Bayesian regression modeling: Theory & practice

Part 1: Bayesian basics & simple linear regression

Michael Franke

Motivation, background, and formalities

Bayesian data analysis

At a glance

- BDA is about what we *should* believe given:
  - some observable data, and
  - our model of how this data was generated (a.k.a. **the data-generating process**)
- our best friend will be **Bayes rule**
  - e.g., for **parameter inference**:

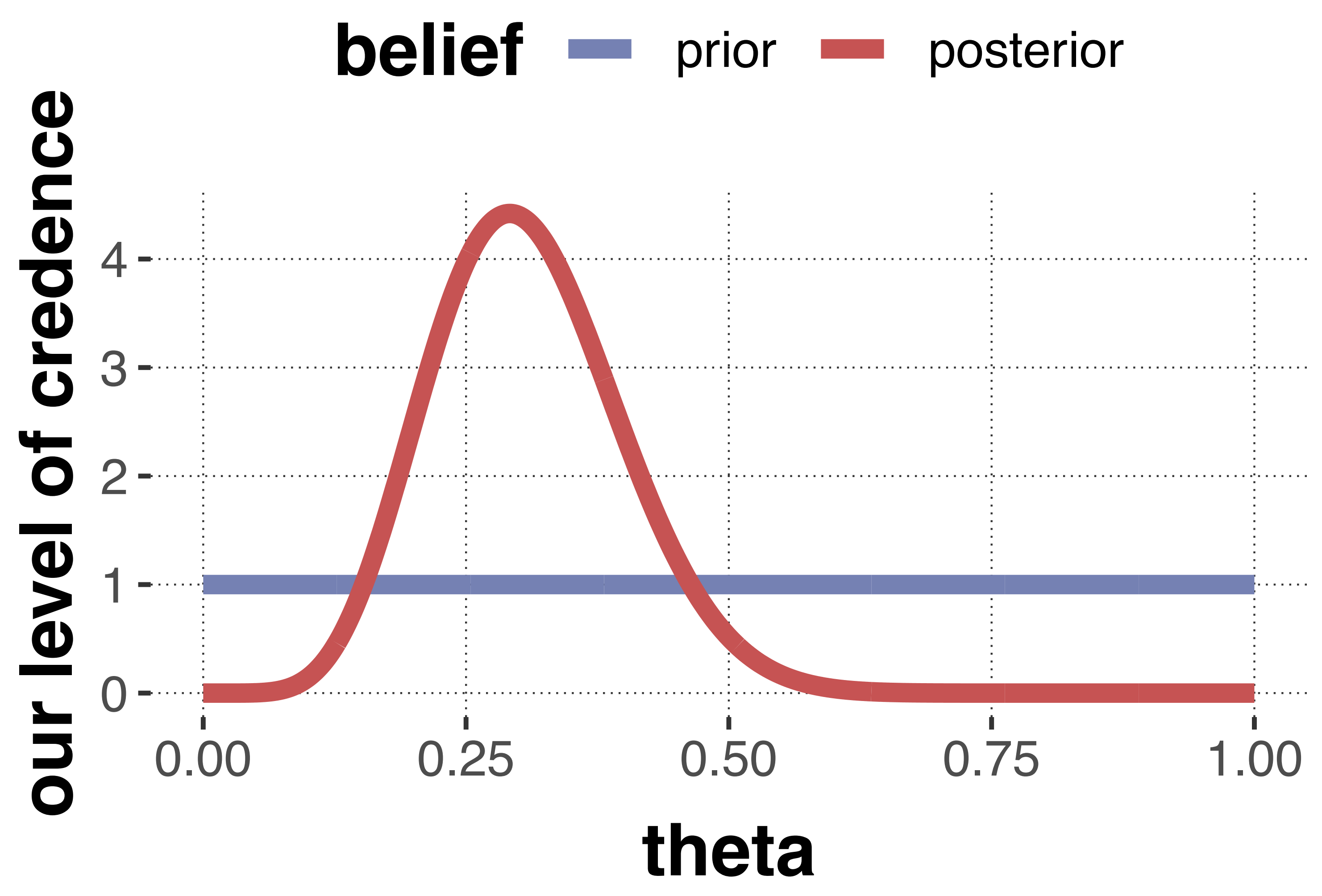
$$\underbrace{P(\theta \mid D)}_{\text{posterior}} \propto \underbrace{P(\theta)}_{\text{prior}} \times \underbrace{P(D \mid \theta)}_{\text{likelihood}}$$

  - or, for **model comparison**:

$$\underbrace{\frac{P(M_1 \mid D)}{P(M_2 \mid D)}}_{\text{posterior odds}} = \underbrace{\frac{P(D \mid M_1)}{P(D \mid M_2)}}_{\text{Bayes factor}} \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}}$$

Running example: 24/7

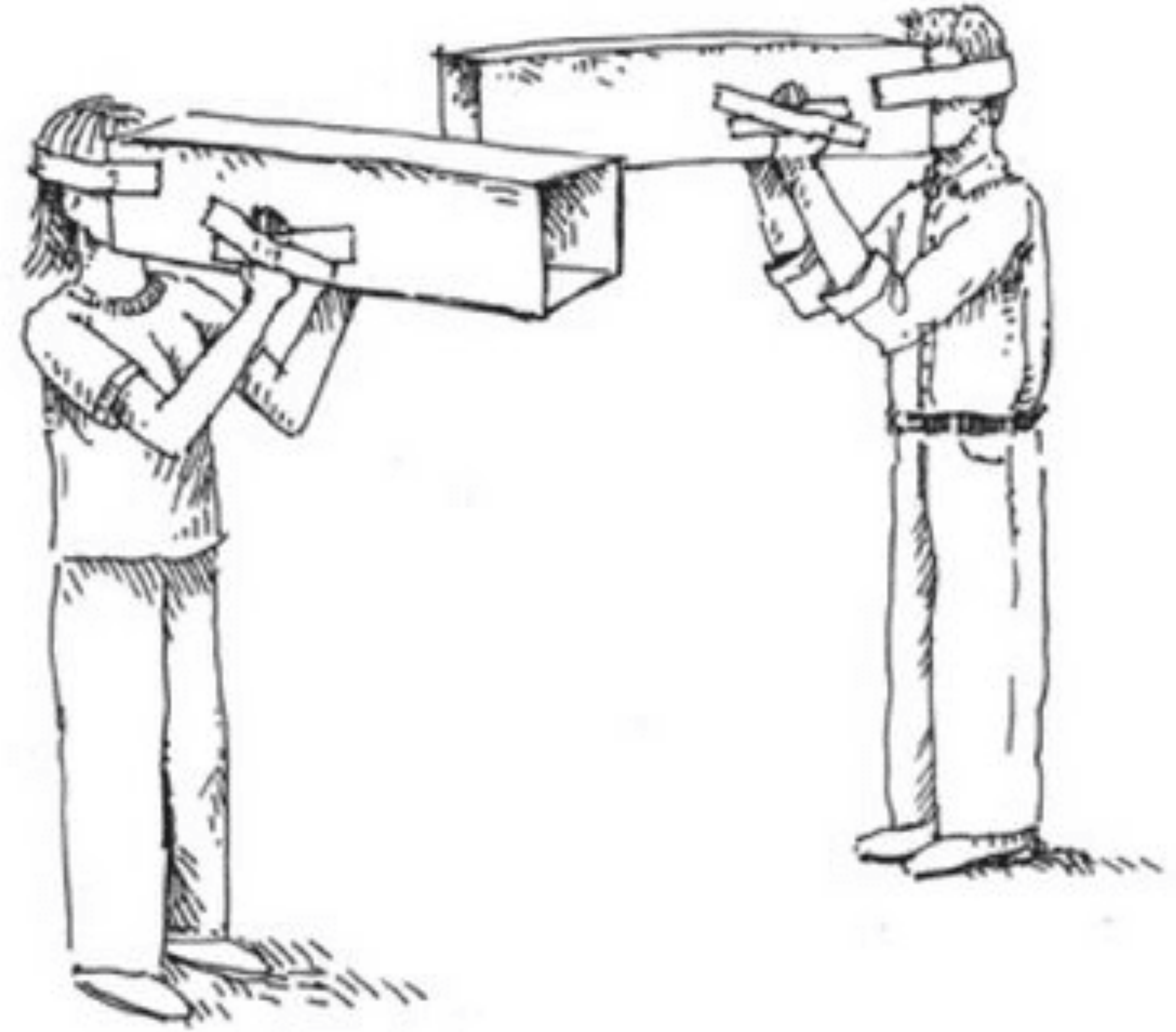
- $\theta \in [0; 1]$ is the bias of a coin
- *a priori* any value of θ is equally likely
- we observe 7 heads in 24 flips
- what should we believe about θ?



Classical frequentist statistics

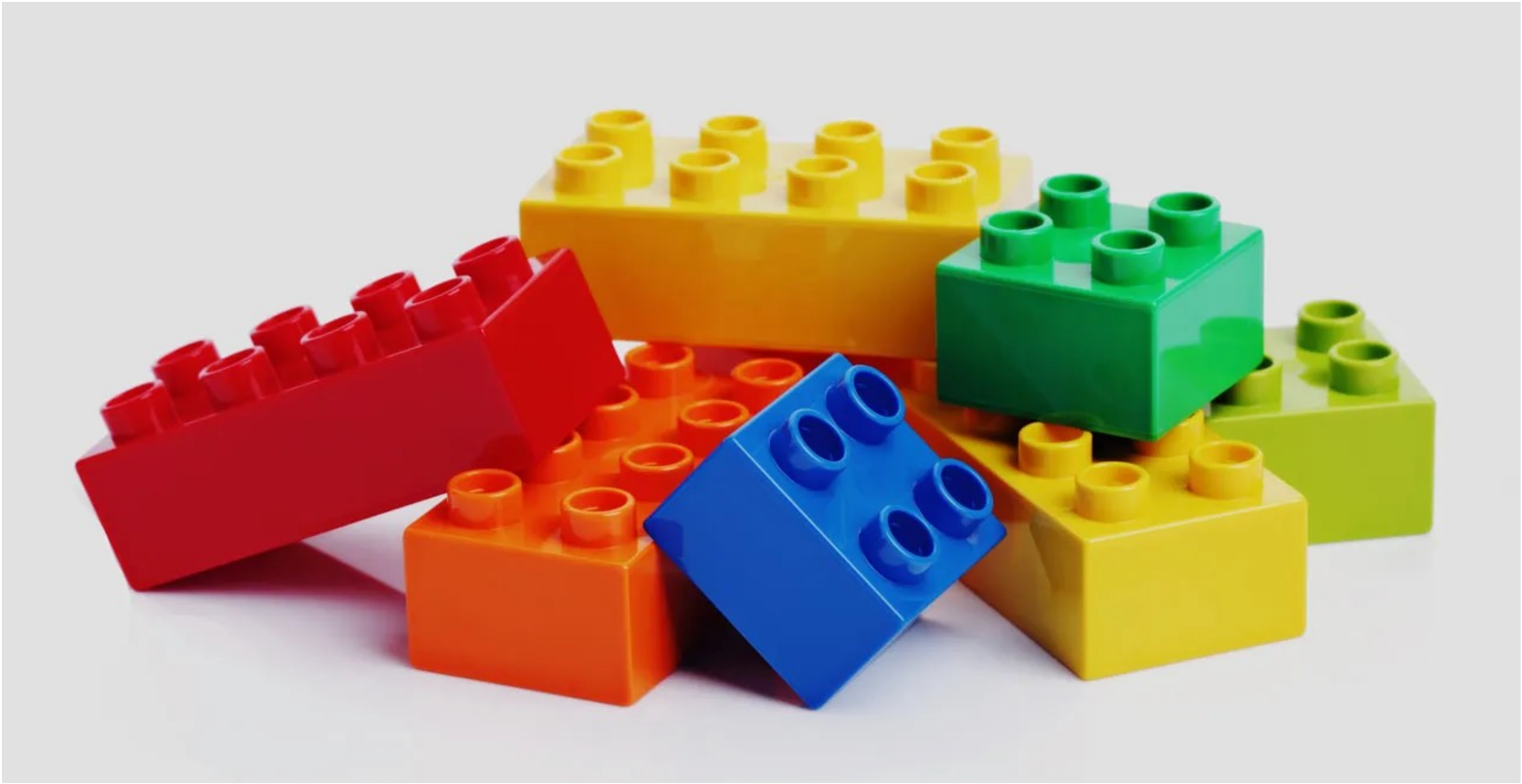
An op-ed

- based on **null-hypothesis significance testing**
  - e.g., is $\theta = 0.5$
- intrinsically married to binary decision-making:
  - accept or reject null-hypothesis
  - prime example of "tyranny of the discontinuous mind"
- relies on "sampling distributions"
  - hidden, and usually simplified assumptions about the data-generating process
  - rely on experimenter intentions, not an objective picture of the DGP
- point-estimates instead of distributions
  - less informative & error-prone
- unprincipled; bag of tricks; hard to customize



Pros
of BDA

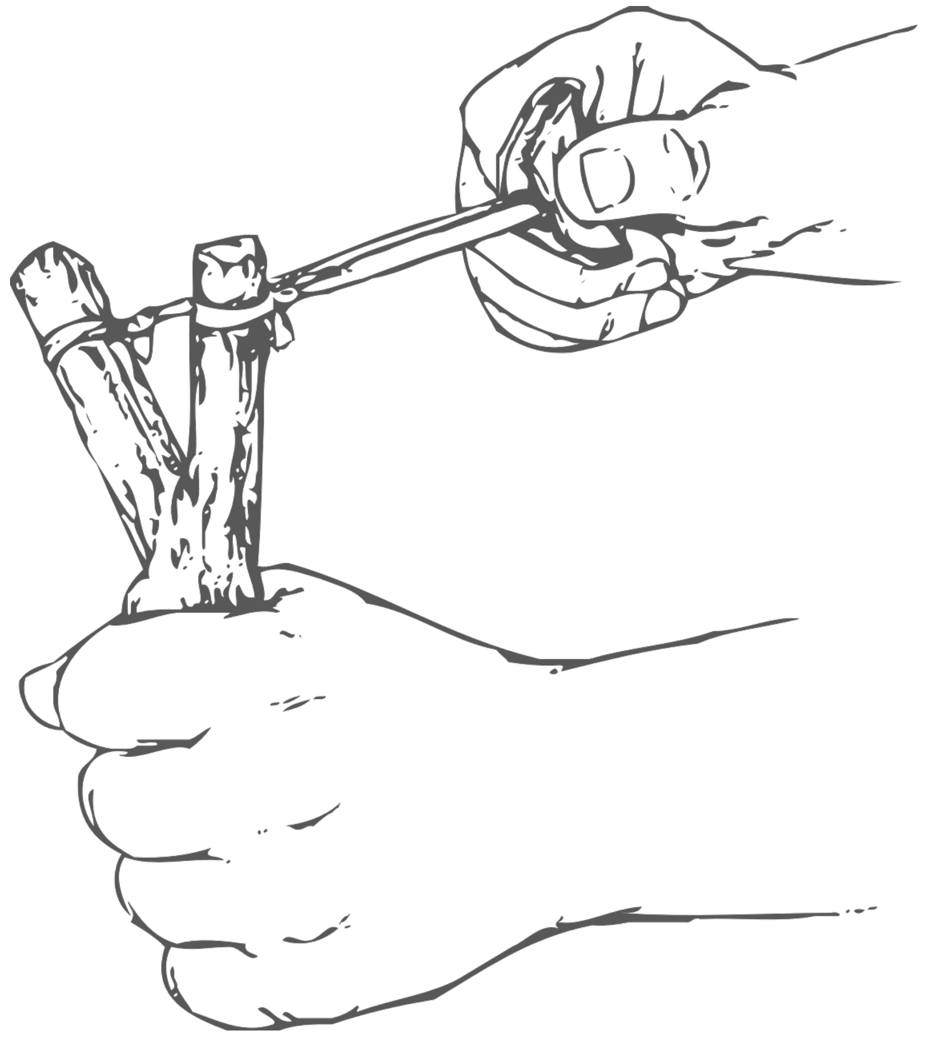
- well-founded & totally general
- easily extensible / customizable
- more informative / insightful
- stimulates view: “models as tools”



Cons
of BDA

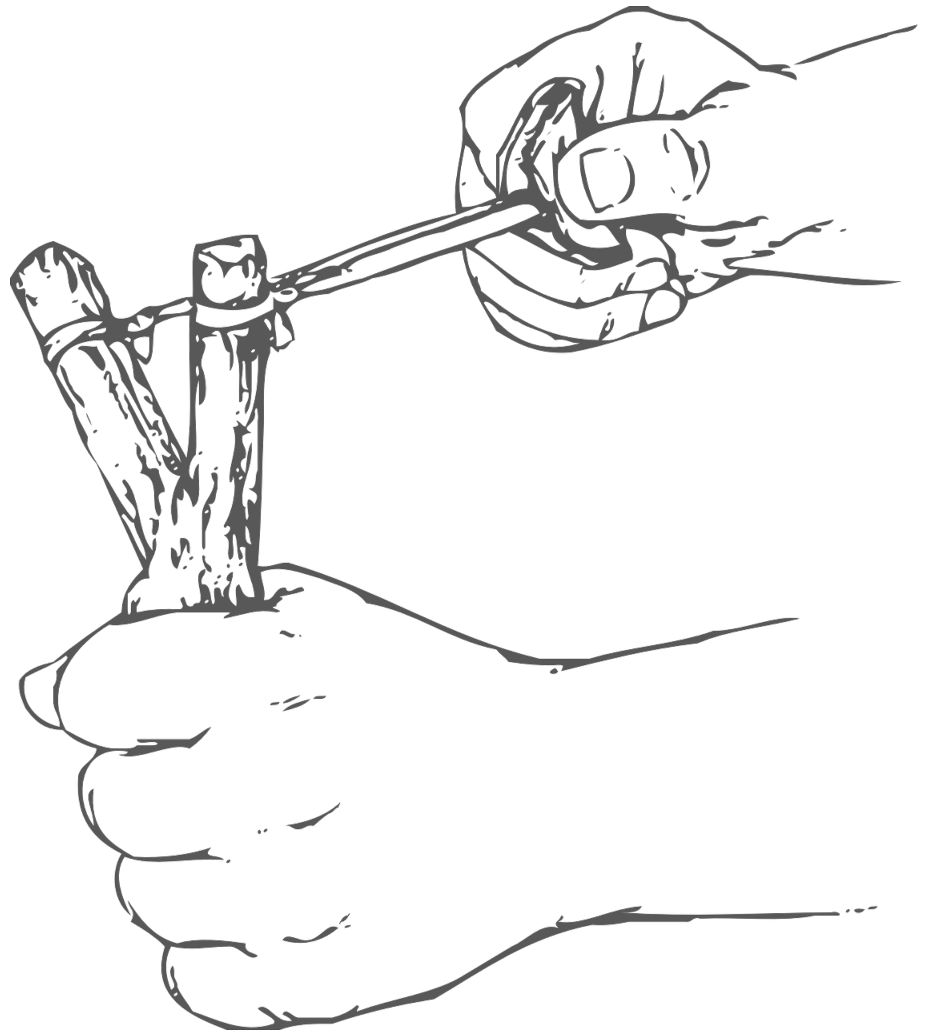
- not yet fully digested by community
- possibly computationally complex
- less ready-made, more hands-on
- requires thinking (wait, that’s a pro!)
  - last two points less valid than 10 years ago

Main learning goals



1. understand key concepts of Bayesian data analysis
    a. priors, posteriors & likelihood
    b. prior & posterior predictives
    c. Bayes factors
    d. Bayesian computation (MCMC)
2. be able to apply hierarchical generalized linear regression modeling
    a. determine the appropriate (kind of) model for a given problem
    b. implement, run and interpret the Bayesian model
    c. draw conclusions regarding evidence for/against research questions

Organization



- class from 9:00 – 14:30
- practical exercises for in class and at home
  - no homework, no need to hand in exercises, no grades
- final take-home exam
  - released on FILL ME
  - due on FILL ME
  - **no group-work! individual submissions only!**

Schedule

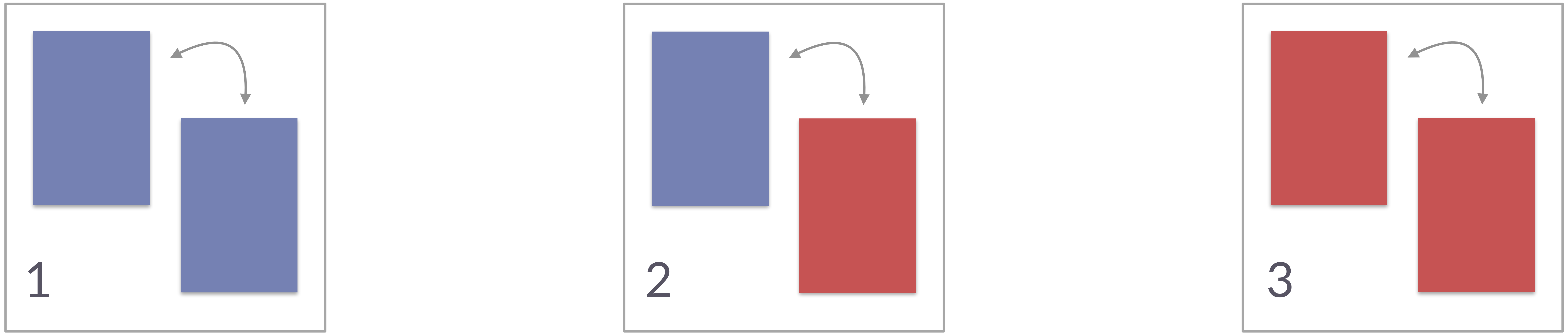
	Day 1	Day 2	Day 3	Day 4	Day 5
Slot 1	basics of BDA	priors & predictions	generalized lin. model	MCMC	Model comparison
Slot 2	simple lin. regression	categorical predictors	🛠	hierarchical regression	🛠
Slot 3	🛠	🛠		🛠	

Bayesian Basics

Three-card problem

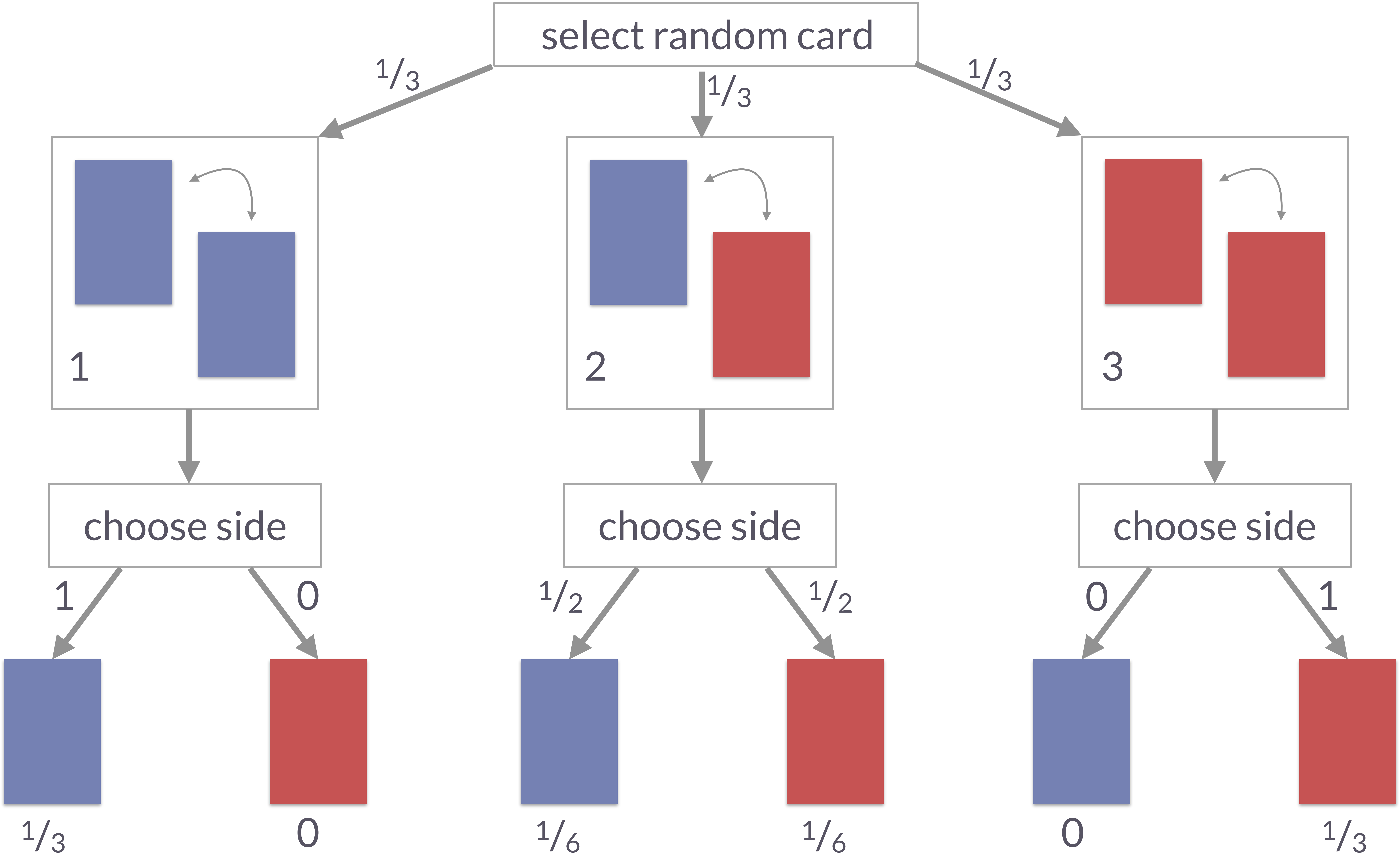
problem statement

- Sample a card (uniformly at random).
- Choose a side of that card to reveal (uniformly at random).
- What's the probability that the side you do not see is **BLUE**, given that the side you see is **BLUE**?



Three-card problem

data-generating process



Conditional probability and Bayes rule

for the three-card problem

- conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Bayes rule

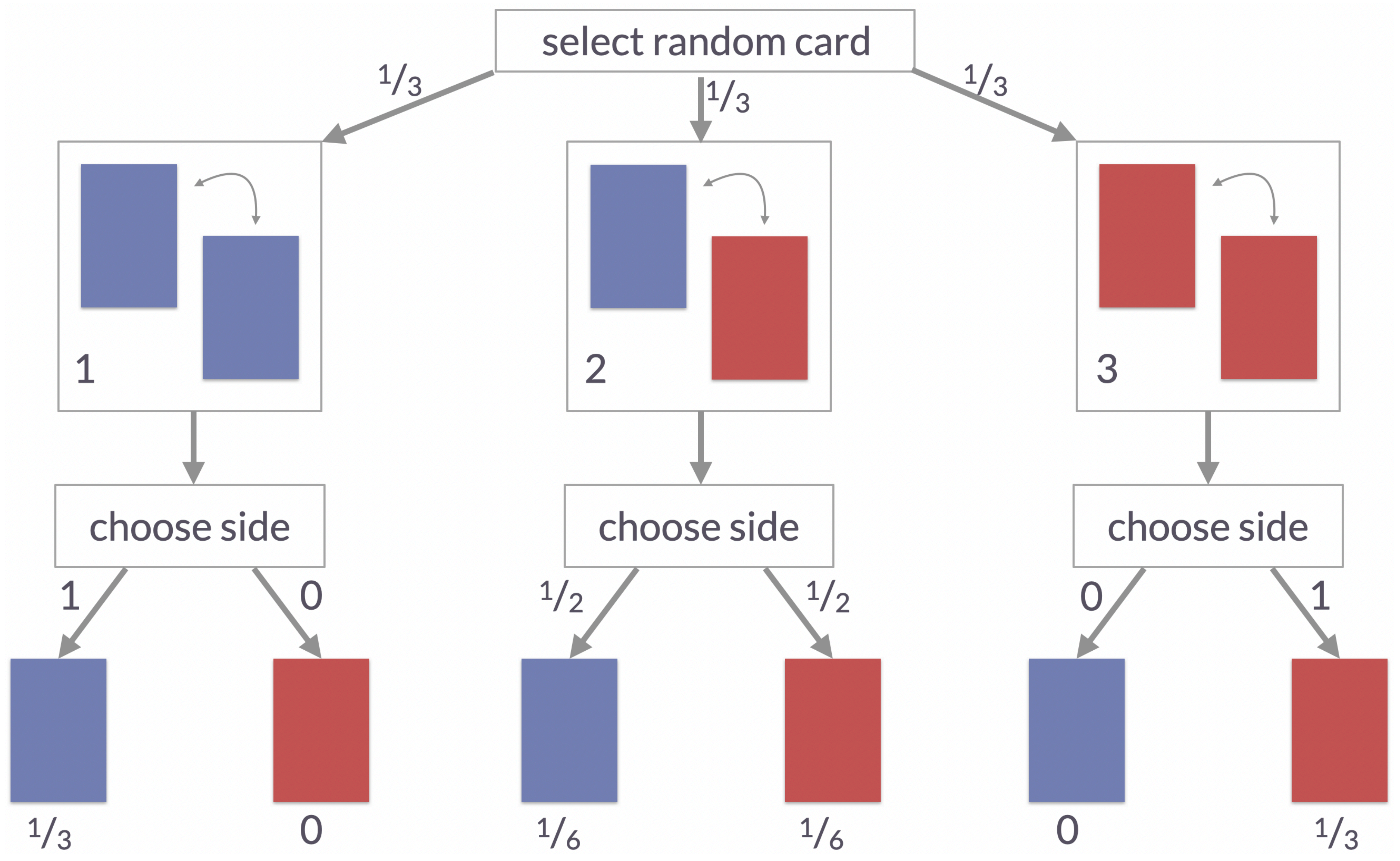
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Applied to three-card problem:

$$P(\text{card 1} \mid \text{blue}) = \frac{P(\text{blue} \mid \text{card 1})\ P(\text{card 1})}{P(\text{blue})}$$

$$= \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

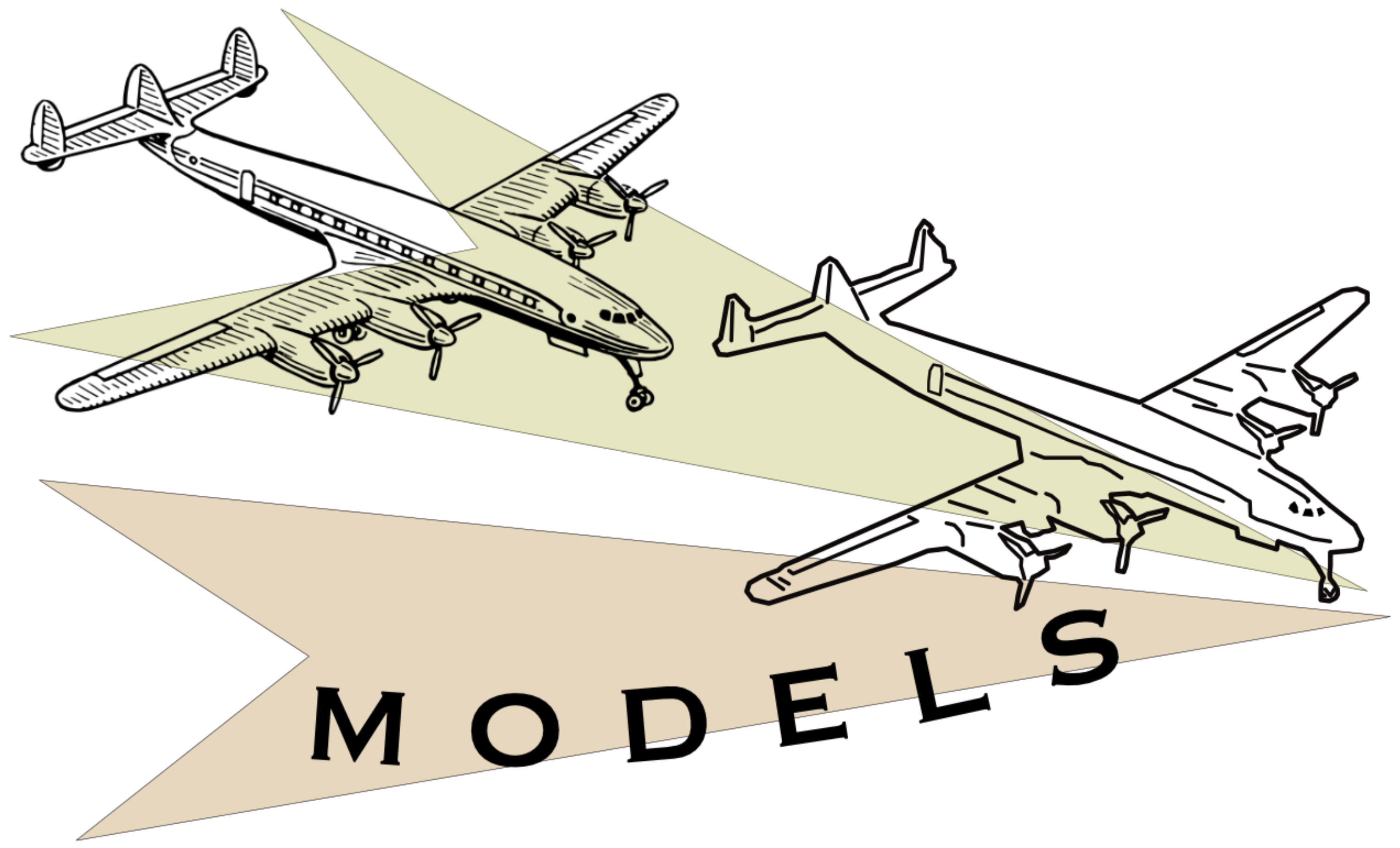
"reasoning from observed effect to latent cause via a model of the data-generating process"



Statistical models

likelihoods from a data-generating process

- A **statistical model** is a condensed formal representation, following common conventional practices of formalization, of the assumptions we make about what the data is and how it might have been generated by some (usually: stochastic) process.
- "All models are wrong, but some are useful." (Box 1979)
- a **Bayesian statistical model** of stochastic process generating data D consists of:
  - a vector of **parameters** θ
  - a **likelihood function**: $P(D \mid \theta)$
  - a **prior** distribution: $P(\theta)$
- among other things, we can use a model for **inference**:
  - **posterior** distribution: $P(\theta \mid D) \propto P(D \mid \theta)\ P(\theta)$



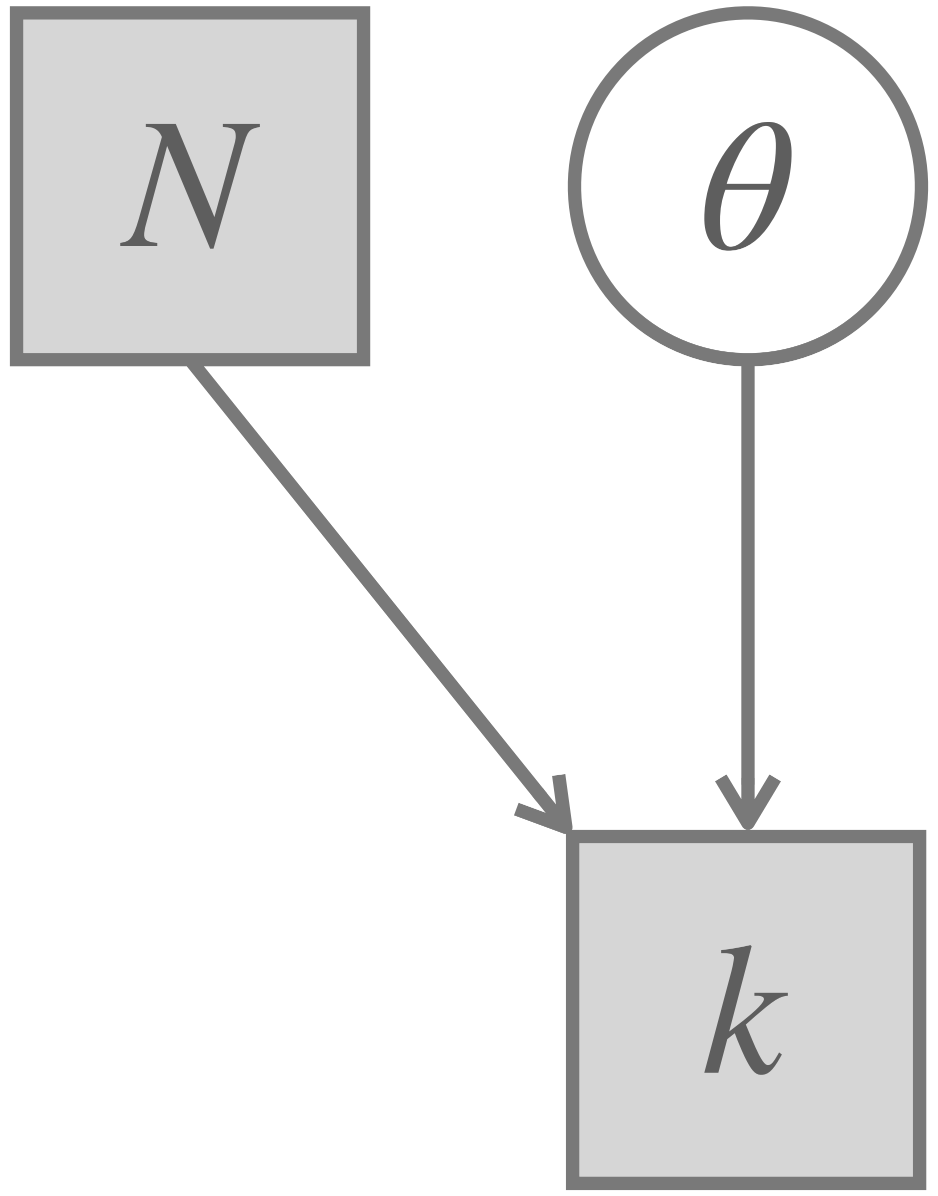
Binomial model

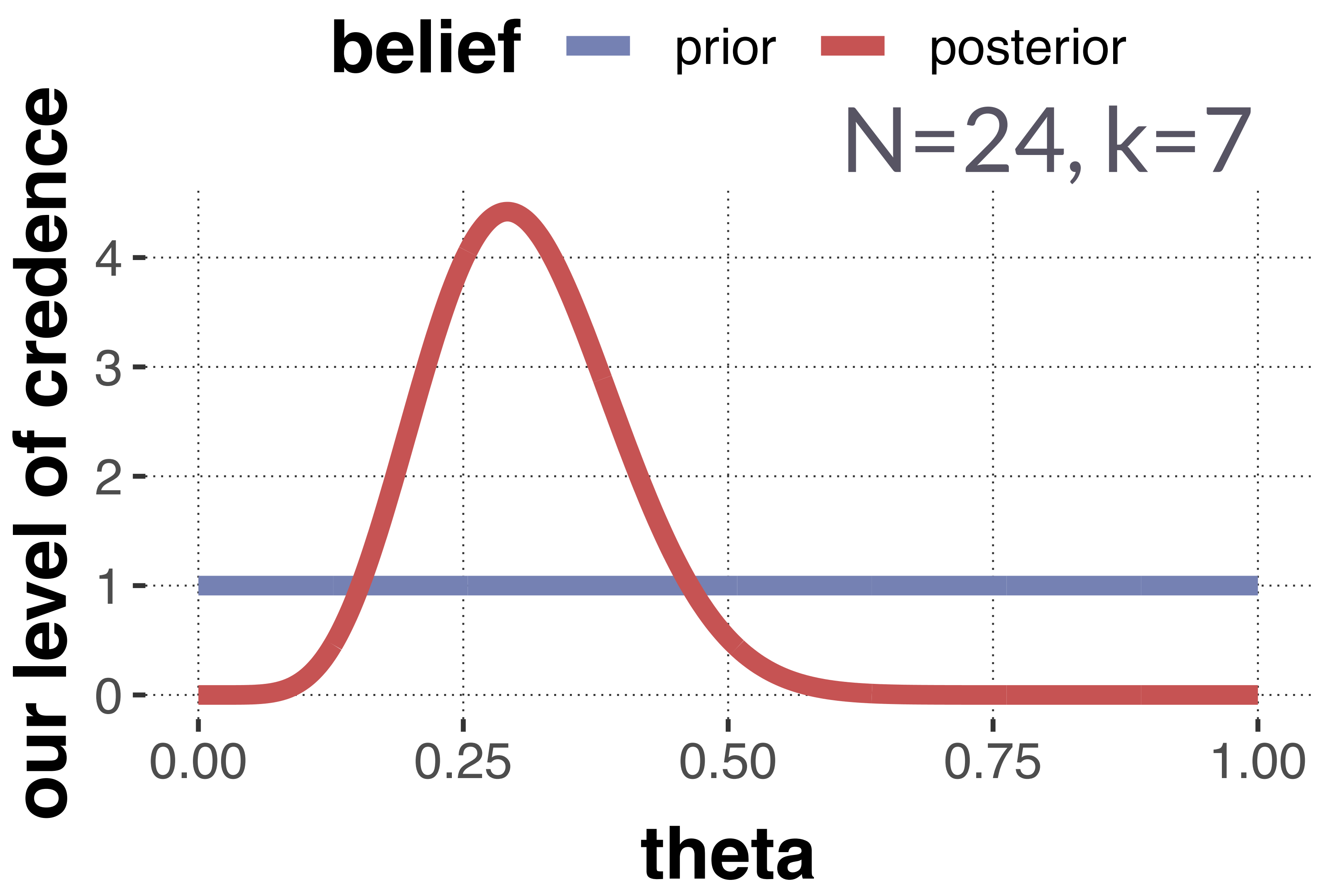
the 'coin-flip' model

- data: pair of numbers $D = \{k, N\}$
  - N is the number of tosses
  - k is the number of heads (successes)
- variable:
  - θ is the number of heads (successes)
- uninformed prior:

  $\theta \sim \text{Beta}(1,1)$
- likelihood function:

  $k \sim \text{Binomial}(\theta, N)$
- conventions for model graphs:
  - circles / squares: continuous / discrete variables
  - white / gray nodes: latent / observed variables





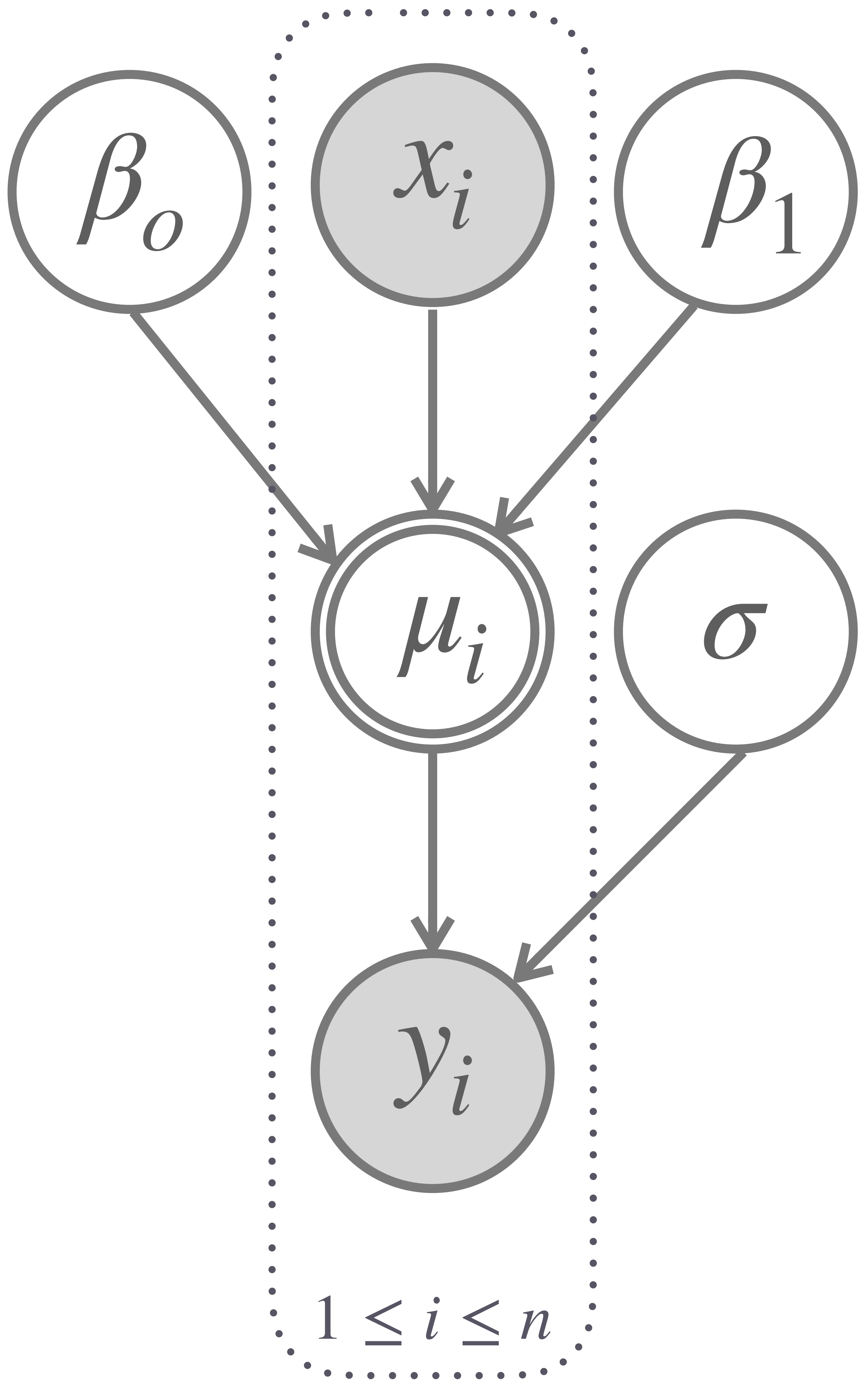
Simple linear regression model

for a single predictor variable

- data: n pairs of numbers $D = \{\langle x_1, y_1 \rangle, \dots \langle x_n, y_n \rangle\}$
  - x_i is the i-th observation of the **independent / predictor variable**
  - y_i is the i-th observation of the **dependent / to-be-predicted variable**
- parameters:
  - β_0 is the **intercept** parameter
  - β_1 is the **slope** parameter
  - σ is the standard deviation of a normal distribution
- derived variable: [shown in node w/ double lines]
  - μ_i is the linear predictor for observation i
- priors (uninformed):

  $\beta_0, \beta_1 \sim \text{Uniform}(-\infty, \infty)$ $\qquad$ $\log(\sigma^2) \sim \text{Uniform}(-\infty, \infty)$
- likelihood:

  $y_i \sim \text{Normal}(\mu_i, \sigma)$ $\qquad$ $\mu_i = \beta_0 + x_1 \cdot \beta_1$

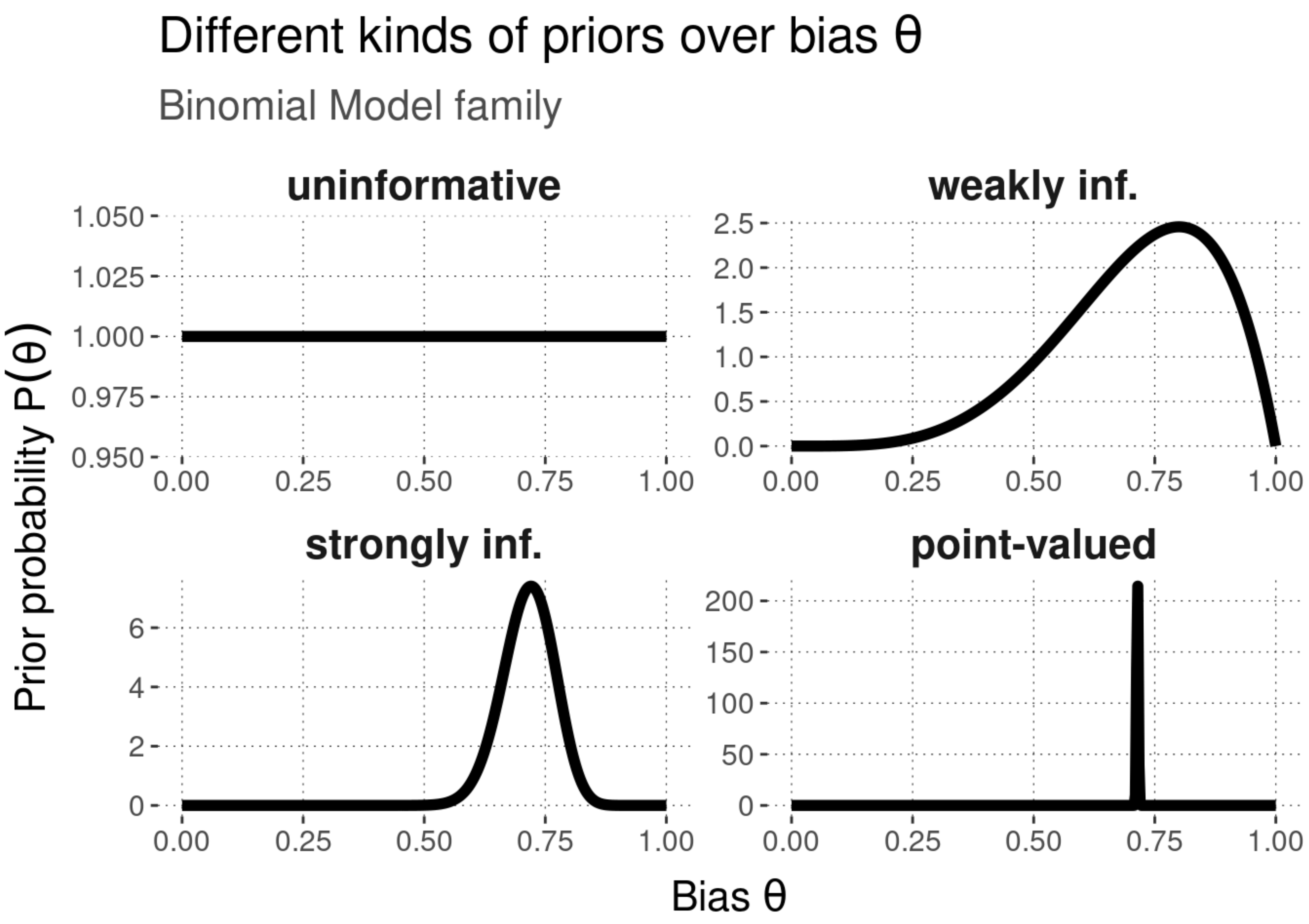


Likelihood, prior & posterior

for the coin-flip model

Kinds of priors

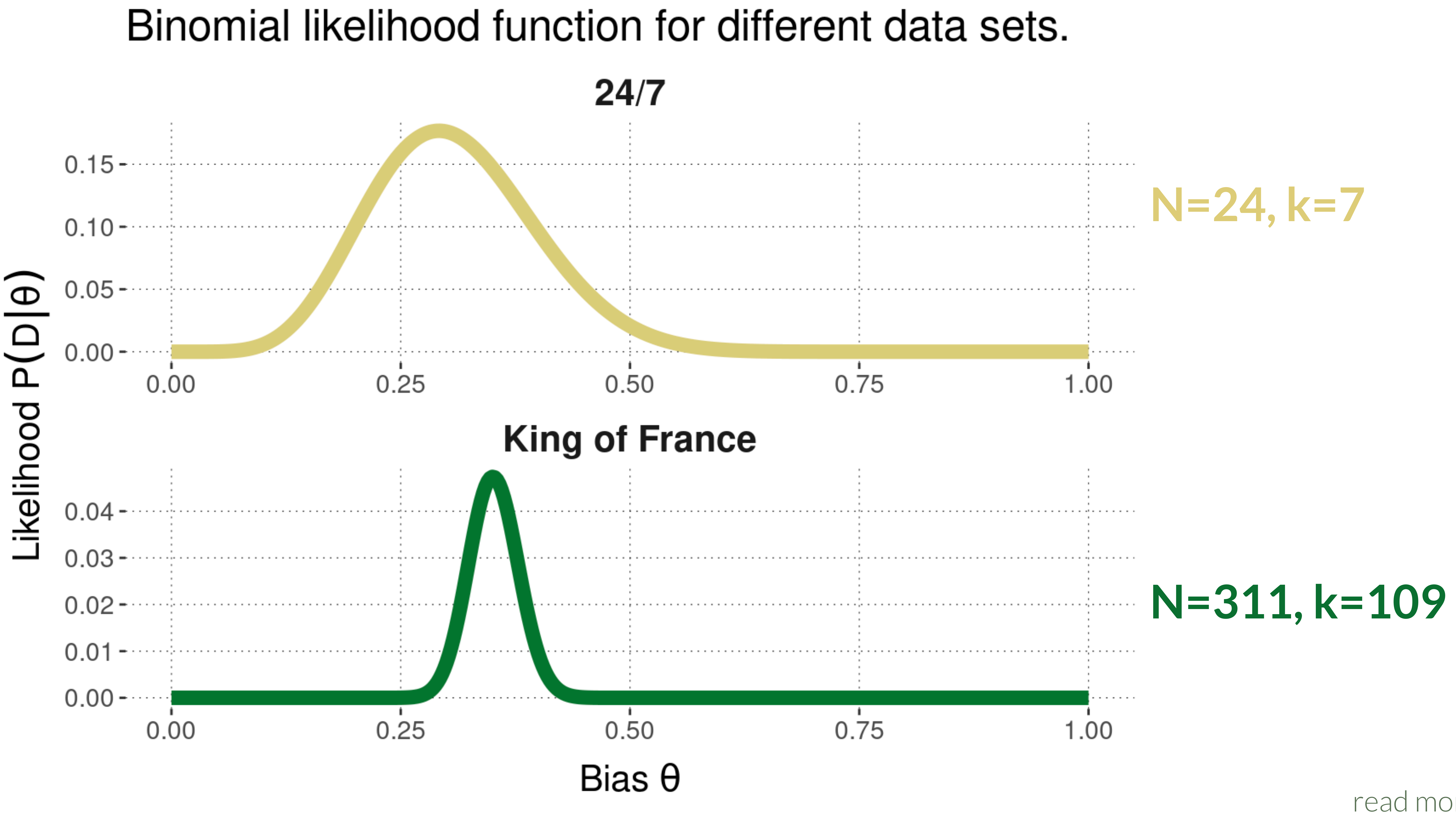
for a Binomial ('coin flip') model



read more here

Binomial likelihoods

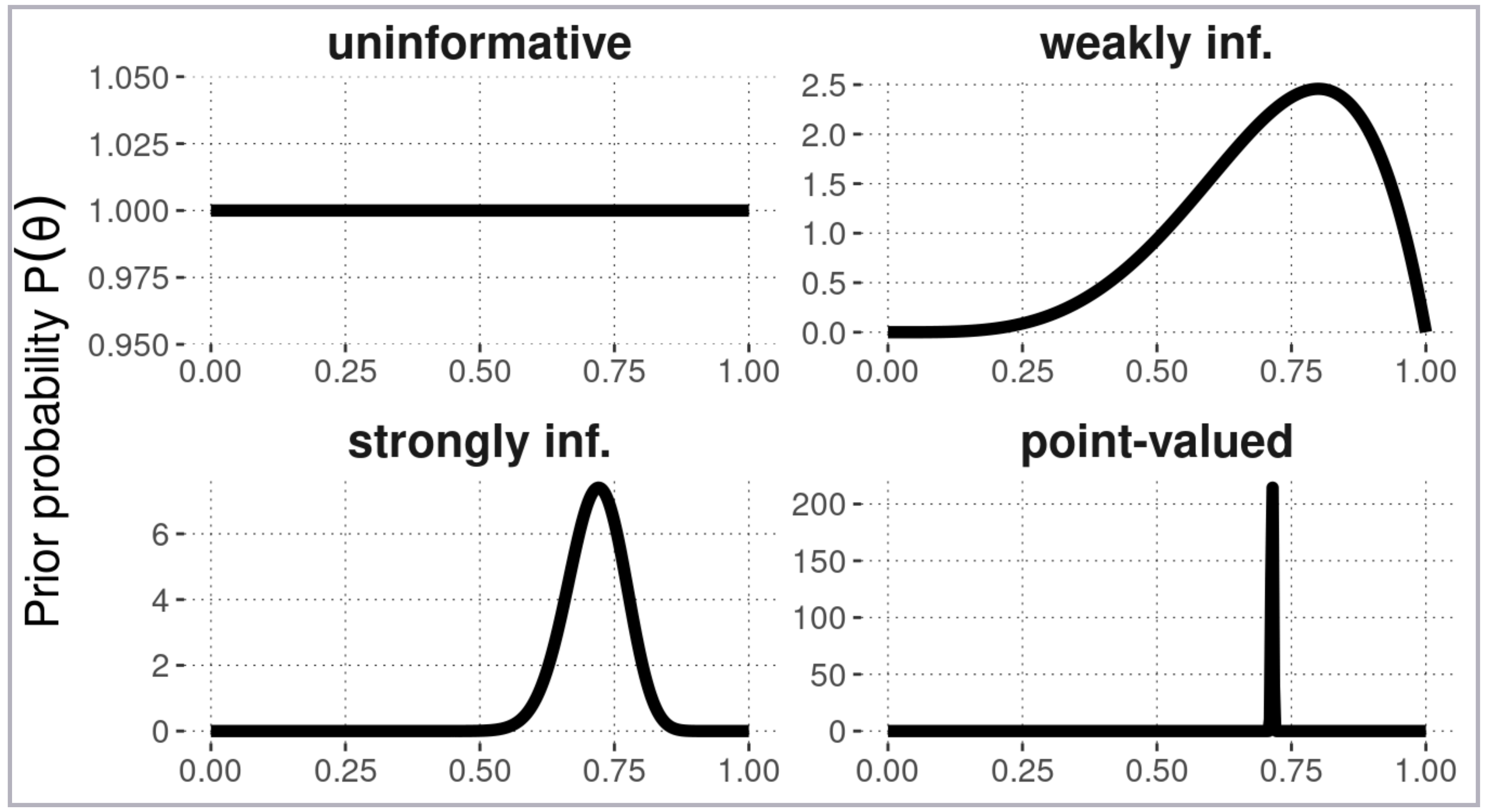
two data sets

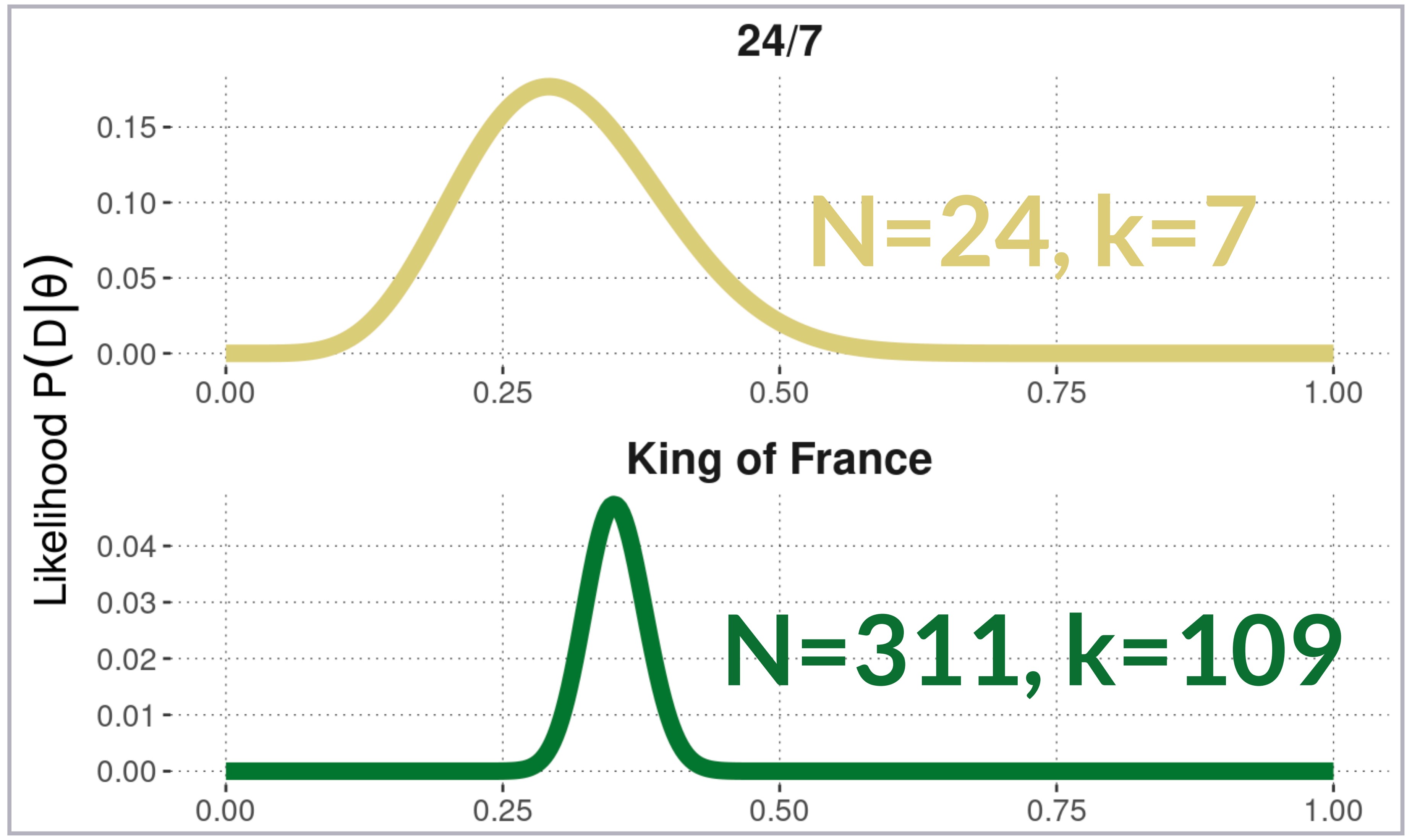


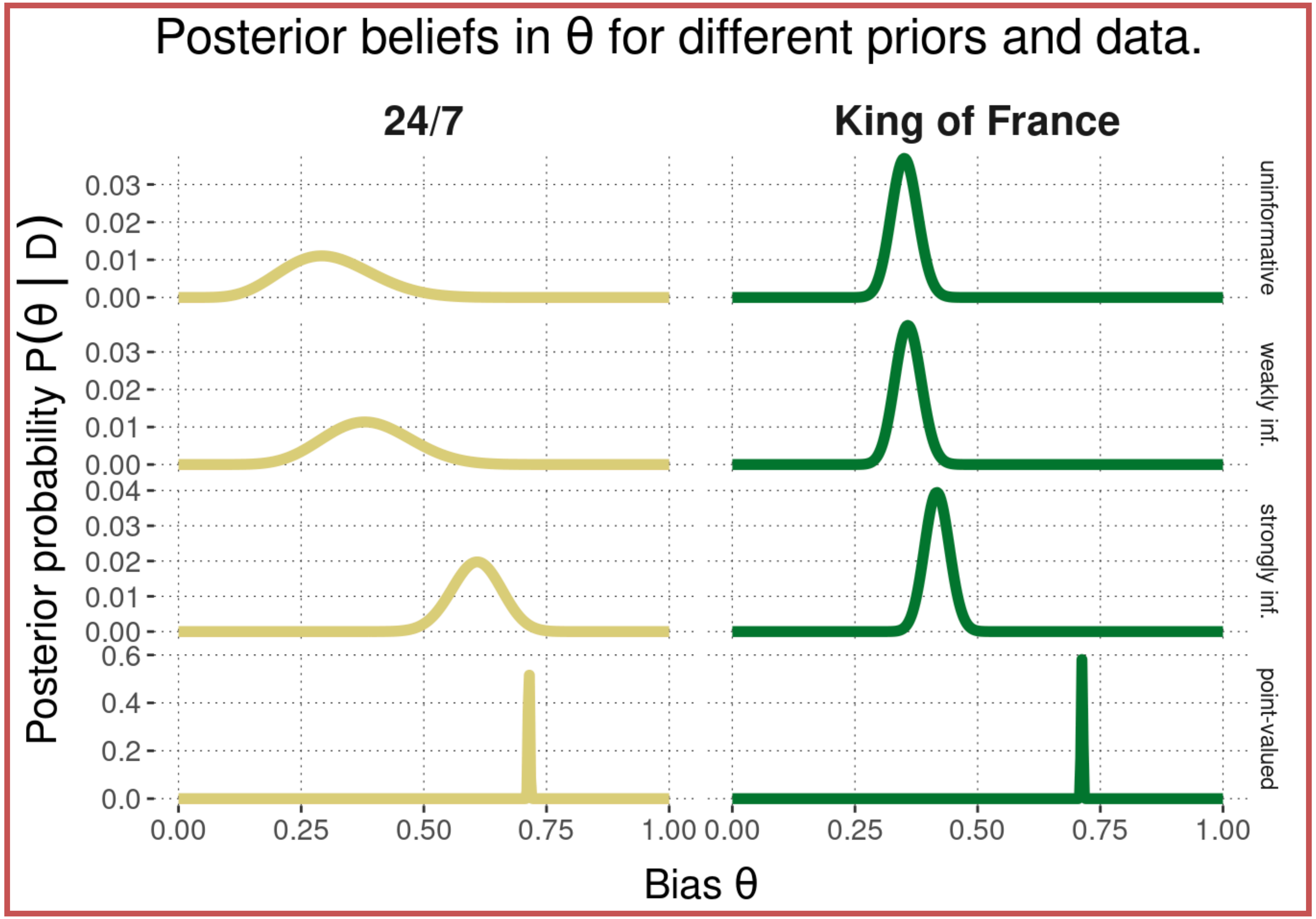
read more here

Posterior distributions

for different priors and likelihoods







read more here

Computing posterior distributions

problem of computational complexity

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

✓fast & easy

✓fast & easy

❌☠ possibly intractable ☠❌

read more here

Posteriors from conjugacy

closed-form posteriors from clever choice of priors

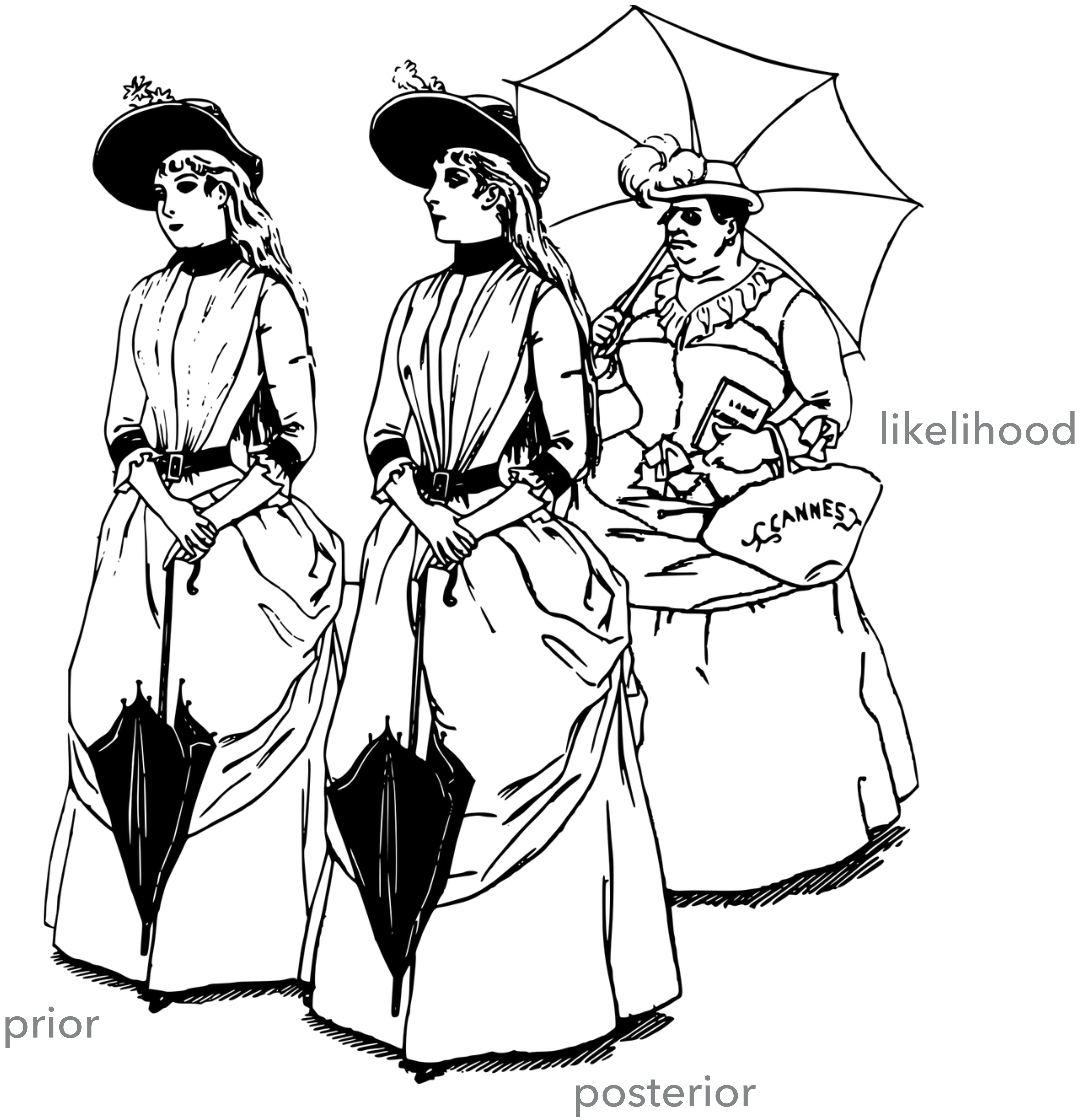
- prior $P(\theta)$ is a **conjugate prior** for likelihood $P(D \mid \theta)$ iff prior $P(\theta)$ and posterior $P(\theta \mid D)$ are the same kind of probability distribution, e.g.:
  - prior: $\theta \sim \text{Beta}(1,1)$
  - posterior: $\theta \mid D \sim \text{Beta}(8,18)$
- **claim:** the beta distribution is a conjugate prior for the binomial likelihood function
  - proof:

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

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

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

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

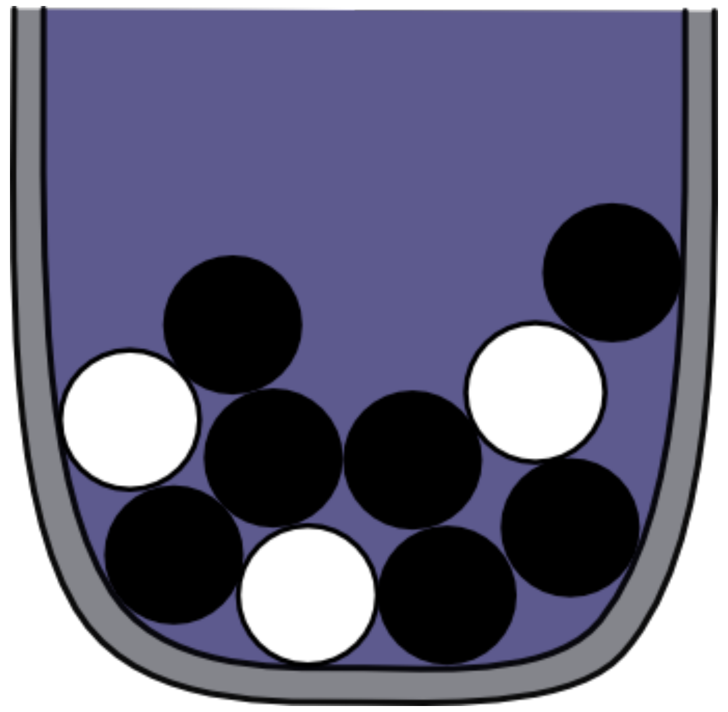


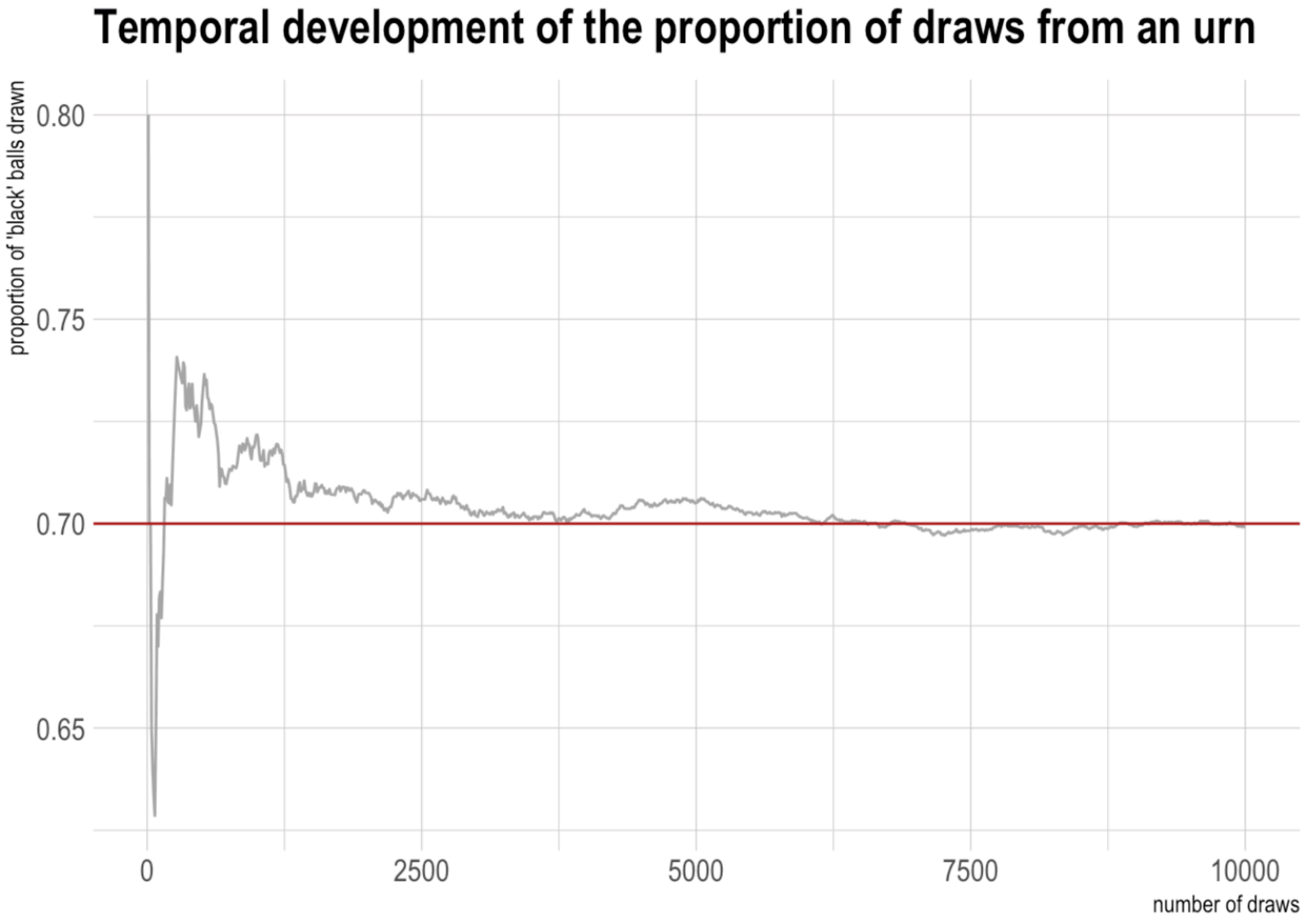
read more here

Approximating distributions via sampling

our go-to solution for approximating posterior distributions beyond conjugacy

- we can approximate any probability distribution by either:
  - a large set of representative samples; or
  - an oracle that returns a sample if needed.

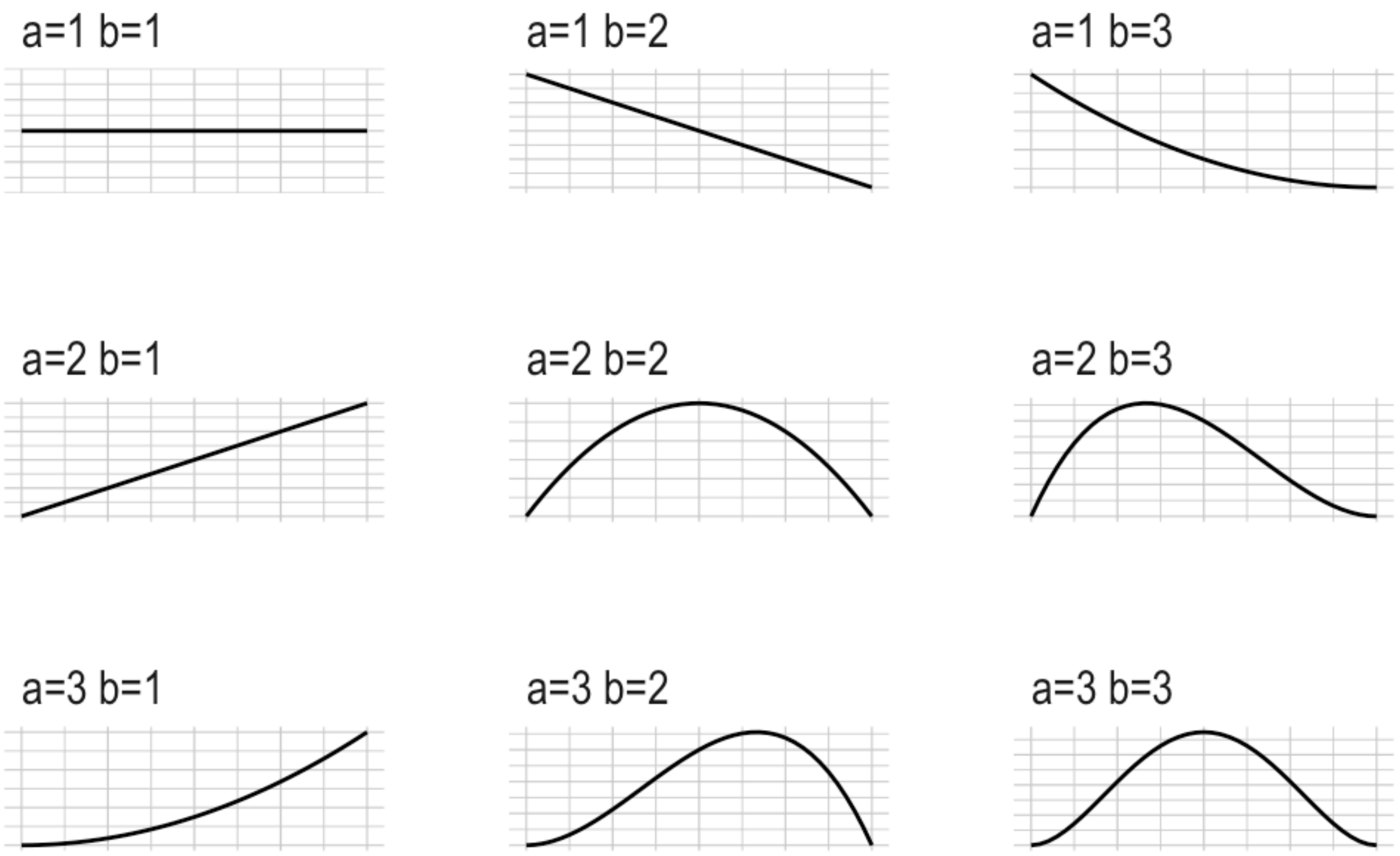




Bayesian parameter estimation

Sequential updating
for the beta-binomial model

- sequence of updating does not matter
  - any order of single-observation updates
  - any 'chunking': whole data set, different subsets in whatever sequence (as long as disjoined)
- "today's posterior is tomorrow's prior"



read more here

Sequential updating

general proof

- **claim:** if $\{D_1, D_2\}$ is a partition of D, then $P(\theta \mid D) \propto P(\theta \mid D_1)\ P(D_2 \mid \theta)$
- **sketch of proof:**

$$
\begin{aligned}
P(\theta \mid D) &= \frac{P(\theta)\ P(D \mid \theta)}{\int P(\theta')\ P(D \mid \theta')\mathrm{d}\theta'} \\
&= \frac{P(\theta)\ P(D_1 \mid \theta)\ P(D_2 \mid \theta)}{\int P(\theta')\ P(D_1 \mid \theta')\ P(D_2 \mid \theta')\mathrm{d}\theta'} && \text{[from multiplicativity of likelihood]} \\
&= \frac{P(\theta)\ P(D_1 \mid \theta)\ P(D_2 \mid \theta)}{\frac{k}{k}\int P(\theta')\ P(D_1 \mid \theta')\ P(D_2 \mid \theta')\mathrm{d}\theta'} && \text{[for random positive k]} \\
&= \frac{\frac{P(\theta)\ P(D_1 \mid \theta)}{k}\ P(D_2 \mid \theta)}{\int \frac{P(\theta')\ P(D_1 \mid \theta')}{k}\ P(D_2 \mid \theta')\mathrm{d}\theta'} && \text{[rules of integration; basic calculus]} \\
&= \frac{P(\theta \mid D_1)\ P(D_2 \mid \theta)}{\int P(\theta' \mid D_1)\ P(D_2 \mid \theta')\mathrm{d}\theta'} && \text{[Bayes rule with } k = \int P(\theta)P(D_1 \mid \theta)\mathrm{d}\theta\text{]}
\end{aligned}
$$

read more here

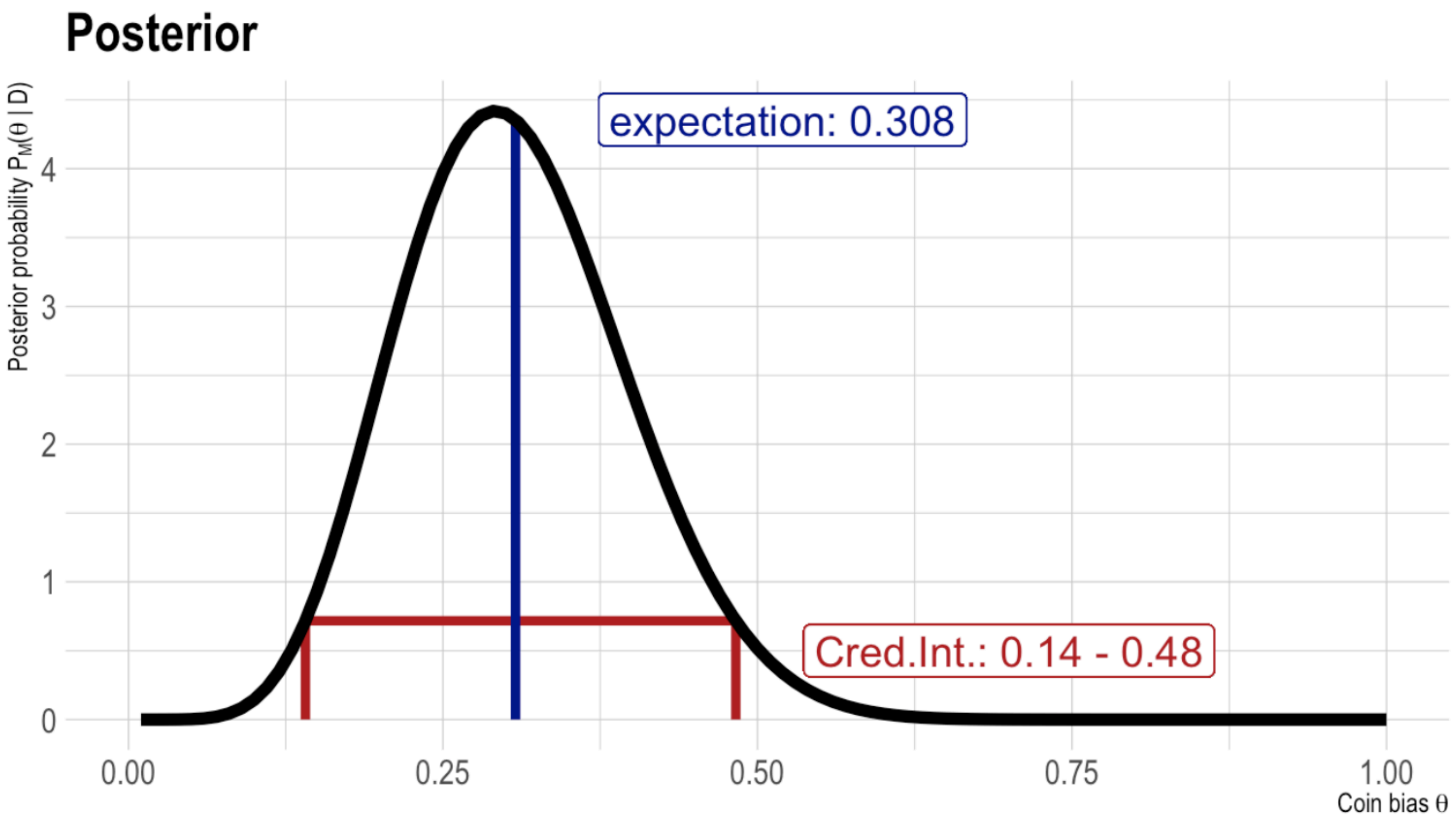
Parameter estimation

point- and interval-valued estimates

- Bayes' rule for parameter estimation:

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

- common point estimates ("best" values):
  - maximum likelihood estimate (MLE)
  - maximum a posteriori (MAP)
  - posterior mean / expected value
- common interval estimates (range of "good" values):
  - confidence intervals
  - credible intervals



read more here

Point-valued estimates

MLE, MAP and (posterior) expected value

- **MLE:**

  $$\arg\max_\theta P(D \mid \theta)$$

  - doesn't take prior into account (not Bayesian)
  - not necessarily unique

- **MAP:**

  $$\arg\max_\theta P(\theta \mid D)$$

  - local / does not consider full distribution (not fully Bayesian)
  - increasingly uninformative in larger parameter spaces
  - not necessarily unique

- **posterior mean / expected valued**

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

  - holistic / depends on full distribution ("genuinely Bayesian")
  - always unique (for proper priors/posteriors)

read more here

Bayesian hypothesis testing /w posterior credible intervals

!!! caveat: it is controversial whether this is the best (Bayesian) approach to hypothesis testing !!!

- consider an interval-based hypothesis: $\theta \in I$
  - e.g., inequality-based: "coin is biased towards heads" $\theta > 0.5$
  - e.g. a **region of practical equivalence [ROPE]:** an ϵ-region around some θ^*: $I = [\theta^* - \epsilon, \theta^* + \epsilon]$
- if $[l; u]$ is a posterior credible interval for θ, we consider this:
  - **reason to accept** hypothesis I if $[l; u]$ is contained entirely in I;
  - **reason to reject** hypothesis I if $[l; u]$ and I have no overlap;
  - **withhold judgement** otherwise.
- this approach is "categorical" (accept, reject, suspend) and not quantitative

read more here

Posterior plausibility of interval-based hypotheses

this is NOT a testing approach, just one way of quantifying support

- consider an interval-based hypothesis $\theta \in I$ as before
- the **posterior plausibility** of I given a model M and the data D is just the posterior probability: $P(\theta \in I \mid D)$
- **not** a notion of observational evidence:
  - if prior is high for I and data is uninformative, posterior plausibility can be high
- good-enough first heuristic when priors are "unbiased" regarding I
- more on hypothesis testing later

read more here

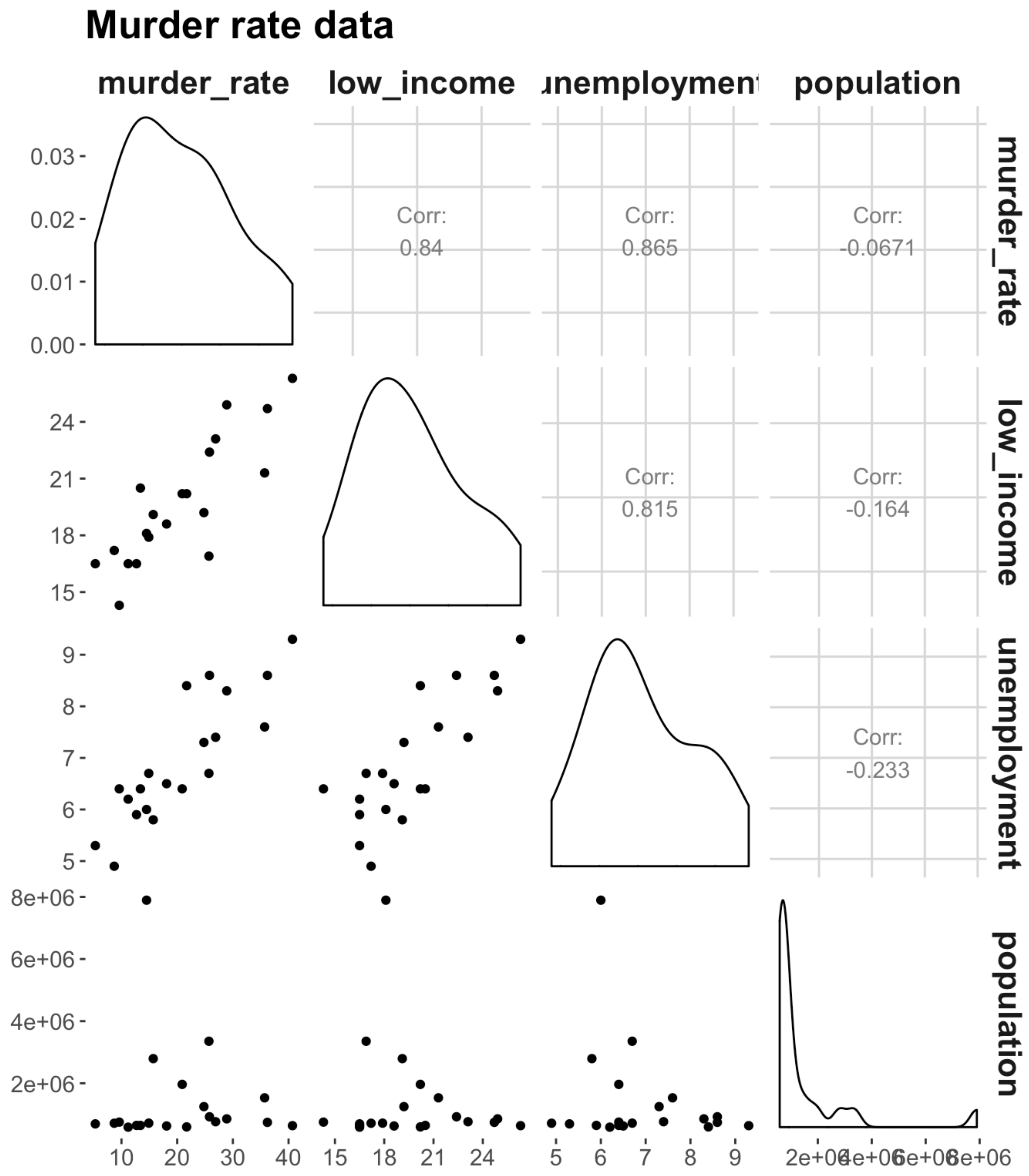
Simple linear regression

likelihood & Bayesian posterior

Murder data

annual murder rate, average income, unemployment rates and population

```
## # A tibble: 20 x 4
##    murder_rate low_income unemployment population
##          <dbl>      <dbl>        <dbl>      <dbl>
##  1        11.2       16.5          6.2     587000
##  2        13.4       20.5          6.4     643000
##  3        40.7       26.3          9.3     635000
##  4         5.3       16.5          5.3     692000
##  5        24.8       19.2          7.3    1248000
##  6        12.7       16.5          5.9     643000
##  7        20.9       20.2          6.4    1964000
##  8        35.7       21.3          7.6    1531000
##  9         8.7       17.2          4.9     713000
## 10         9.6       14.3          6.4     749000
## 11        14.5       18.1          6      7895000
## 12        26.9       23.1          7.4     762000
## 13        15.7       19.1          5.8    2793000
## 14        36.2       24.7          8.6     741000
## 15        18.1       18.6          6.5     625000
## 16        28.9       24.9          8.3     854000
## 17        14.9       17.9          6.7     716000
## 18        25.8       22.4          8.6     921000
## 19        21.7       20.2          8.4     595000
## 20        25.7       16.9          6.7    3353000
```



annual murders per million inhabitants

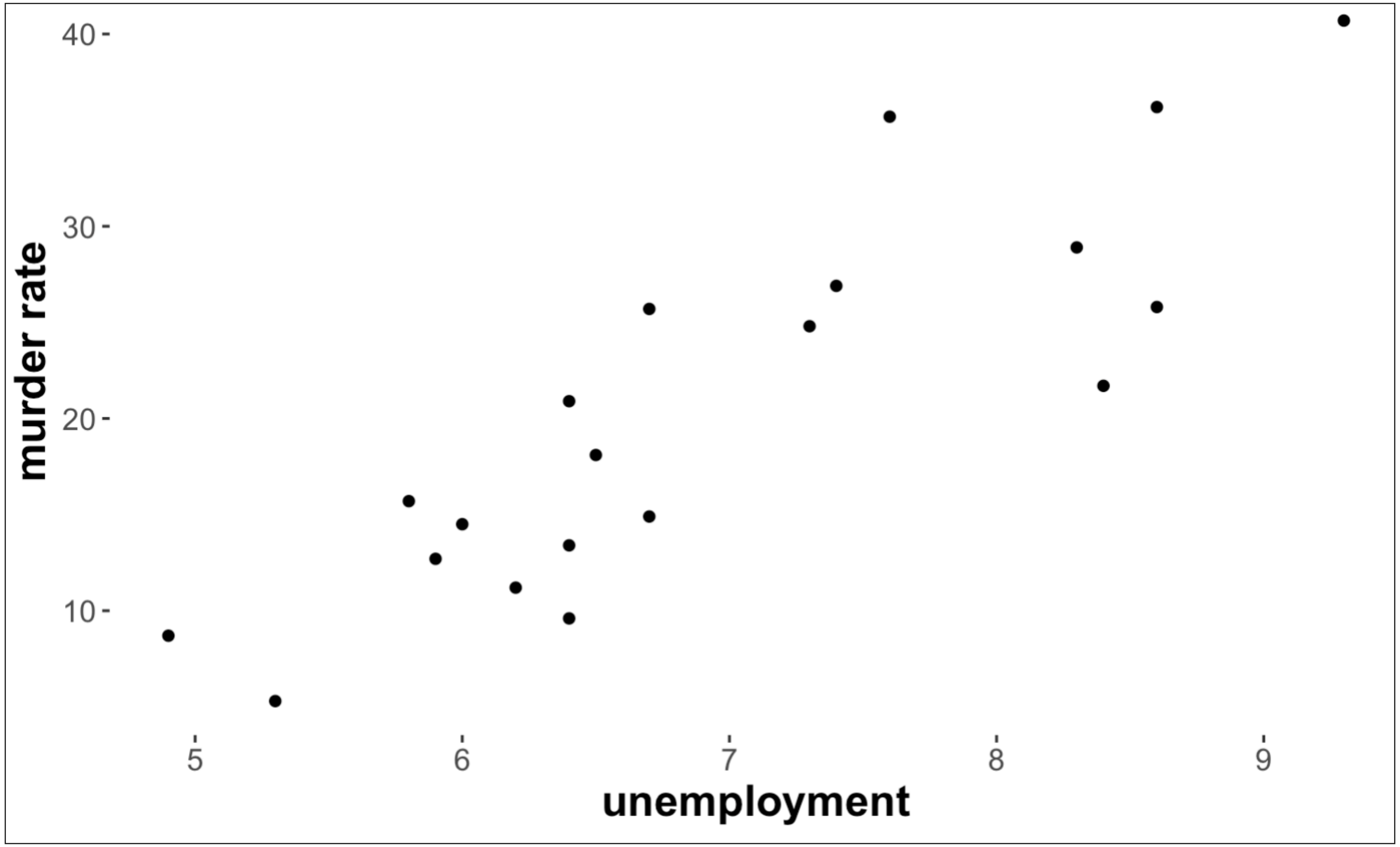
percentage inhabitants with low income

percentage inhabitants who are unemployed

total population

Predicting murder rate based on unemployment rate

some wild linear guessing



We are to predict the murder rate y_i of a randomly drawn city i. **We know that city's unemployment rate, x_i, but nothing more.**

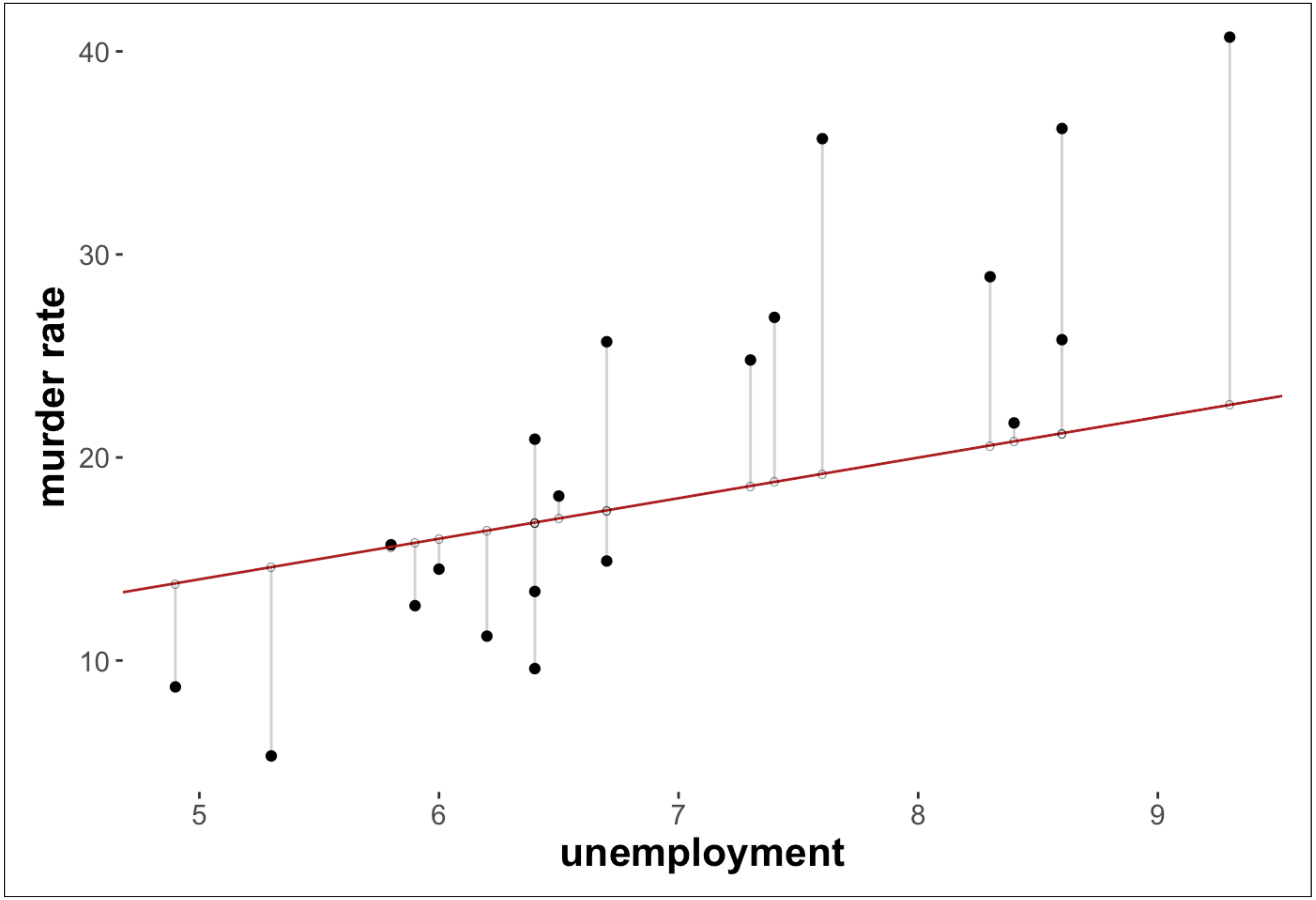
Let's just assume the following **linear relationship** to make a prediction b/c why not?!?

$$\hat{y}_i = 4 + 2x_i$$

How good is this prediction?

How good is any given prediction?

quantifying distance or likelihood



Distance-based approach

Residual Sum-of-Squares:

$$\text{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

- no predictions about spread around linear predictor

Likelihood-based approach:

Normal likelihood:

$$\text{LH} = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mu = \hat{y}_i, \sigma)$$

- fully predictive

Likelihood-based simple linear regression

- likelihood:

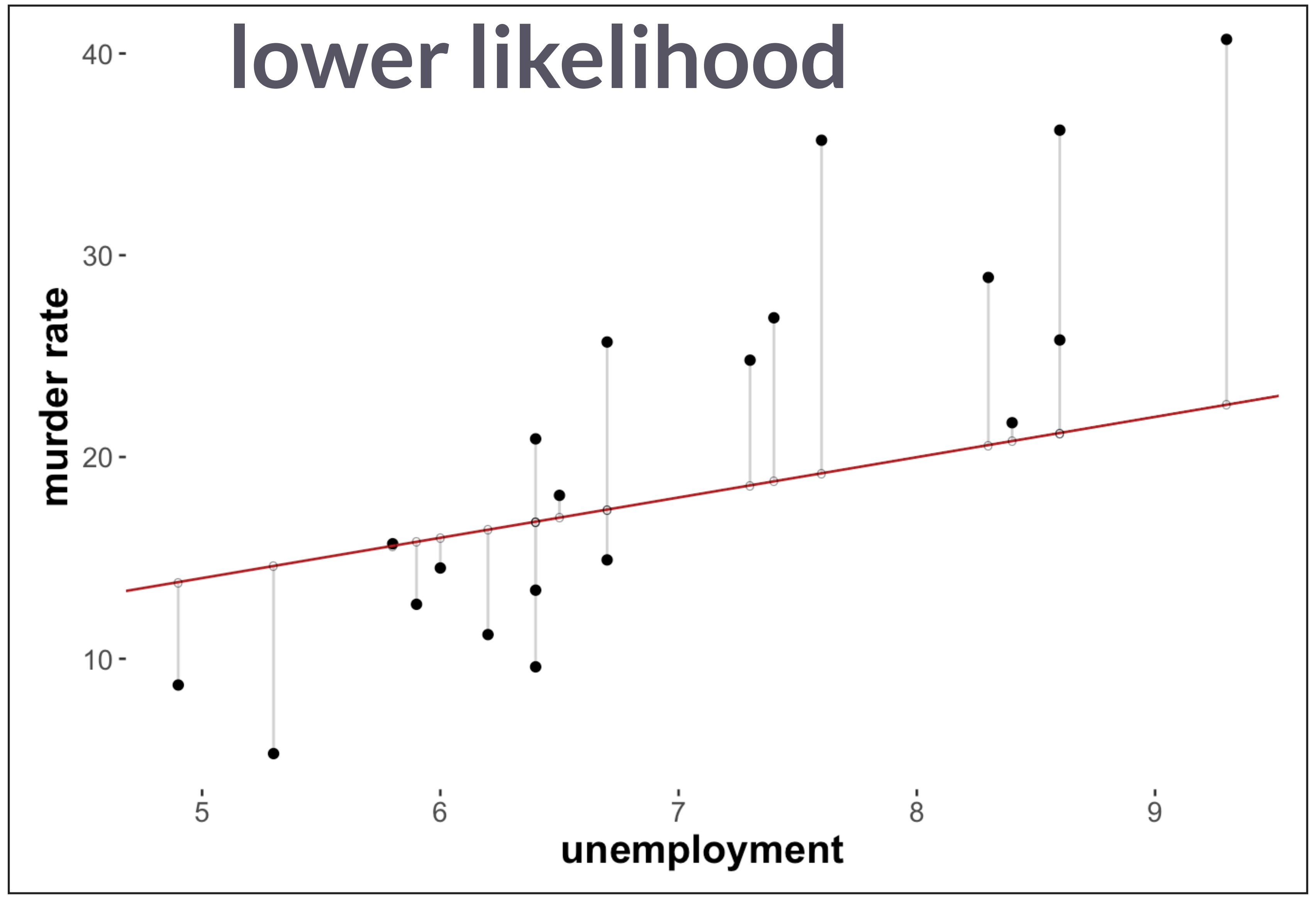
  $$y_i \sim \text{Normal}(\mu_i, \sigma)$$
  $$\mu_i = \beta_0 + x_1 \cdot \beta_1$$

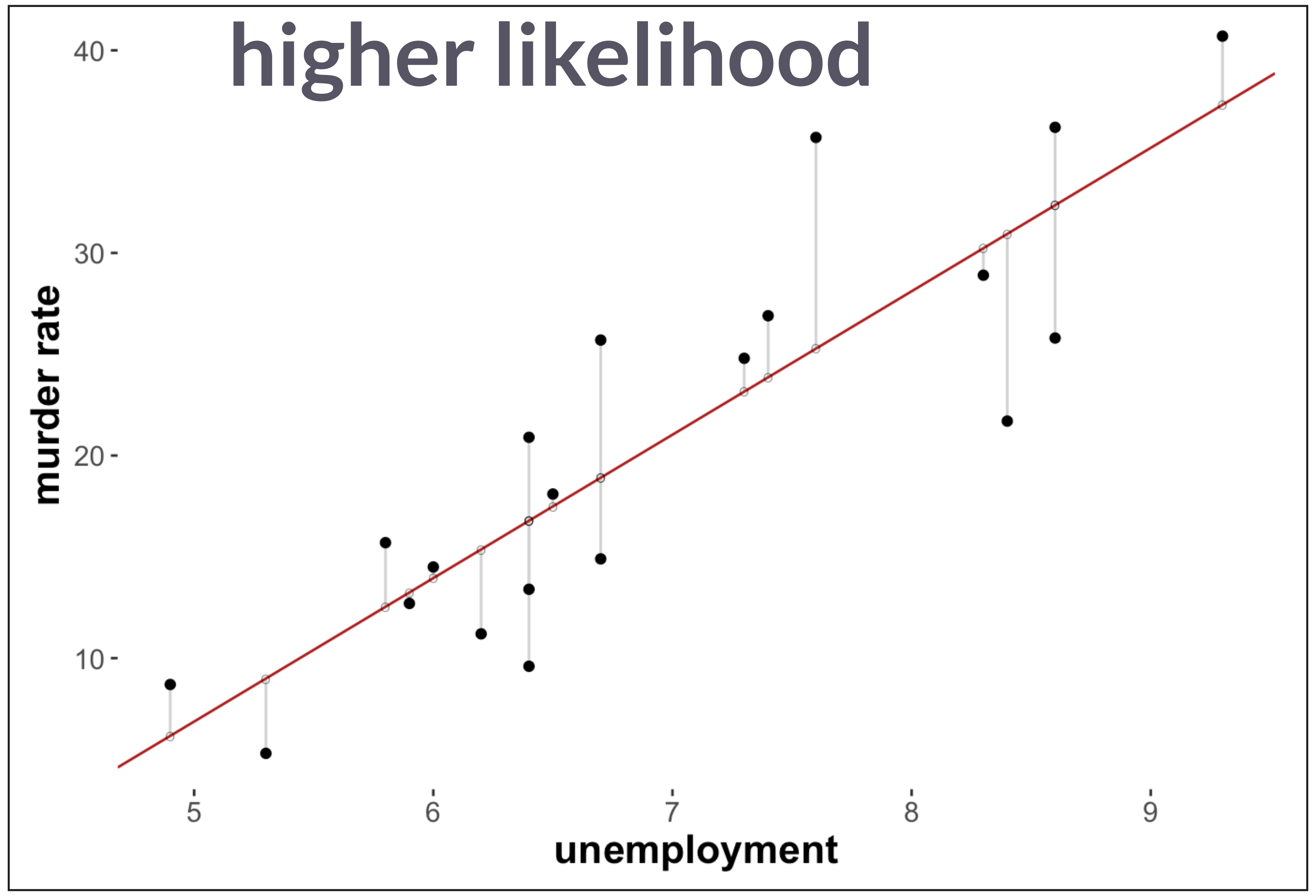
- differential likelihood:
  - parameter triples $\langle \beta_0, \beta_1, \sigma \rangle$ can be better or worse
  - higher vs. lower likelihood $P(D \mid \beta_0, \beta_1, \sigma)$
- maximum-likelihood solution:

  $$\arg\max_{\beta_0, \beta_1, \sigma} P(D \mid \beta_0, \beta_1, \sigma)$$

  - standard (frequentist) solution
  - MLE corresponds to MAP for "flat" priors
- Bayesian approach: full posterior distribution

  $$P(\beta_0, \beta_1, \sigma \mid D) \propto P(D \mid \beta_0, \beta_1, \sigma)\ P(\beta_0, \beta_1, \sigma)$$





Bayesian linear regression in R

using BRMS and Stan

- R package BRMS provides high-level interface for Bayesian linear regression
- models are specified with R's formula syntax
- returns samples from the posterior distribution
  - alternatives: MAPs, variational inference
- runs probabilistic programming language Stan in the background
  - powerful, cutting-edge tool for Bayesian computation
  - strong, non-commercial development team
  - many interfaces: stand-alone, R, Python, Julia, ...

```
fit_brms_murder <- brm(
  # specify what to explain in terms of what
  #  using the formula syntax
  formula = murder_rate ~ unemployment,
  # which data to use
  data = murder_data
)
```

Stan

Navigating BRMS output

```r
summary(fit_brms_murder)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: murder_rate ~ unemployment
   Data: murder_data (Number of observations: 20)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      -28.48      7.32   -42.05   -13.79 1.00     3014     2362
unemployment     7.07      1.04     4.97     9.04 1.00     2978     2451

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     5.42      0.96     3.88     7.63 1.00     2664     2196

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

model & data we used

information about sampling
(more on this later)

main model parameters

additional model parameters

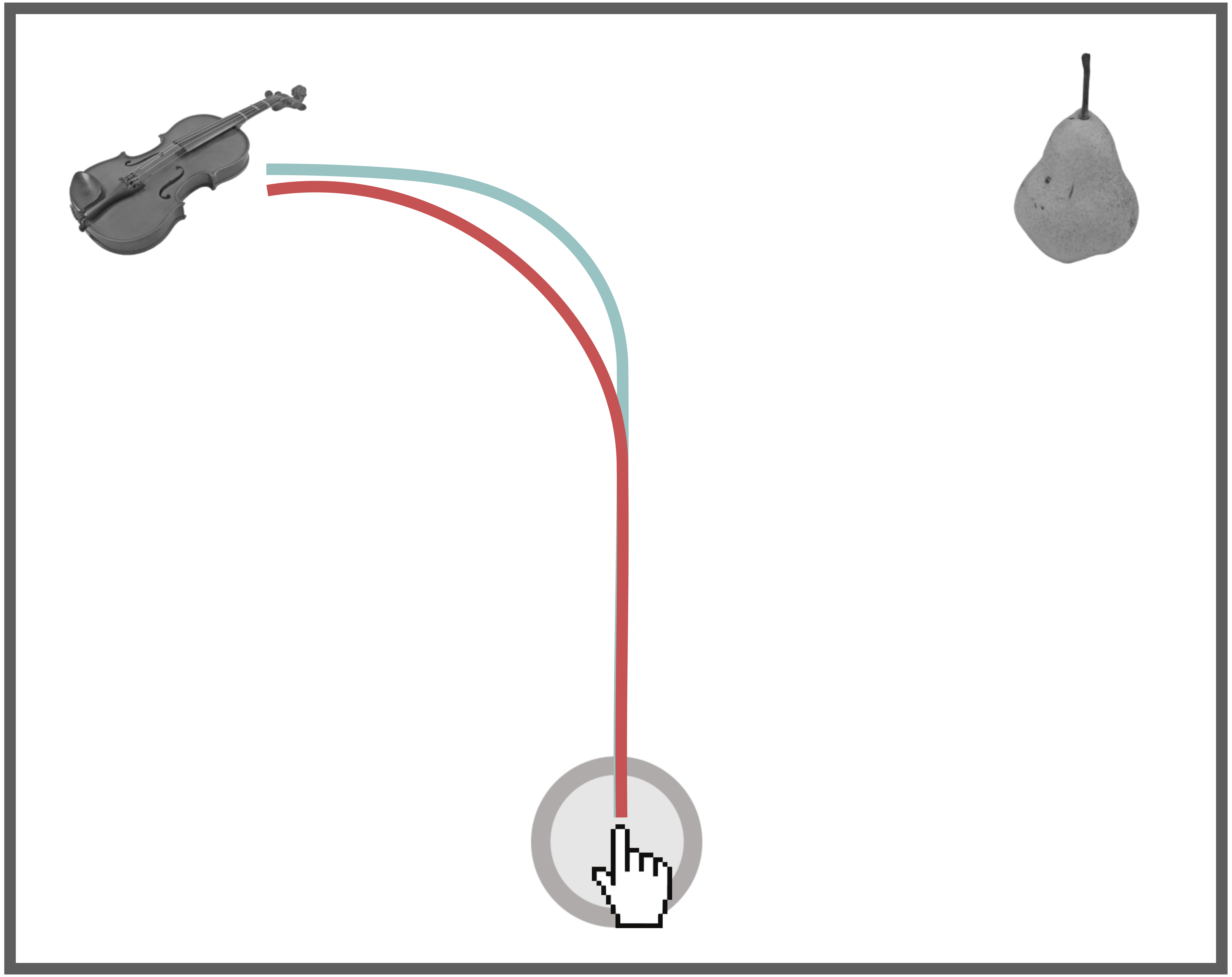
information about sampling
(more on this later)

Mouse-tracking data on typicality in category decisions

Mouse-tracking

Hand-movement during decision making

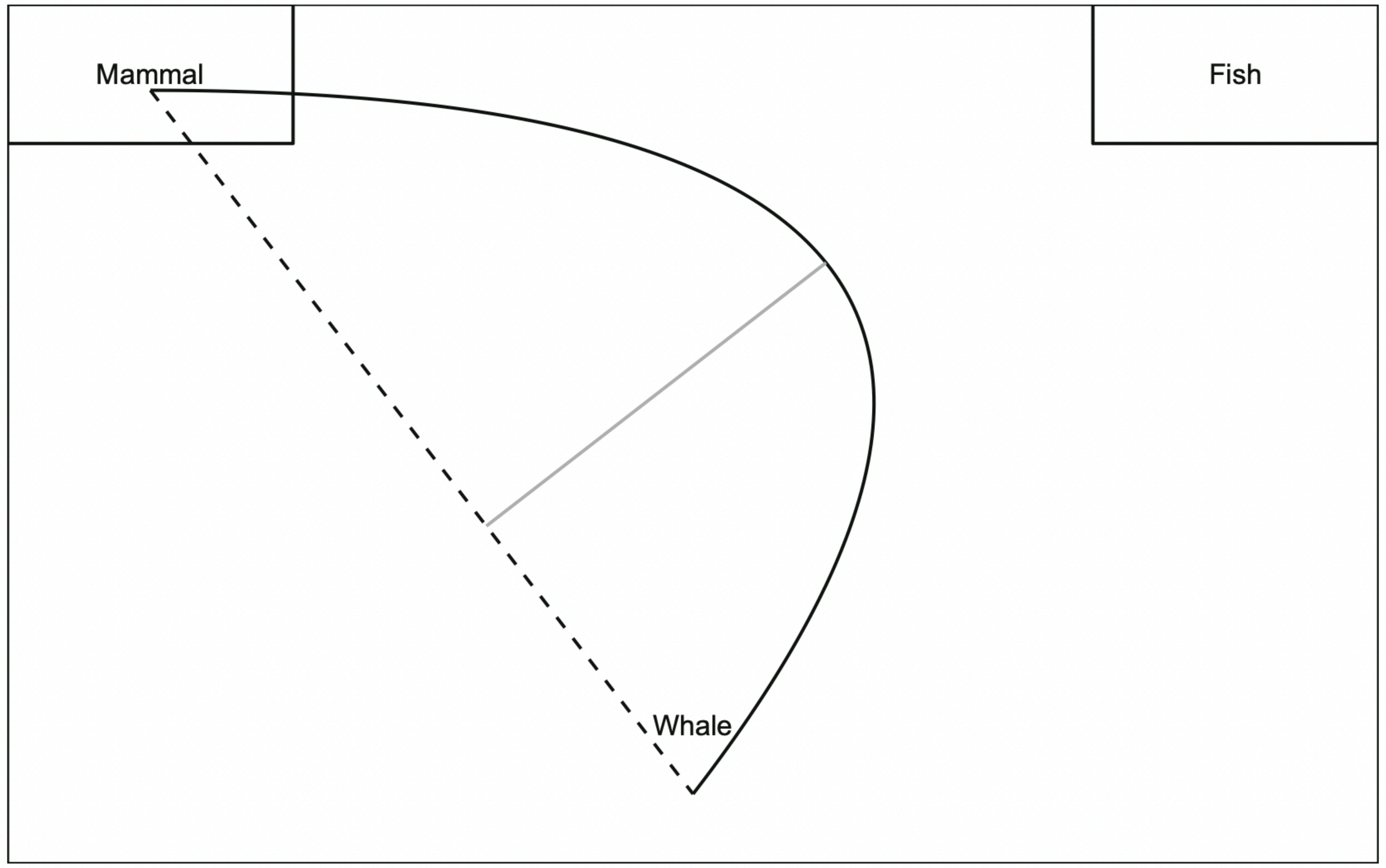
- general idea: motor-execution provides information about the ongoing decision process
  - uncertainty
  - gradual evidence accumulation
  - change-of-mind
  - time-point of decision
  - ...
- many subtle design decisions
  - click vs touch
  - move horizontally or vertically
  - ...

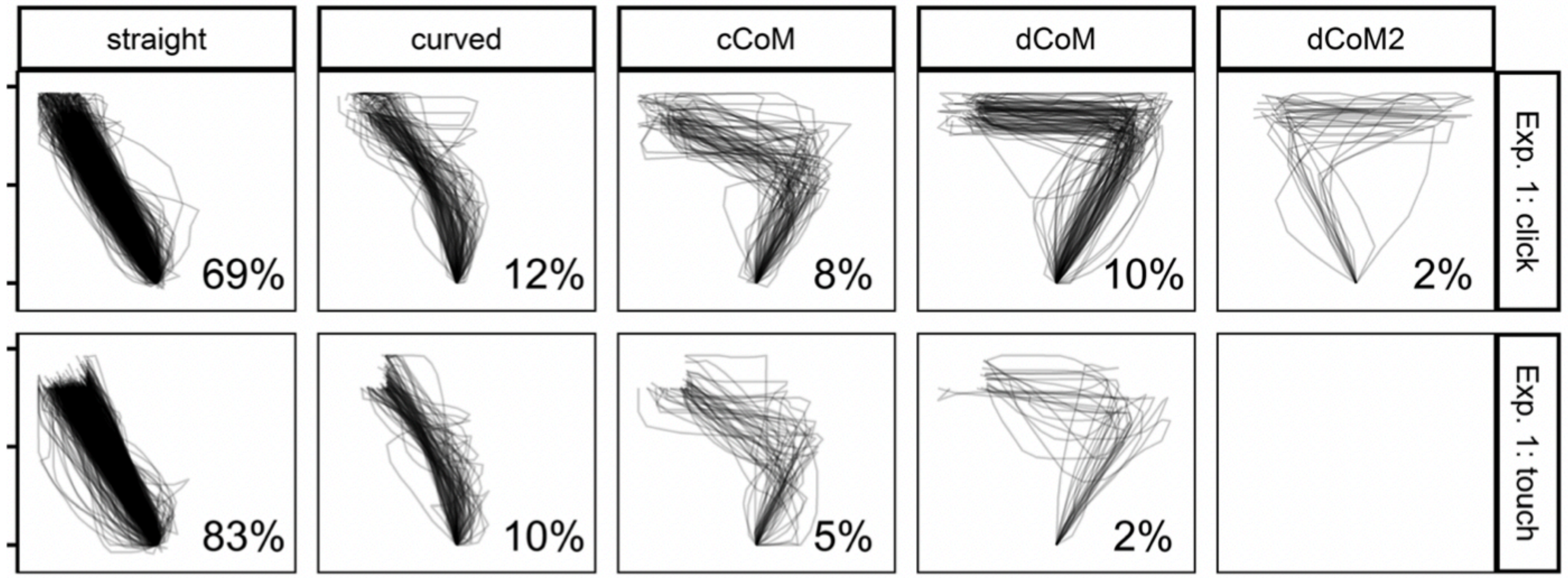


Mouse-tracking

common measures of mouse-trajectories

- raw data are lists of triples
  - (time, x-position, y-position)
- commonly used measures
  - area-under the curve (AUC)
    - area between the mouse trajectory and a straight line from start to selected option
  - maximal deviation (MAD)
    - maximum distance between trajectory and straight line from start to selected option
  - correctness
    - whether choice of option was correct or not
  - reaction time (RT)
    - how long did the movement last in total
  - type of trajectory
    - result of clustering analysis based on shape of the trajectories (usually some 3-5 categories)
  - x-flips
    - number of times the trajectory crossed the vertical middle line (at x = 0)





Running example

category recognition for typical vs atypical exemplars

- materials & procedure
  - participants read an animal name (e.g. 'dolphin')
  - they choose the true category the animal belongs to (e.g., 'fish' or 'mammal')
  - some trigger words are typical others atypical representatives of the true category
- methodological investigation:
  - two groups: **click vs touch** to select category
- **hypothesis:** typical exemplars are easier to categorize than atypical ones
  - fewer mistakes
  - smaller RTs, AUC, MAD
  - less x-flips
  - less "change-of-mind" curve types
- **research question (methods):** any differences between click & touch selection?

variables used in the data set

`trial_id` = unique id for individual trials
`MAD` = maximal deviation into competitor space
`AUC` = area under the curve
`xpos_flips` = the amount of horizontal direction changes
`RT` = reaction time in ms
`prototype_label` = different categories of prototypical movement strategies
`subject_id` = unique id for individual participants
`group` = groups differ in the response design (click vs. touch)
`condition` = category membership (Typical vs. Atypical)
`exemplar` = the concrete animal
`category_left` = the category displayed on the left
`category_right` = the category displayed on the right
`category_correct` = the category that is correct
`response` = the selected category
`correct` = whether or not the `response` matches `category_correct`

Kieslich et al.'s (2019) replication of Dale et al.'s (2003) experiment