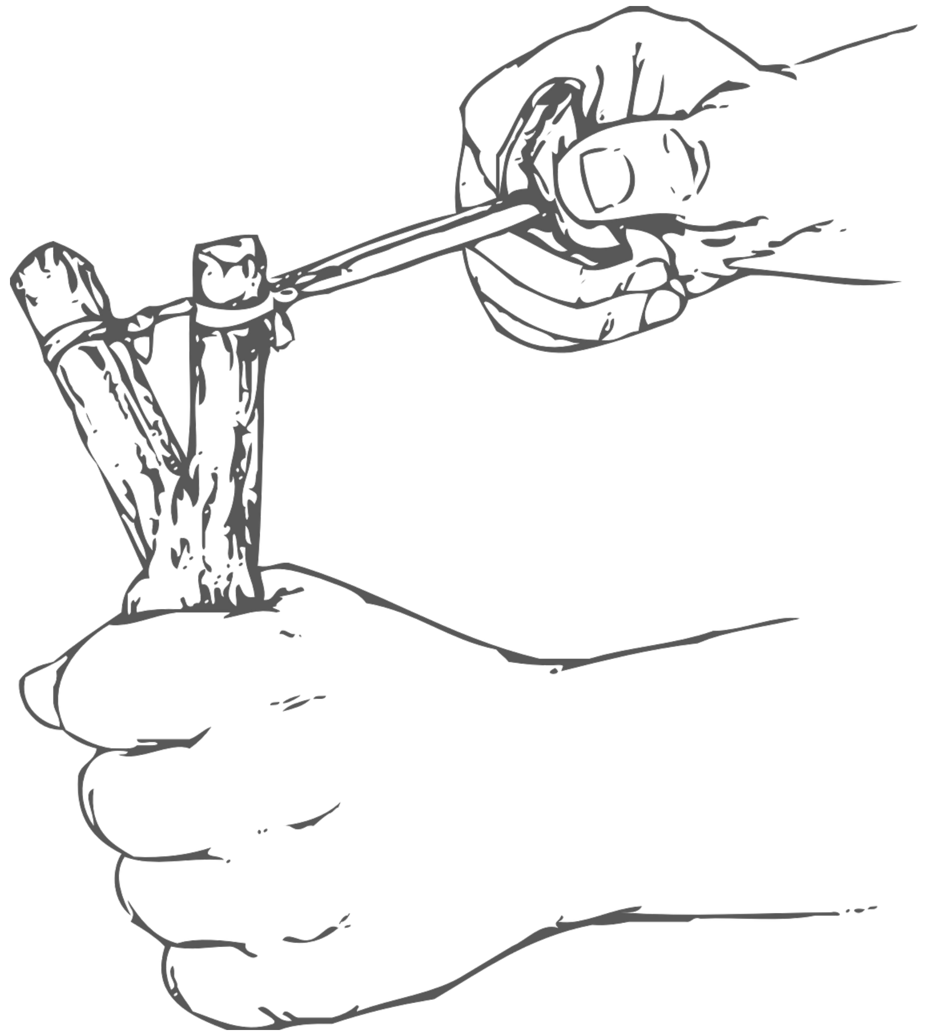
Bayesian data analysis: Theory & practice

Part 4b: Model checking

Michael Franke

Main learning goals



1. understand the role of model checking in statistical inquiry
   a. assessing implications of priors
   b. inspecting posterior predictives
2. apply common methods of posterior predictive checking
   a. visual
   b. Bayesian p-values

Three pillars of BDA

1. parameter estimation / inference [which parameter values are credible given data and model?]

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

2. predictions [which future data observations are likely given my model?]

a. prior

$$P(D_{\text{pred}}) = \int P(\theta)\ P(D_{\text{pred}} \mid \theta)\ \mathrm{d}\theta$$

b. posterior

$$P(D_{\text{pred}} \mid D_{\text{obs}}) = \int P(\theta \mid D_{\text{obs}})\ P(D_{\text{pred}} \mid \theta)\ \mathrm{d}\theta$$

3. model comparison [which model of two models is more likely to have generated the data?]

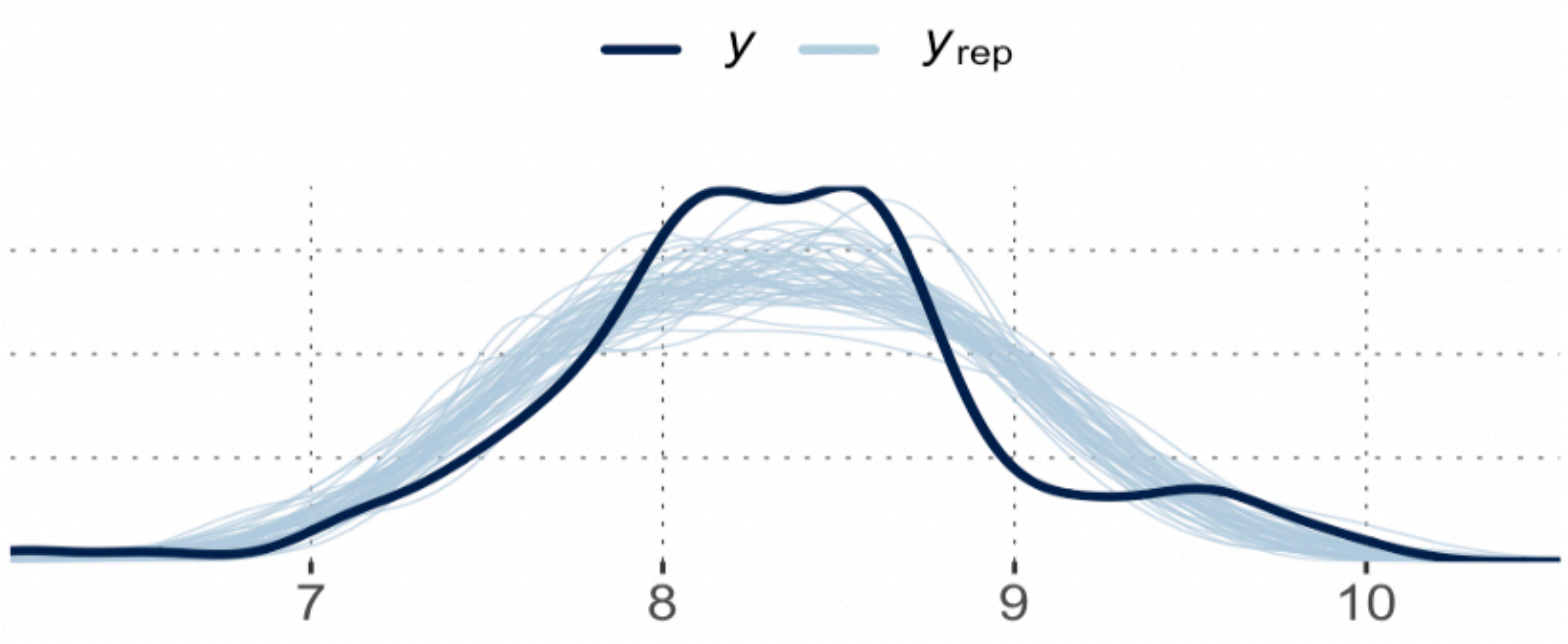
$$\underbrace{\frac{P(M_1 \mid D)}{P(M_2 \mid D)}}_{\text{posterior odds}} = \underbrace{\frac{P(D \mid M_1)}{P(D \mid M_2)}}_{\text{Bayes factor}} \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{prior odds}}$$

Visual posterior predictive checks

for world-temperature data

- black line:
  - distribution of observed temperature
- each of the 50 blue lines:
  - distribution of temperatures predicted for same years given a sample from the posterior

```
pp_check(fit_worldTemp, ndraws = 20)
```



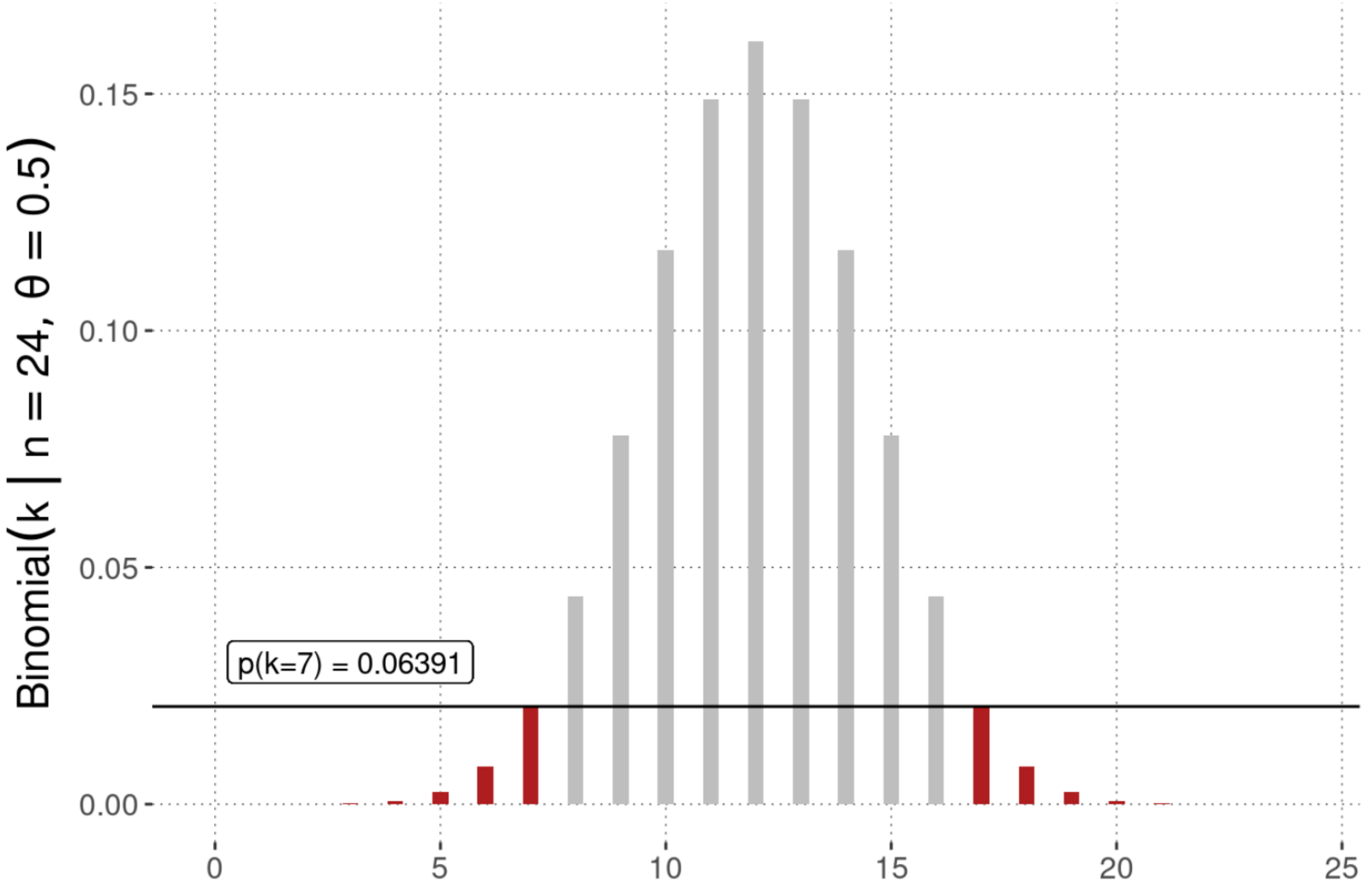
Recap: *p*-values

Monte Carlo approximation

- fix null hypothesis $\theta = \theta^*$
- derive **sampling distribution** $P(D \mid \theta^*)$
  - how likely is each possible observation under θ^*
- fix a **test statistic** $t(D)$
  - real number measuring some relevant aspect of D
- consider observed data D_{obs}
- p-value from MC simulation:
  - sample $d_1, \ldots, d_n \sim P(D \mid \theta^*)$
  - calculate:

$$p(D_{\text{obs}}) \approx \frac{1}{n} \sum_{i=1}^{n} \left[P(d_i \mid \theta^*) \leq P(D_{\text{obs}} \mid \theta^*) \right]$$

$$p\left(D_{\text{obs}}\right) = P\left(T^{|H_0} \succeq^{H_{0,a}} t\left(D_{\text{obs}}\right)\right)$$



read more here

Bayesian *p*-values

Monte Carlo approximation

- fix a model with $P(D \mid \theta)$ and $P(\theta)$
  - latter can be prior or posterior
  - gives prior / posterior predictive p-values
- gives **predictive distribution** $P_M(D)$
- fix a **test statistic** $t(D)$
  - real number measuring some relevant aspect of D
- consider observed data D_{obs}
- p-value from MC simulation:
  - sample $d_1, \ldots, d_n \sim P_M(D)$
  - calculate:

$$p(D_{\text{obs}}) \approx \frac{1}{n} \sum_{i=1}^{n} \left[P_M(d_i) \leq P_M(D_{\text{obs}}) \right]$$


demo
visual PPCs and Bayesian p