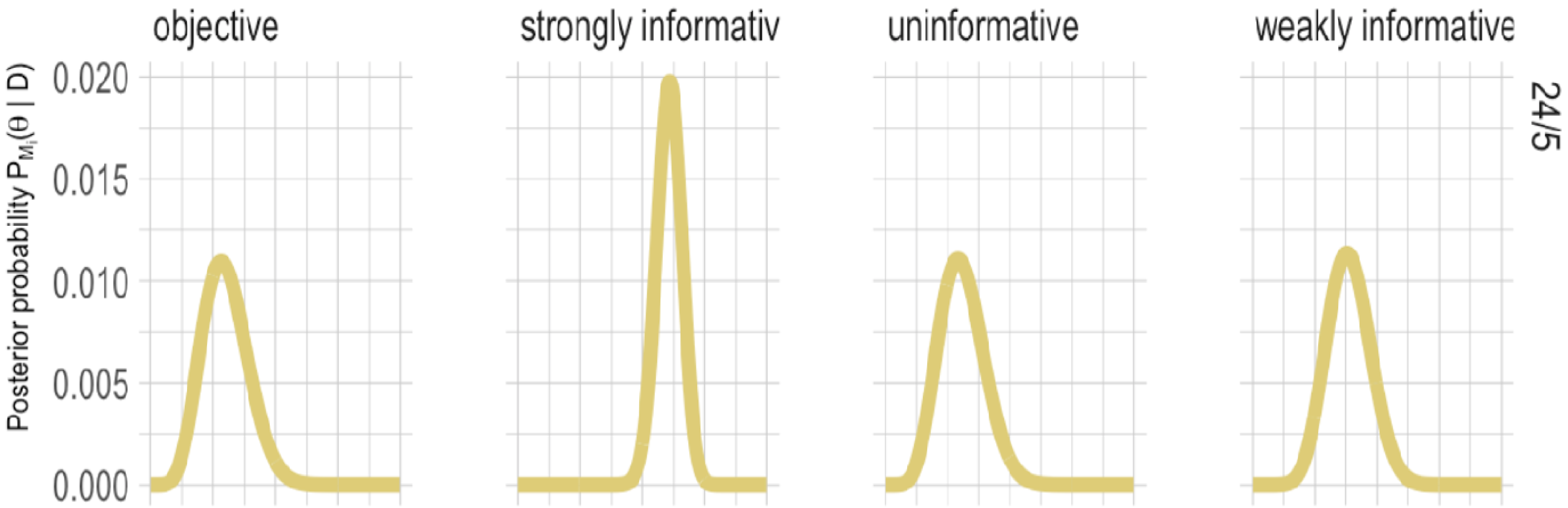
# PRIOR


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

# 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

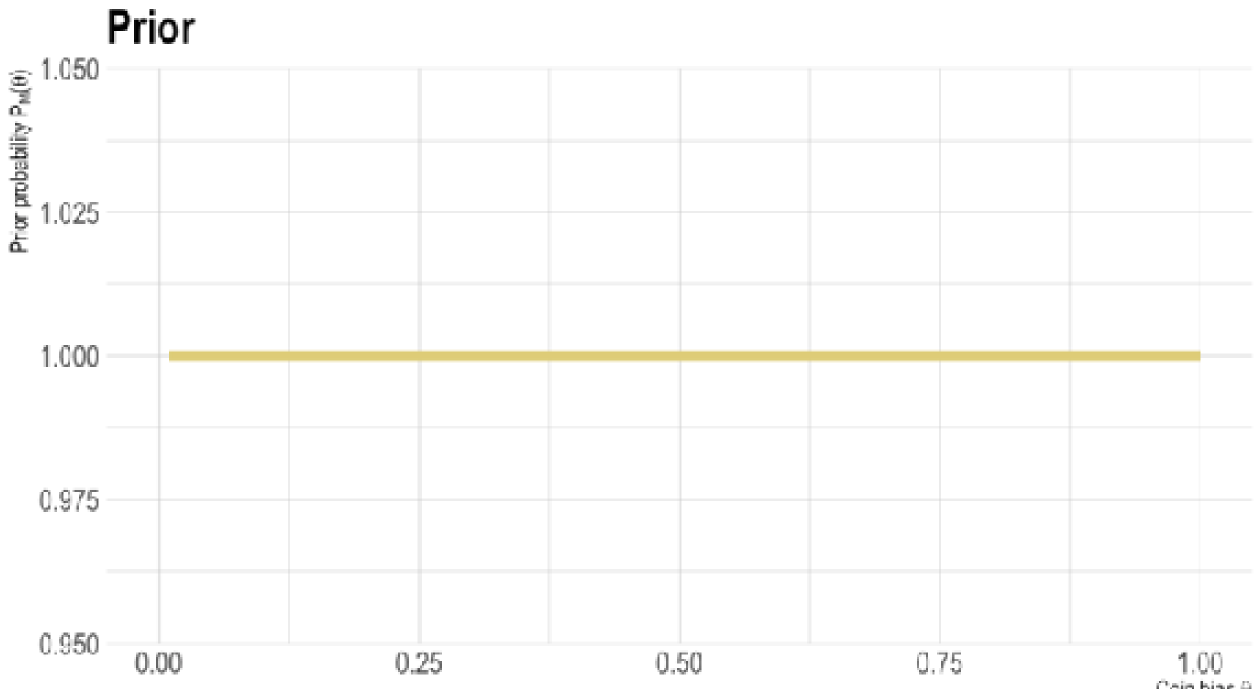
# POSTERIOR

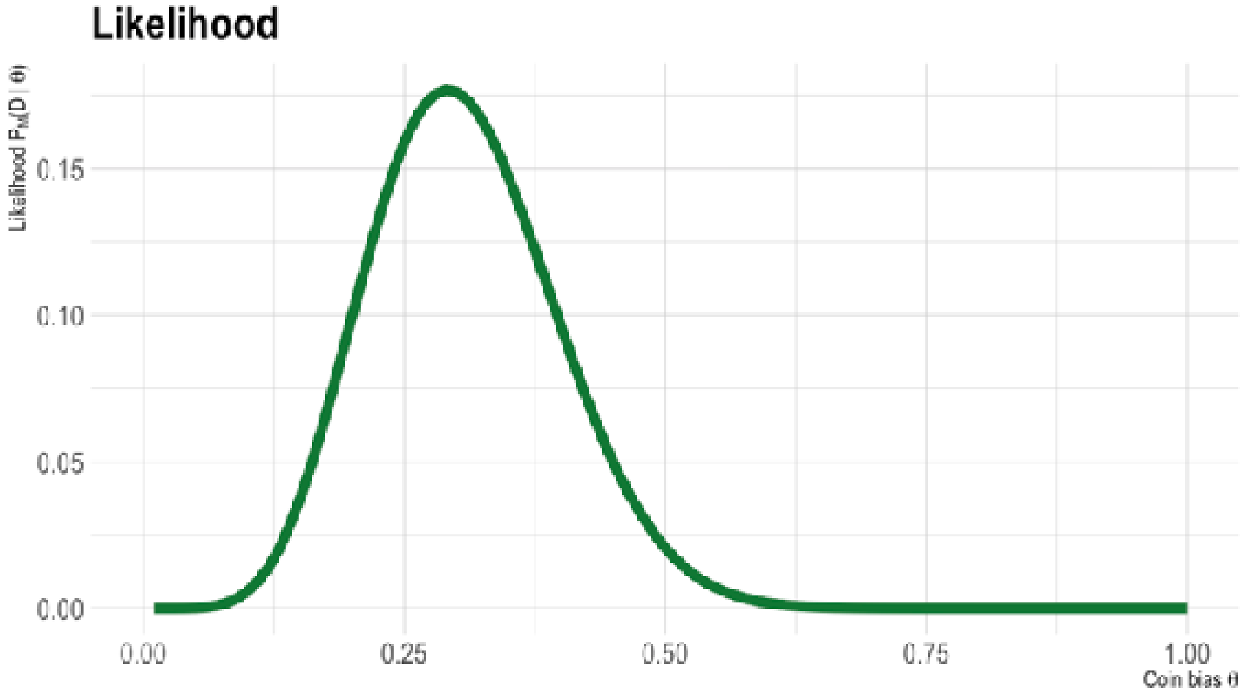

objective
strongly informativ
uninformative
weakly informative
Posterior probability $P_{M_i}(\theta \mid D)$
0.020
0.015
0.010
0.005
0.000
24/5

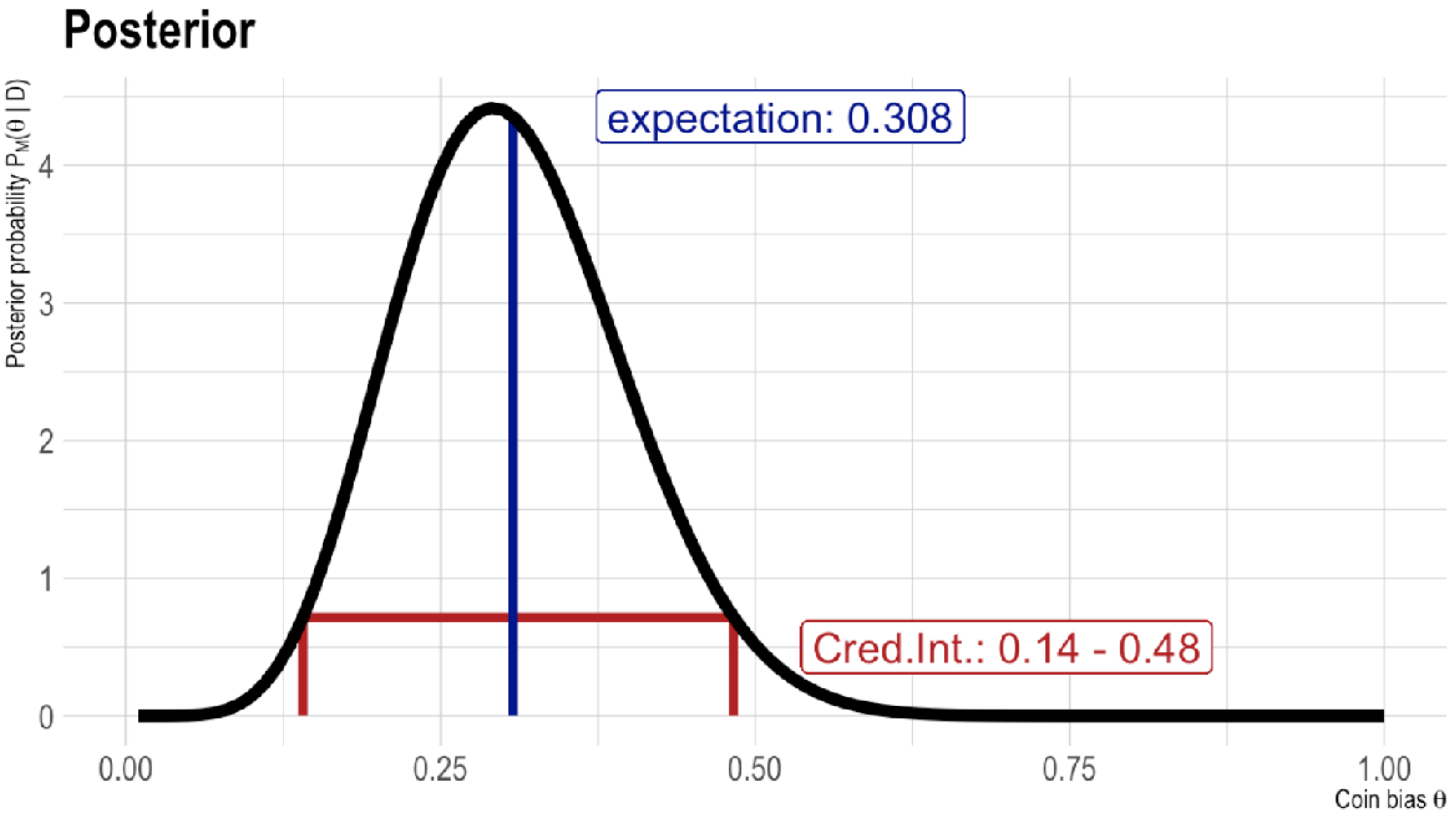
# Bayesian point- & interval-estimates

# EXAMPLE

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



# POSTERIOR MEAN &MAP

- posterior mean:

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

- maximum a posteriori:

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

- 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

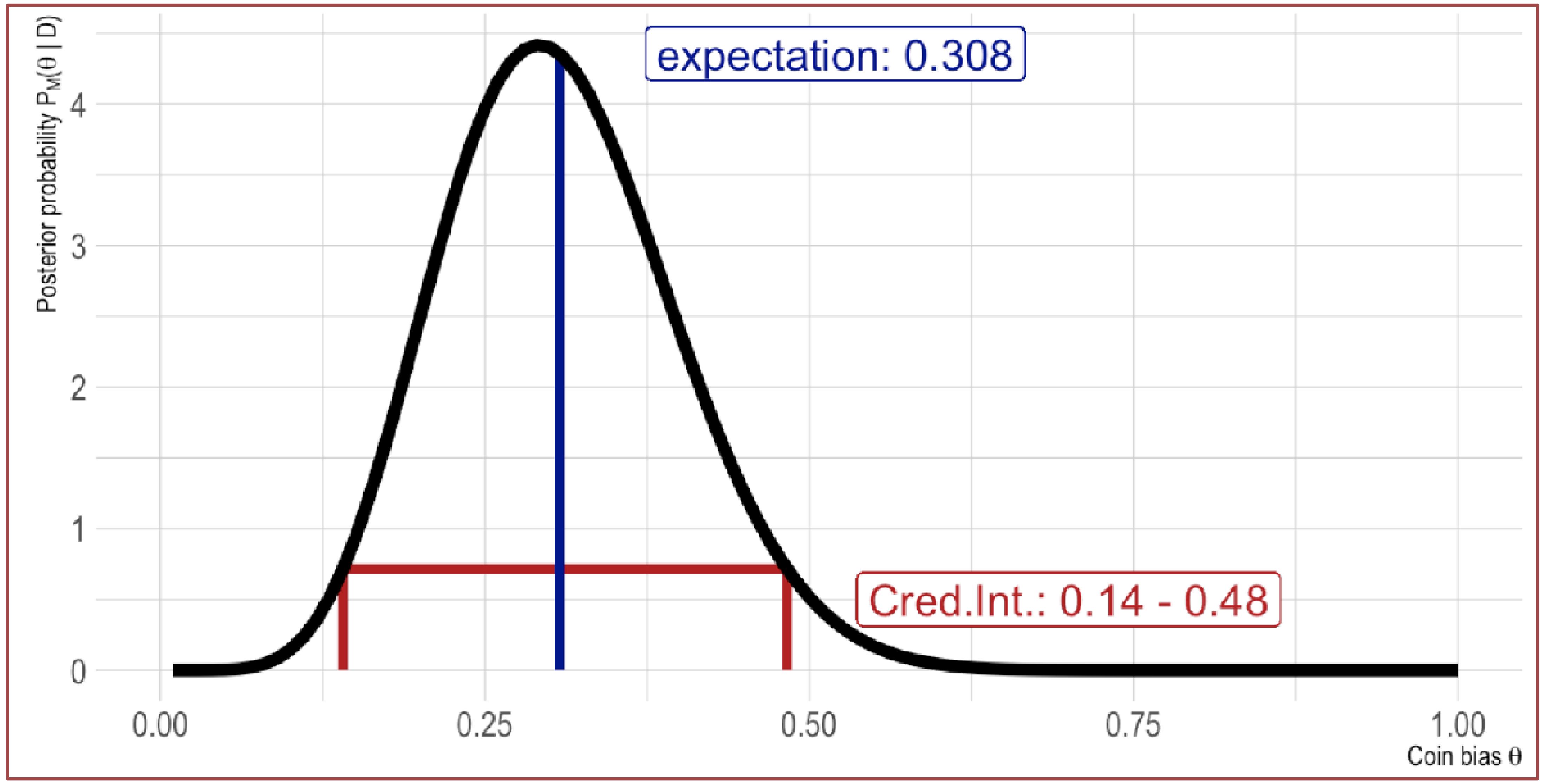
- interval $[l; u]$ is a $\gamma\%$ credible interval for a random variable $X$ if

  (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:

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

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

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

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

# EXERCISES

a. Fill in the blanks in the code below to get a plot of the posterior distribution for the coin flip scenario with $k = 20$, $N = 24$, making use of conjugacy and starting with a uniform Beta prior.

```
theta = seq(0, 1, length.out = 401)

as_tibble(theta) %>%
  mutate(posterior =  ____ ) %>%
  ggplot(aes(___, posterior)) +
   geom_line()
```

# EXERCISES

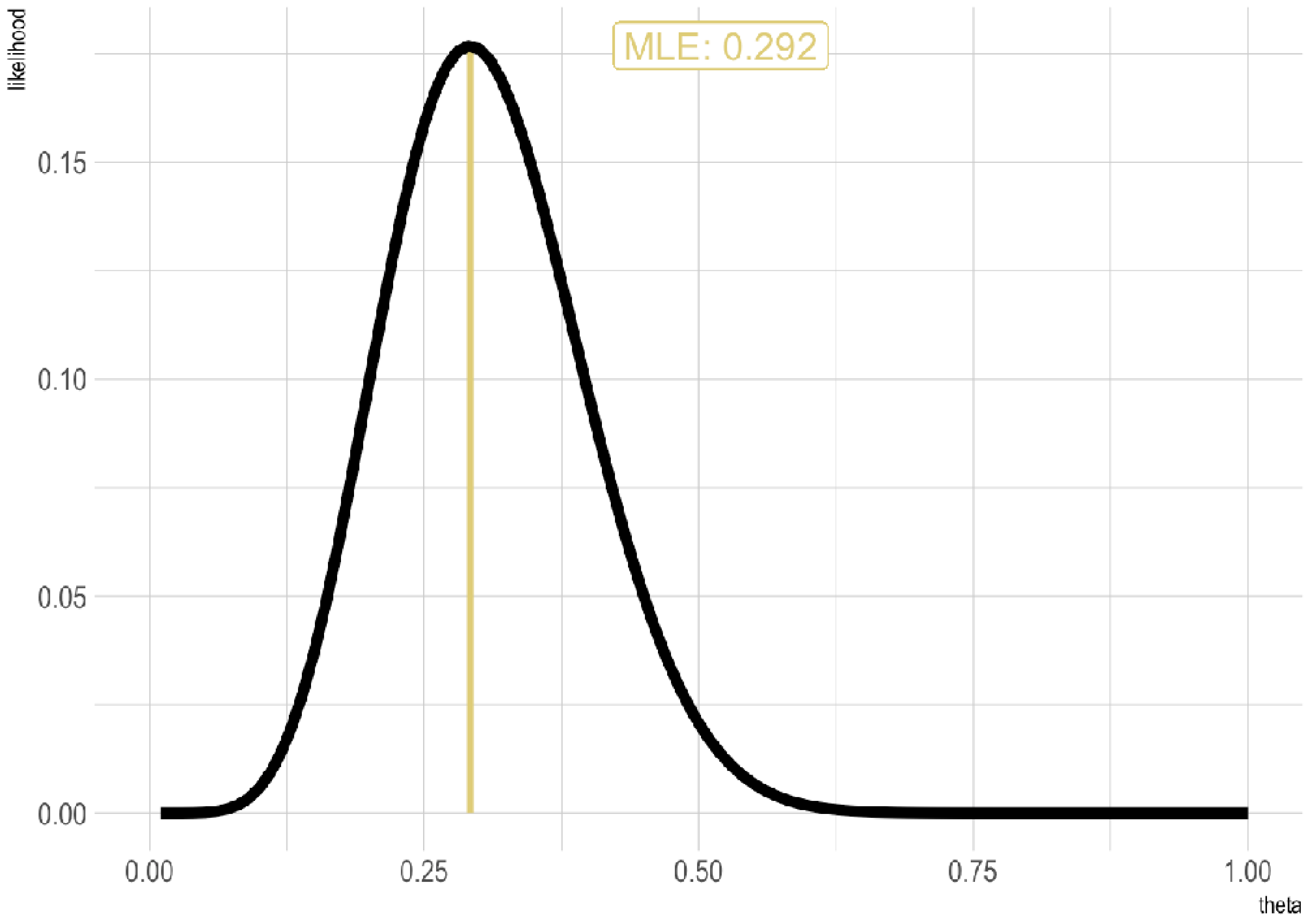
b. Suppose that Jones flipped a coin with unknown bias 30 times. She observed 20 heads. She updates her beliefs rationally with Bayes rule. Her posterior beliefs have the form of a beta distribution with parameters $\alpha = 25$, $\beta = 15$. What distribution and what parameter values of that distribution capture Jones' prior beliefs before updating her beliefs with this data?

# frequentist estimation

# MAXIMUM LIKELIHOOD ESTIMATE

- maximum likelihood estimate:

$$\hat{\theta} = \underset{\theta}{\arg\max}\ P(D \mid \theta)$$



# EXERCISES

Can you think of a situation where MLE and MAP are the same? HINT: Think which prior eliminates the difference between them!

# CONFIDENCE INTERVAL (MATH)

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

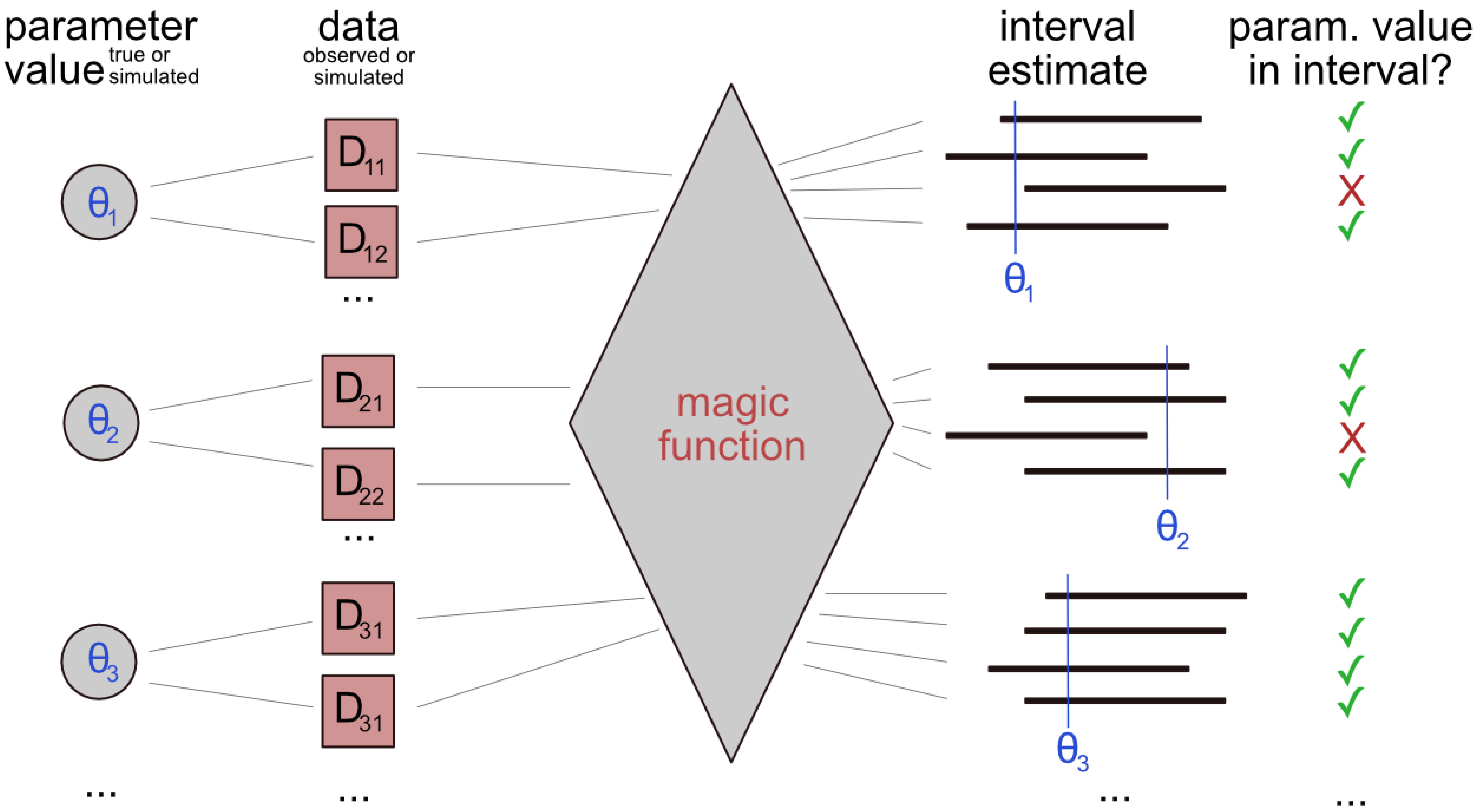
$$[g_l(D_{\text{obs}}), g_u(D_{\text{obs}})]$$

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

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

- and where $\theta_{\text{true}}$ is the true value

# CONFIDENCE INTERVAL (ALGORITHM)



# CONFIDENCE INTERVAL (ALGORITHM)

- fix number of coin flips $N$ (not really necessary, but easier)
- suppose the true coin bias is $\theta_{\text{true}}$ (but we don't know it)
- we have a magic function $MF$ : $\quad k \mapsto [u_k; l_k]$
- 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 $\theta_{\text{true}}$ is inside of $MF(k)$ = $[u_k; l_k]$ in $\gamma\%$ of the sampled $k$

# addressing point-valued hypotheses with estimation

# ADDRESSING POINT-VALUED HYPOTHESES (BAYES)

- $\Theta_i = \theta_i^*$ is out point-valued hypothesis
- a region of practical equivalence [ROPE] is an $\epsilon$ -region around $\theta_i^*$:
  $\mathrm{ROPE}(\theta_i^*) = [\theta_i^* - \epsilon, \theta_i^* + \epsilon]$
- for a Bayesian credible interval $[l; u]$ for $\Theta_i$, we:
  - accept the point-valued hypothesis iff $[l; u]$ is contained entirely in $\mathrm{ROPE}(\theta_i^*)$;
  - reject the point-valued hypothesis iff $[l; u]$ and $\mathrm{ROPE}(\theta_i^*)$ have no overlap;
  - withhold judgement otherwise.

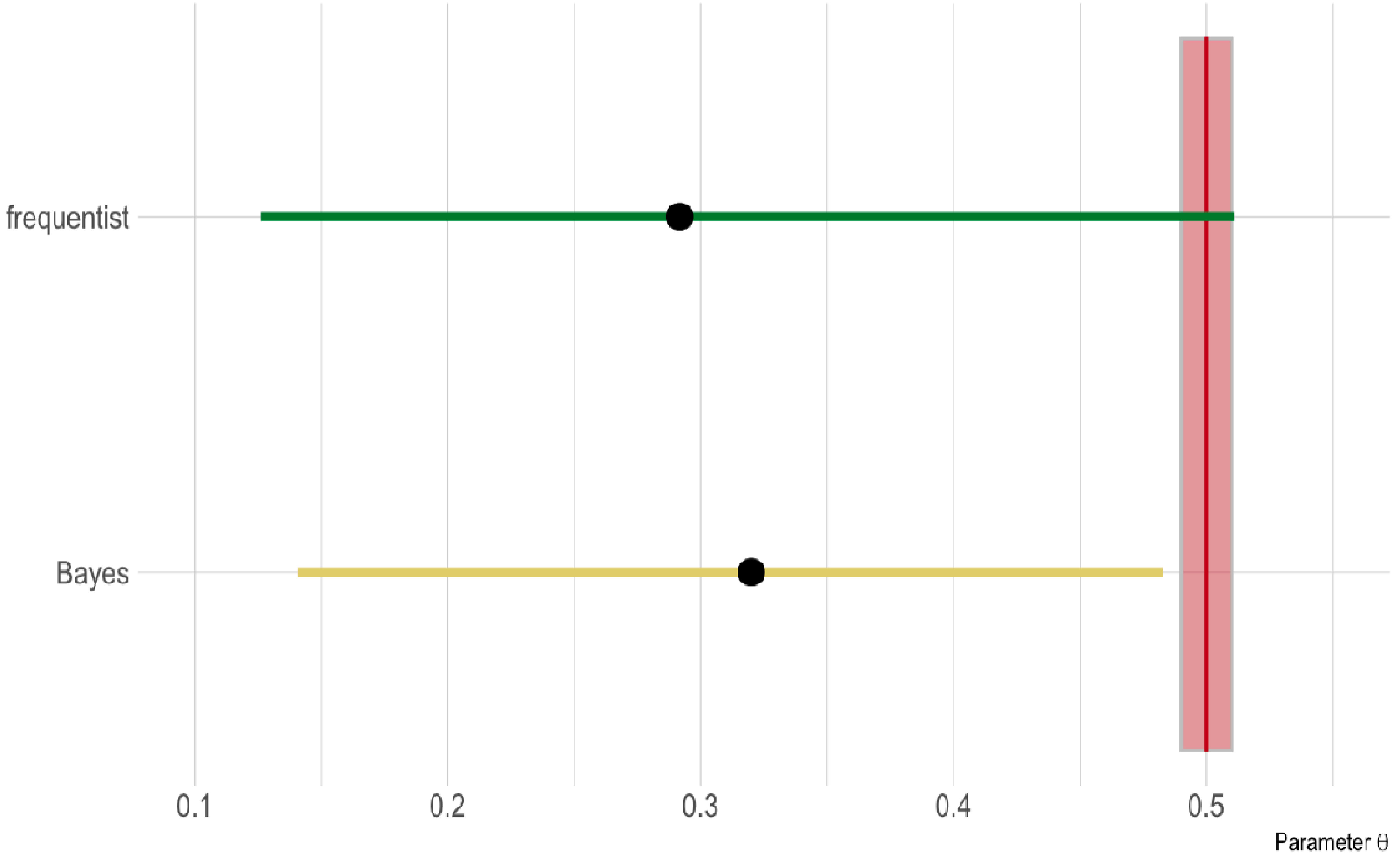
# ADDRESSING POINT-VALUED HYPOTHESES (FREQUENTIST)

- $\Theta_i = \theta_i^*$ is out point-valued hypothesis
- we do not consider a ROPE
- for a frequentist credible interval $[l; u]$ for $\Theta_i$, we:
  - reject the point-valued hypothesis iff $\theta_i^* \notin [l; u]$; and
  - withhold judgement otherwise.

# EXAMPLE

- 24/7 example, uninformative priors for Bayesian model
- point- and interval estimates:

```
## # A tibble: 2 x 4
##   approach    lower point upper
##   <chr>       <dbl> <dbl> <dbl>
## 1 Bayes       0.141 0.32  0.483
## 2 frequentist 0.126 0.292 0.511
```



# computing estimates

# OPTIMIZING FUNCTIONS

```r
# function for the negative log-likelihood of the given
# data and fixed parameter values
nll = function(y, x, beta_0, beta_1, sd) {
  # negative sigma is logically impossible
  if (sd <= 0) {return( Inf )}
  # predicted values
  yPred = beta_0 + x * beta_1
  # negative log-likelihood of each data point
  nll = -dnorm(y, mean=yPred, sd=sd, log = T)
  # sum over all observations
  sum(nll)
}
```

```r
fit_lh = optim(
  # initial parameter values
  par = c(1.5, 0, 0.5),
  # function to optimize
  fn = function(par) {
    with(avocado_data,
         nll(average_price, total_volume_sold,
             par[1], par[2], par[3])
    )
  }
)
fit_lh$par
```

```
## [1]  1.425080e+00 -2.247373e-08  3.950978e-01
```

```r
lm(average_price ~ total_volume_sold, avocado_data)$coef
```

```
##       (Intercept) total_volume_sold
##      1.425096e+00     -2.247455e-08
```