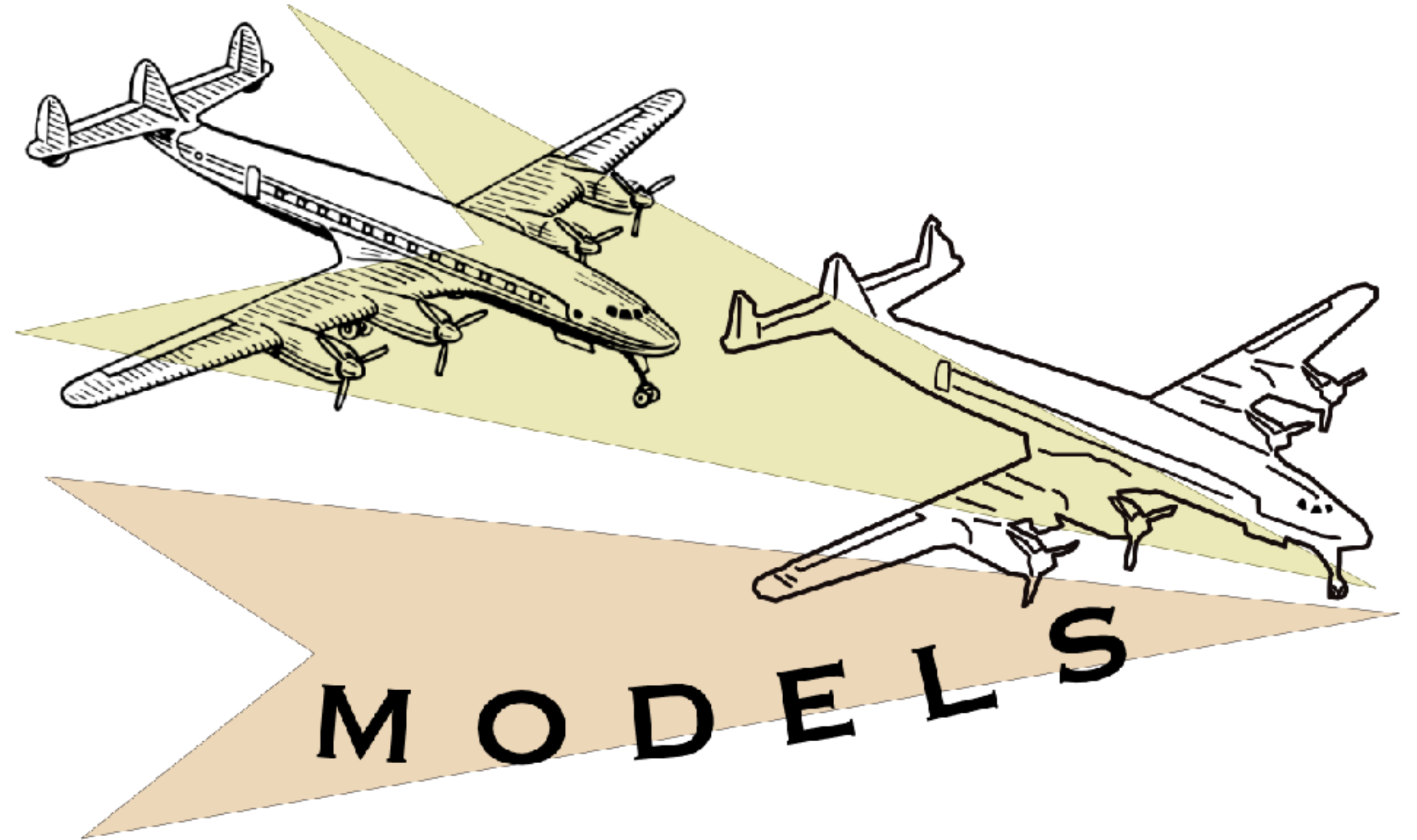


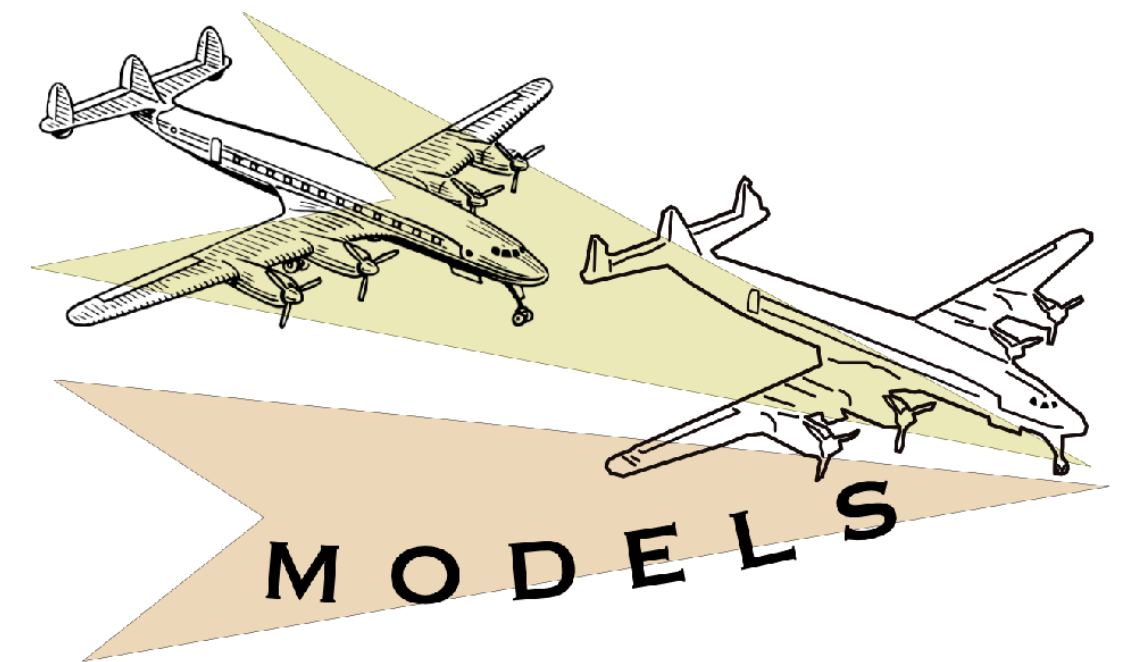
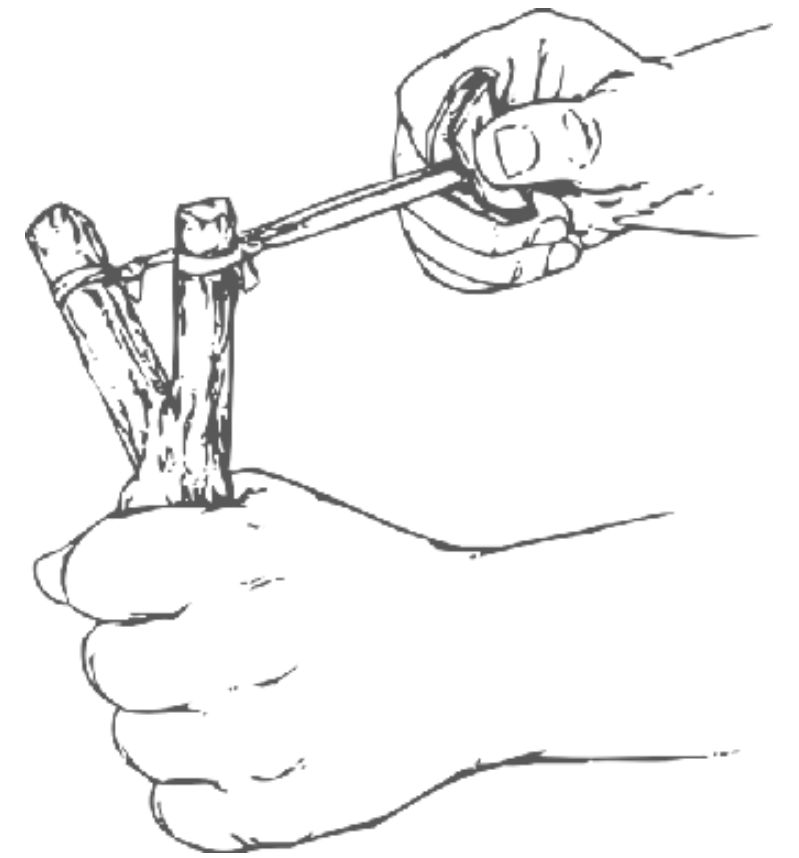
DATA ANALYSIS

MODELS



LEARNING GOALS

- ▶ become acquainted with statistical models
- ▶ understand what parameters are and what priors can do
 - ▶ likelihood function, parameters, prior, prior distribution
- ▶ understand notation to communicate models
 - ▶ formulas & graphs

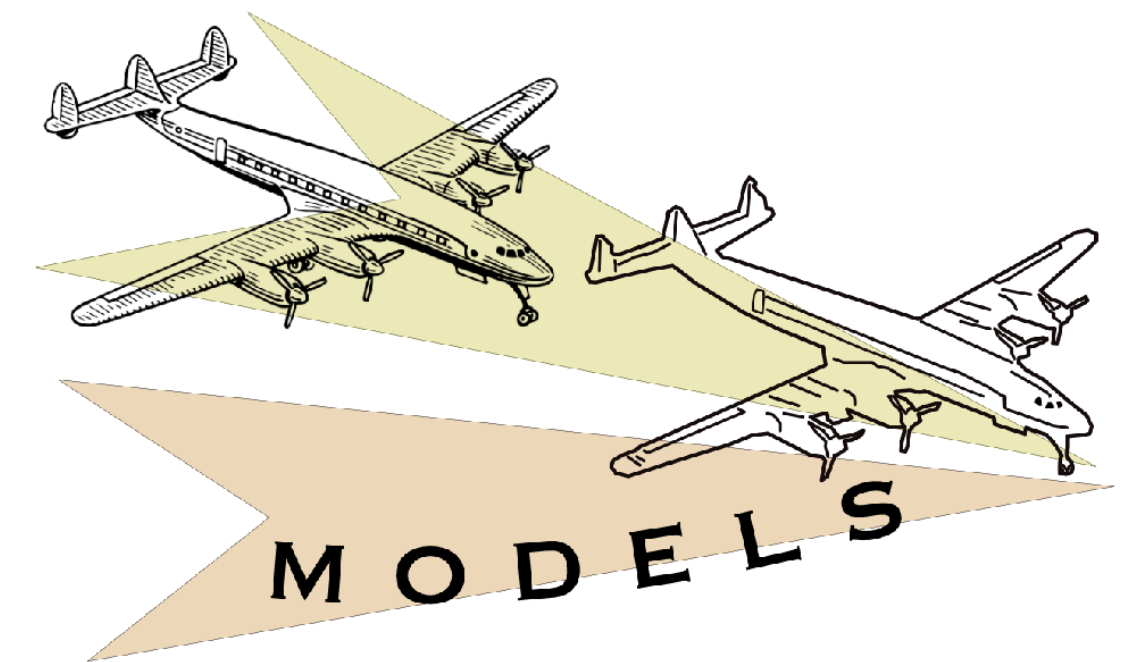




Statistical models

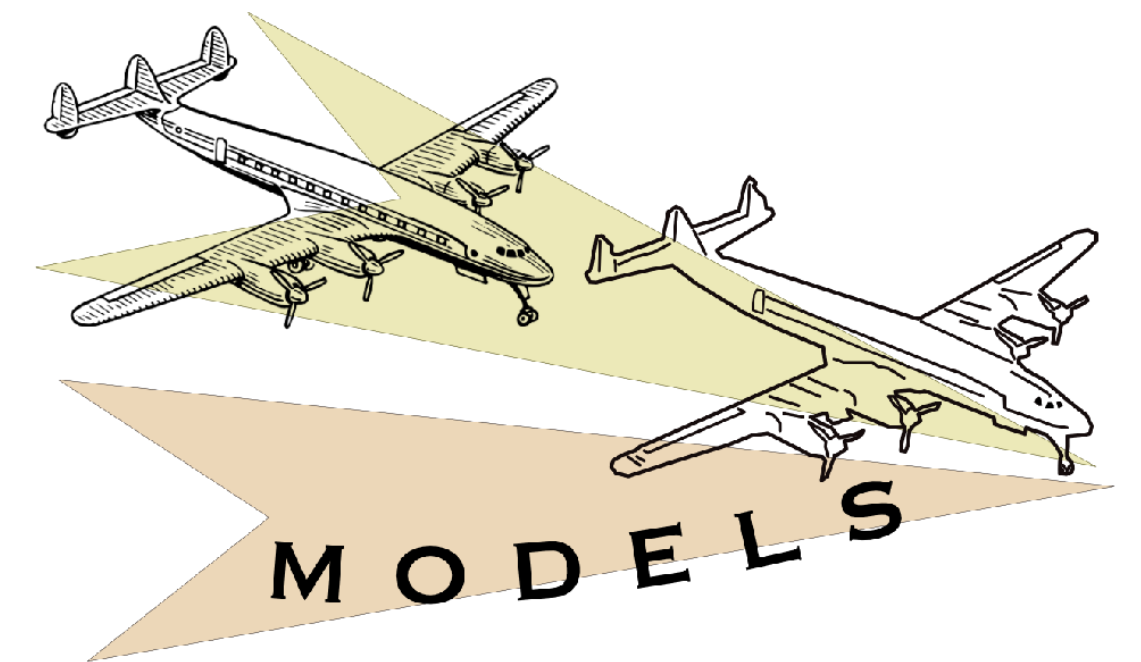
STATISTICAL MODELS

- ▶ 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.



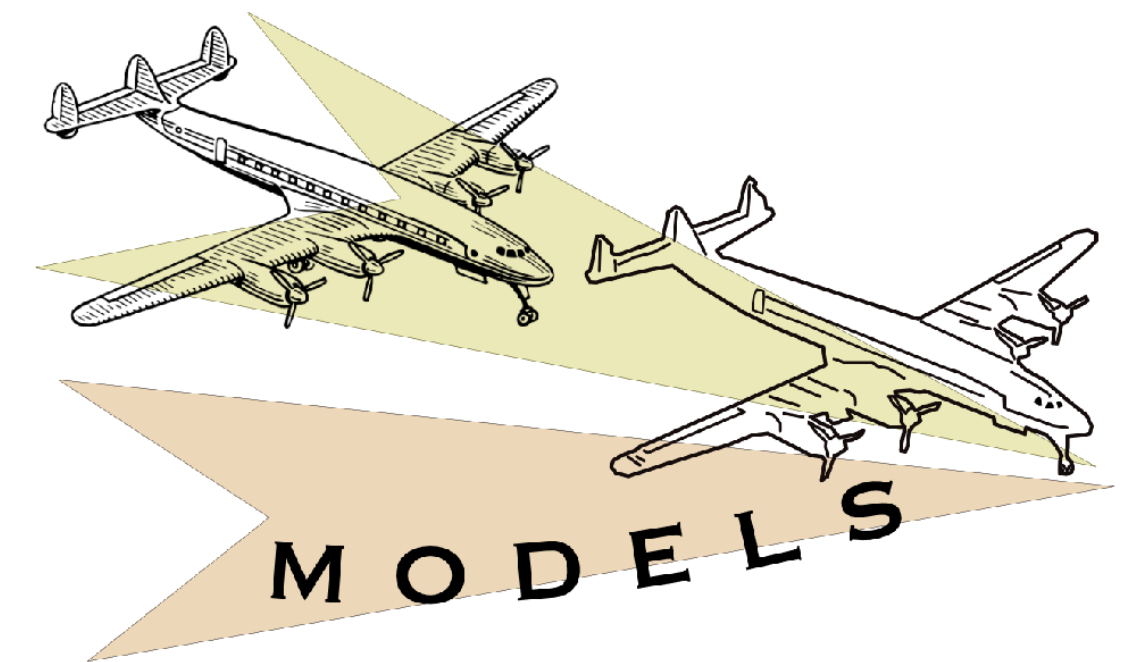
PRAGMATISM IN MODELING

All models are wrong, but some are useful. — Box (1979)



DEFINITION

- ▶ a **statistical model** M of random process R generating data D consists of:
 - ▶ a partition into D_{IV} and D_{DV} of a subset of D
 - ▶ a **likelihood function**: $P_M(D \mid \theta)$
 - ▶ [if Bayesian] a **prior**: $P_M(\theta)$





First examples

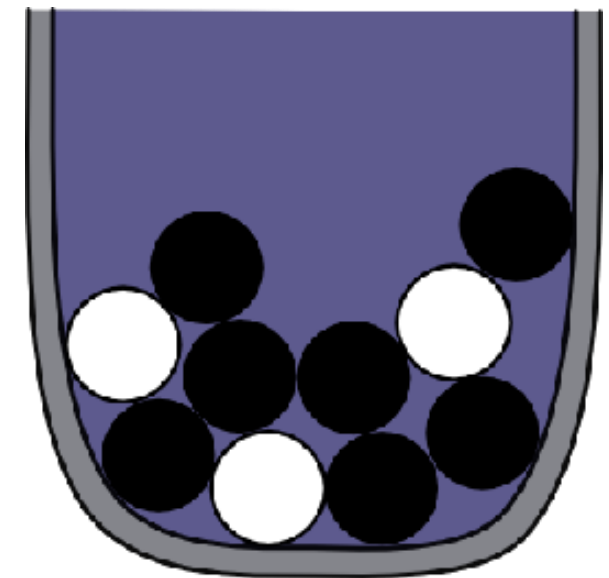
SINGLE DRAW FROM AN URN

- ▶ urn contains $N = 10$ balls
- ▶ unknown number $k \in \{0, 1, \dots, 10\}$ of black balls (rest white)
- ▶ likelihood function is uncontroversial

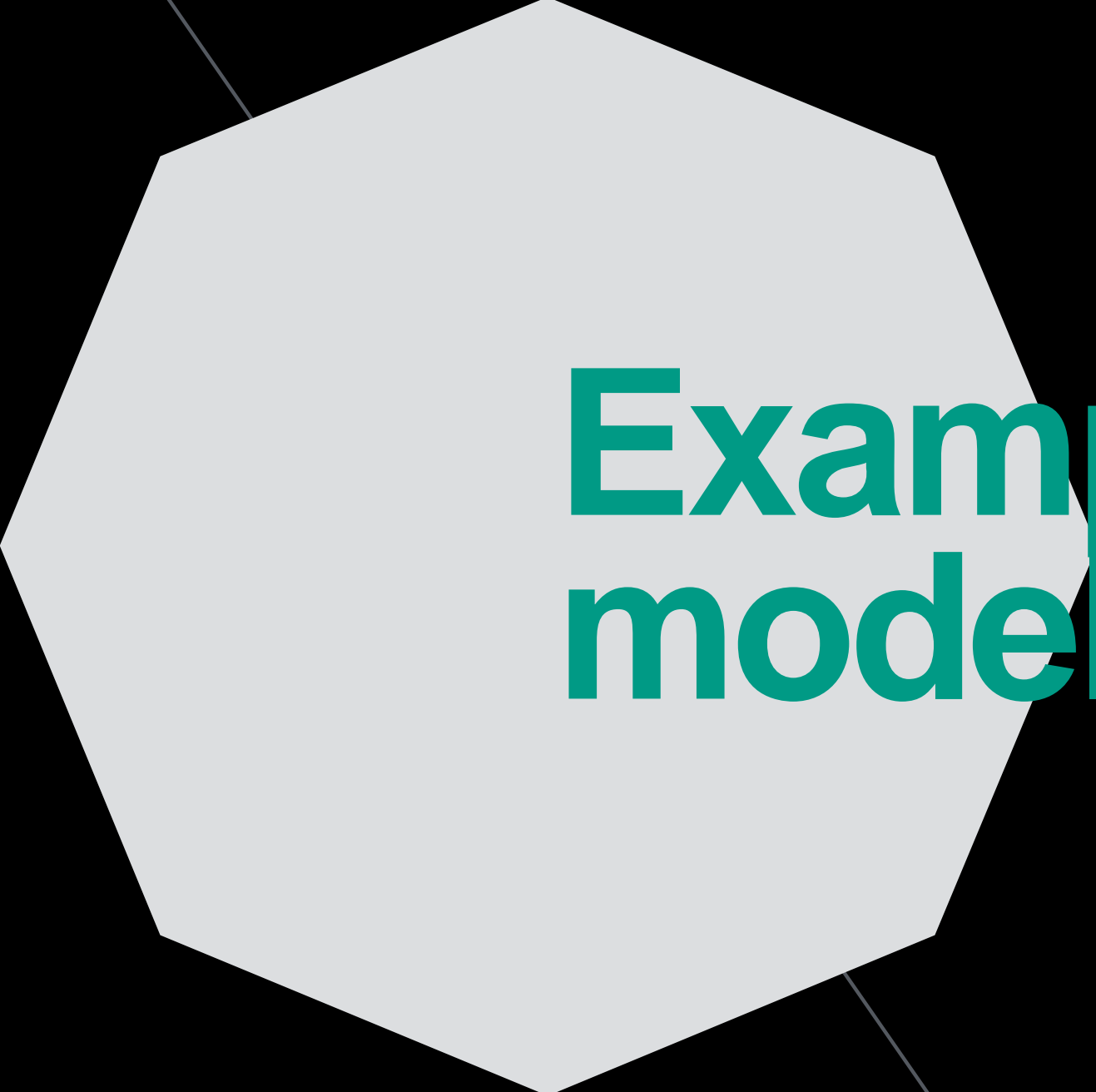
$$P_M(D = \text{black} \mid k) = \frac{k}{N}$$

- ▶ possible prior:

$$P_M(k = i) = \frac{1}{11}, \quad \text{for all } i \in \{0, 1, \dots, 10\}$$



one out of eleven
possible urns



**Example
models**

Notation & graphical representation

Recall that the *Binomial Model* has a binomial likelihood function:

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

And a Beta distribution as a prior, e.g., with shape parameters set so that all values of θ_c are equally likely.

$$P_M(\theta_c) = \text{Beta}(\theta_c, 1, 1)$$

Formula Notation

To concisely represent models, we use a special notation, which is very intuitive when we think about sampling. Instead of the above notation for the prior we write:

$$\theta_c \sim \text{Beta}(1, 1)$$

The symbol “ \sim ” is often read as “is distributed as”. You can also think of it as meaning that θ_c is sampled from a $\text{Beta}(1, 1)$ distribution.

Similarly, for the likelihood function, we just write:

$$k \sim \text{Binomial}(\theta_c, N).$$

Graph Notation

- known or unknown (= latent) variable
 - *shaded nodes*: observed variables
 - *unshaded nodes*: unobserved / latent variables
- kind of variable:
 - *circular nodes*: continuous variables
 - *square nodes*: discrete variables
- kind of dependency:
 - *single line*: stochastic dependency
 - *double line*: deterministic dependency

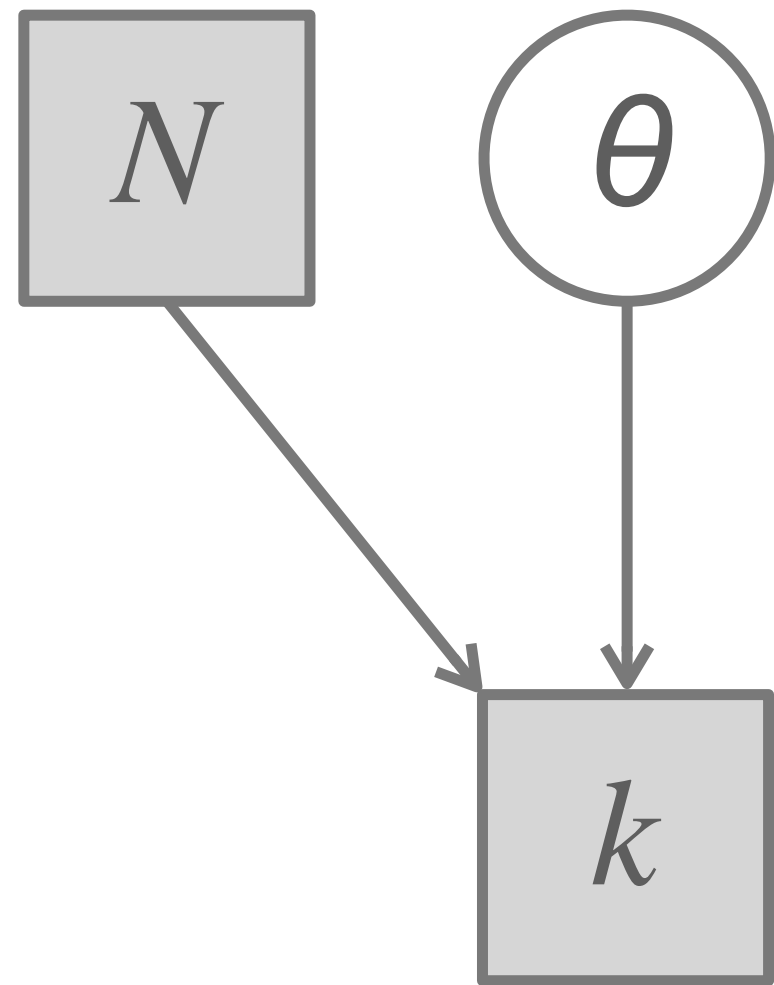
Graph Notation

For the Binomial Model this results in the relevant variables:

- number of trials (N)
- number of success (k)
- probability for success (θ_c)

Of these, N and k are observed and discrete variables, and θ_c is a latent continuous variable. Clearly, the number of heads k depends on the coin bias θ_c as well as on the number of trials N . This yields a graphical and formulaic notation as in Figure 8.1.

BINOMIAL MODELS



$$\theta \sim \text{Beta}(\dots)$$

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