Predictive model selection and uncertainty

Vik Shirvaikar (joint work with Stephen Walker and Chris Holmes)

June 2025

How do we view uncertainty?

Frequentists

The parameter is **fixed** and the data is **random**

"If I ran this experiment several times, 95% of the confidence intervals I build using this procedure would contain the true mean."

Bayesians

The data is **fixed** and the parameter is **random**

"Given the observed data and my prior, there is a 95% posterior probability that the true mean lies within this credible interval."

Can we get the best of both worlds?

Uncertainty directly on a parameter,
without having to interpret results
through the lens of repeated sampling…

…but in a completely data-dependent framework that **doesn't require the specification of a prior** distribution

Outline

1. Prequential forecasting and proper scoring
2. Martingale posterior distributions (Fong et al., 2023)
3. Model uncertainty via predictive resampling
    A. Illustration: density estimation
    B. Illustration: hypothesis testing

Prior works have emphasized the value of prediction

> All this is not to deny the usefulness and convenience of parametric models; clearly, the extraordinarily appealing notion of exchangeability when applied to statistical paradigms implies the mathematical existence of the lurking parameter as evidenced by the de Finetti representation theorem. Further, in some situations parameters may be important hypothetical constructs in providing an appropriate lubricant or focus for a model and in planning an experimental program whether or not they can be regarded as real physical entities. However, the time has come for statistical analyses to emphasize the observable and the finite in contrast to the potentially artificial and infinite.

Geisser (1982), “Aspects of the predictive and estimative approaches in the determination of probabilities”

Prior works have emphasized the value of prediction

SUMMARY

The prequential approach is founded on the premiss that the purpose of statistical inference is to make sequential probability forecasts for future observations, rather than to express information about parameters. Many traditional parametric concepts, such as consistency and efficiency, prove to have natural counterparts in this formulation, which sheds new light on these and suggests fruitful extensions.

3. PREQUENTIAL FORECASTING SYSTEMS

So now let $\mathbf{X} = (X_1, X_2, \ldots)$ be a sequence of uncertain quantities. At any time n, the prequential forecaster, with the values $\mathbf{x}^{(n)}$ of $\mathbf{X}^{(n)} = (X_1, X_2, \ldots, X_n)$ to hand, must issue a probability forecast distribution P_{n+1} for the next observation X_{n+1}.

Dawid (1984), “Statistical theory: the prequential approach”

A scoring rule is a summary measure

- Suppose I have a model $\mathscr{M}$ and start to observe some data $x_1, x_2, \ldots$
- How do I evaluate this model's forecasts? I use a **scoring rule**
- For example, I could use a very simple rule where I score one point if my most-likely guess is correct, and zero points otherwise

$$S(x, \mathscr{M}) = \begin{cases} 1, & \text{if } x = \arg\max_y \mathscr{M}(y), \\ 0, & \text{otherwise} \end{cases}$$

Matheson and Winkler (1976), "Scoring Rules for Continuous Probability Distributions"

The marginal likelihood is a scoring rule

- The Bayesian **marginal likelihood** of any model $\mathscr{M}$ can be factorized as a product of one-step-ahead predictive densities

$$p(x_{1:n} \mid \mathscr{M}) = \prod_{i=1}^{n} p_{\mathscr{M}}(x_i \mid x_{1:i-1})$$

- By logging each side, we get a scoring rule!

$$S(x_{1:n}, \mathscr{M}) = \log p(x_{1:n} \mid \mathscr{M}) = \sum_{i=1}^{n} \log p_{\mathscr{M}}(x_i \mid x_{1:i-1})$$

The marginal likelihood is a scoring rule

- It turns out that this particular scoring rule has some nice properties:
  - It's **prequential**, treating the data as if they arrive in sequence, like in reality
  - It's **proper**, because its expectation is maximized by using the true distribution for forecasting — there's no way to "game the system"
  - It's **unique** in guaranteeing coherent model evaluation

$$S(x_{1:n}, \mathscr{M}) = \log p(x_{1:n} \mid \mathscr{M}) = \sum_{i=1}^{n} \log p_{\mathscr{M}}(x_i \mid x_{1:i-1})$$

Gneiting and Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation"
Fong and Holmes (2020), "On the Marginal Likelihood and Cross-Validation"

The Bayes factor is a marginal likelihood (ratio)

- We know that the **Bayes factor** is the ratio of evidence between two models

$$BF = \frac{p(x_{1:n} \mid \mathscr{M}_1)}{p(x_{1:n} \mid \mathscr{M}_2)}$$

- This means the Bayes factor has an interpretation based on prequential scoring

$$\log BF = S(x_{1:n}, \mathscr{M}_1) - S(x_{1:n}, \mathscr{M}_2)$$

I.J. Good (1952), “Rational Decisions”
Dawid (1984), “Statistical theory: the prequential approach”

The BIC is (approximately) the Bayes factor

- Our favorite model selection rule is the **BIC** (because it has “Bayesian” in it)

$$\text{BIC}(\mathscr{M}_j) = -2 \log p(x_{1:n} \mid \mathscr{M}_j, \hat{\theta}_j) + d_j \log(n)$$

- Optimizing the BIC is asymptotically equivalent to optimizing the Bayes factor

$$\log BF \approx -\frac{\text{BIC}(\mathscr{M}_1) - \text{BIC}(\mathscr{M}_2)}{2}$$

- This allows the BIC to inherit key properties of the Bayes factor, specifically **consistency** — the probability of selecting the correct model converges to one

Schwarz (1978), “Estimating the dimension of a model”

Kass and Raftery (1995), “Bayes factors”

The BIC is an optimal predictive summary measure!

- How do we evaluate model predictions?

  ➡ We use a **scoring rule**

  ➡ such as the **marginal likelihood**

  ➡ which is captured by the **Bayes factor**

  ➡ which is approximated by the **BIC**

> the model-selection method which proceeds by maximizing the adjusted prequential likelihood, or equivalently minimizing the "adjusted stochastic complexity" of x^n, $-\log p_j(x^n) - \log \alpha_j$, will be (almost everywhere) consistent.

Dawid (1992), "Prequential Analysis, Stochastic Complexity, and Bayesian Inference"

Outline

1. Prequential forecasting and proper scoring
2. Martingale posterior distributions (Fong et al., 2023)
3. Model uncertainty via predictive resampling
    A. Illustration: density estimation
    B. Illustration: hypothesis testing

Uncertainty comes from missing data

> Clarity is even more easily attainable, perhaps, if it is recognized that an inference about finite population parameters can be interpreted as a predictive inference about a sample, namely, *the unsampled part of a finite population*. This sample is already in existence but has not been observed by the statis-

Roberts (1965), “Probabilistic prediction”

Uncertainty comes from missing data

JOURNAL ARTICLE

Martingale posterior distributions

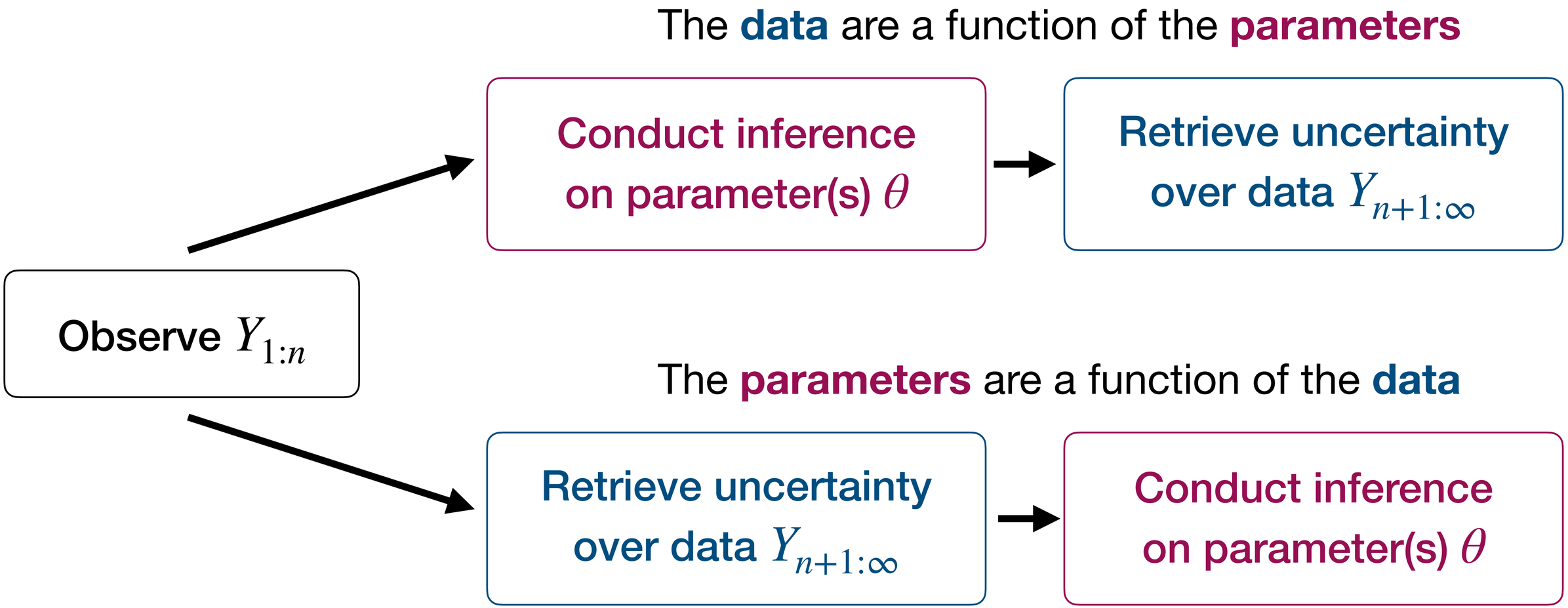
Edwin Fong, Chris Holmes ✉, Stephen G Walker Author Notes

Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 85, Issue 5, November 2023, Pages

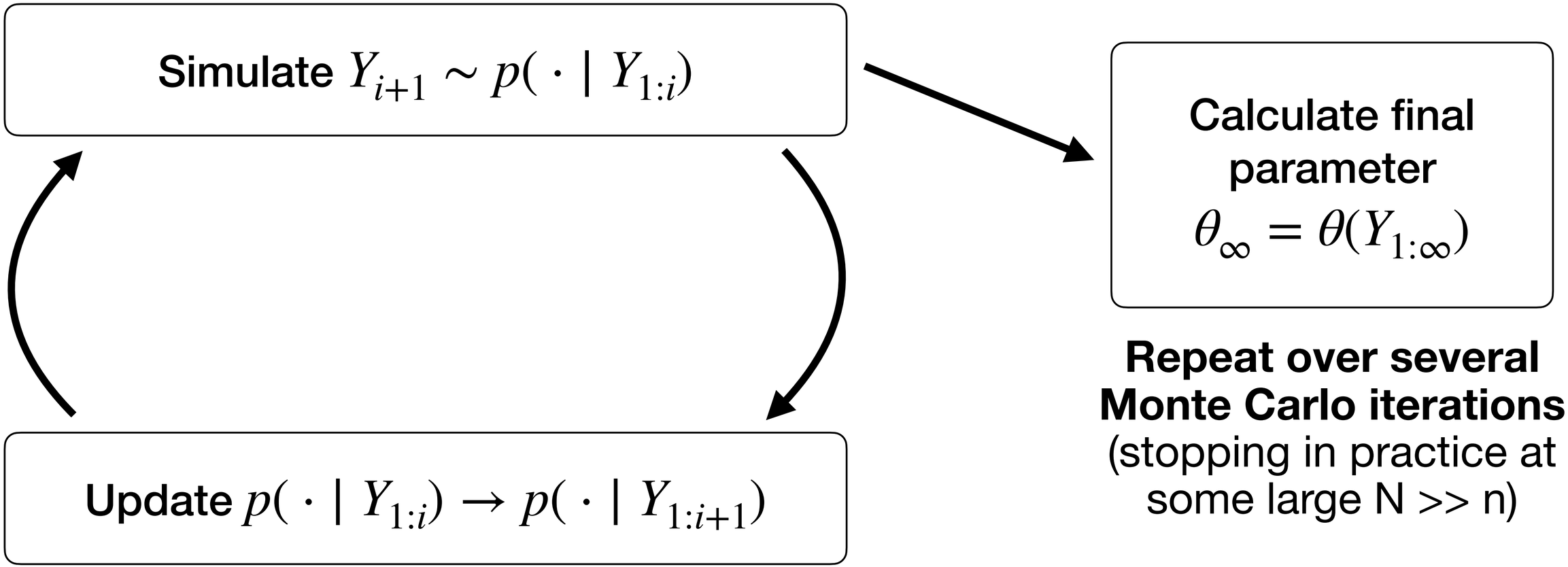
Abstract

The prior distribution is the usual starting point for Bayesian uncertainty. In this paper, we present a different perspective that focuses on missing observations as the source of statistical uncertainty, with the parameter of interest being known precisely given the entire population. We argue that the foundation of Bayesian inference is to assign a distribution on missing observations conditional on what has been observed. In the i.i.d. setting with an observed sample of size n, the Bayesian would thus assign a predictive distribution on the missing $Y_{n+1:\infty}$ conditional on $Y_{1:n}$, which then induces a distribution on the parameter.

Uncertainty comes from missing data



The predictive resampling pipeline



Repeat over several Monte Carlo iterations (stopping in practice at some large N >> n)

Different perspective, same uncertainty

likelihood prior

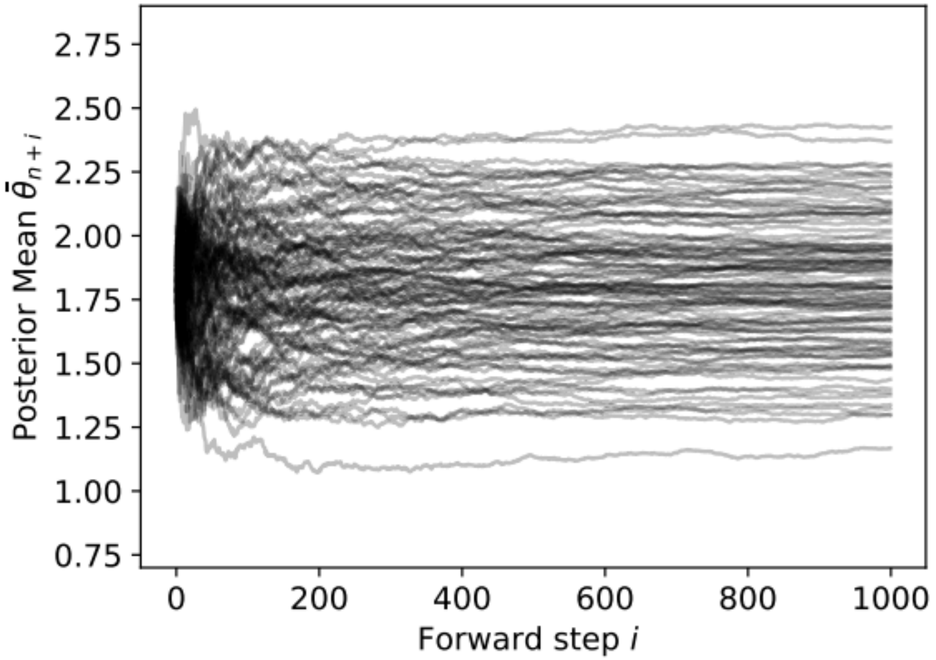
Let $f_\theta(y) = \mathcal{N}(y \mid \theta, 1)$, with $\pi(\theta) = \mathcal{N}(\theta \mid 0, 1)$. Given an observed dataset $y_{1:n}$, the tractable posterior density takes on the form $\pi(\theta \mid y_{1:n}) = \mathcal{N}(\theta \mid \bar{\theta}_n, \bar{\sigma}_n^2)$ where

$$\bar{\theta}_n = \frac{\sum_{i=1}^n y_i}{n+1}, \quad \bar{\sigma}_n^2 = \frac{1}{n+1}.$$

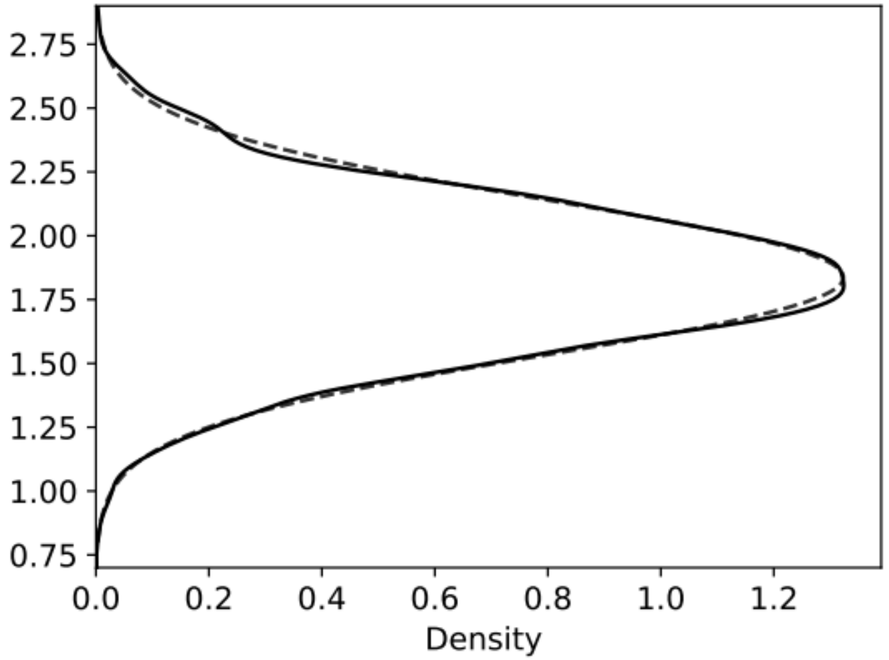
The posterior predictive density then takes on the form $p(y \mid y_{1:n}) = \mathcal{N}(y \mid \bar{\theta}_n, 1 + \bar{\sigma}_n^2)$. For observed data, we generated $y_{1:n} \overset{\text{iid}}{\sim} f_\theta(y)$ for $n = 10$ with $\theta = 2$, giving $\bar{\theta}_n = 1.84$.

Fong et al. (2023), “Martingale posterior distributions”

Different perspective, same uncertainty



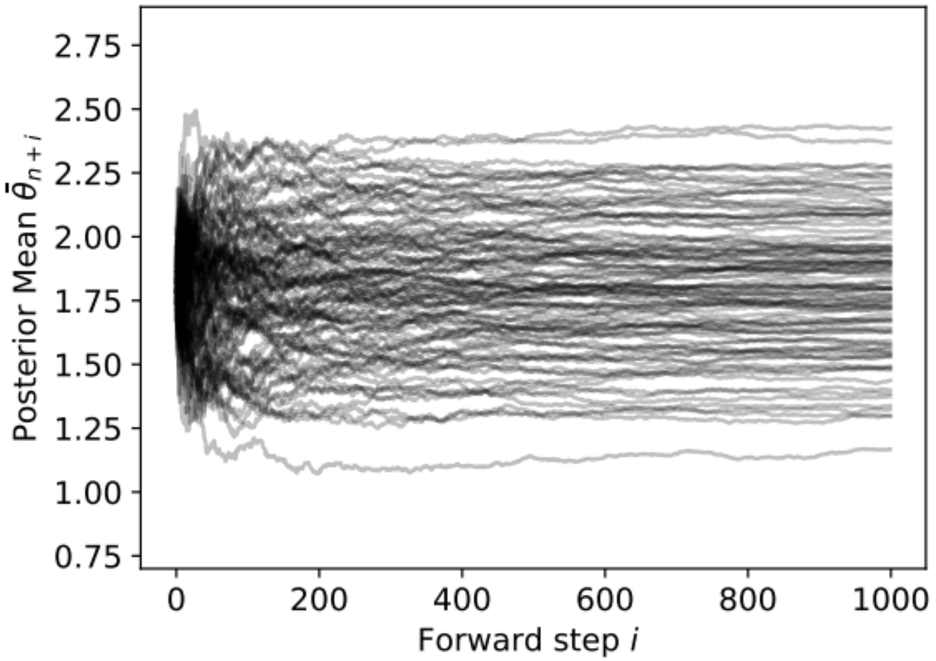
The **predictive** uncertainty from the imputation of additional data…

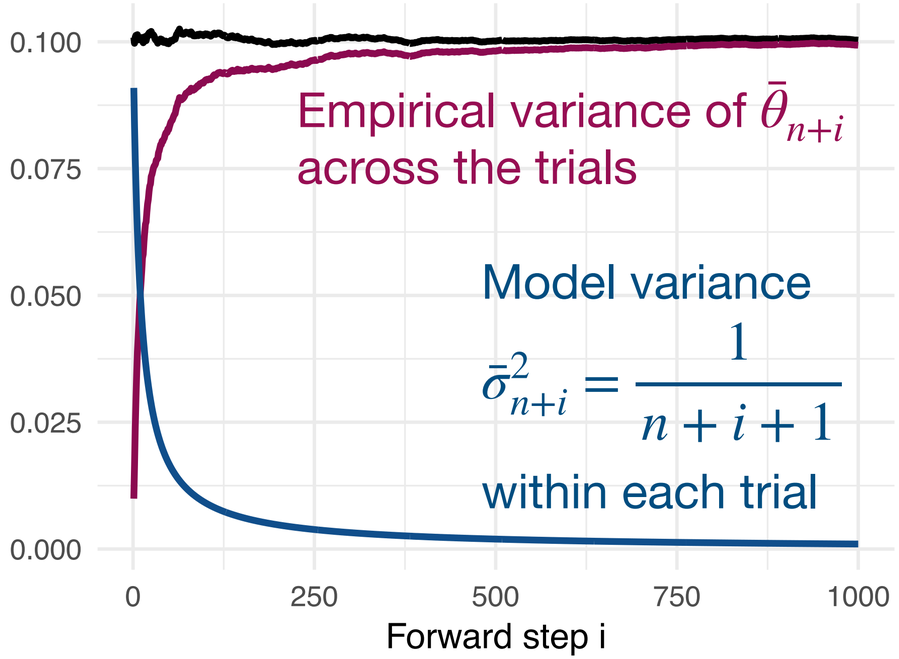


…is equivalent to the usual **posterior** uncertainty from Bayesian methods

Fong et al. (2023), “Martingale posterior distributions”

What type of uncertainty is this?





The **total** variance can be decomposed into **aleatoric** and **epistemic**

Fong et al. (2023), “Martingale posterior distributions”

Outline

1. Prequential forecasting and proper scoring
2. Martingale posterior distributions (Fong et al., 2023)
3. Model uncertainty via predictive resampling
    A. Illustration: density estimation
    B. Illustration: hypothesis testing

Acknowledging model structure

Let our model $\mathscr{M} = \{S, \theta\}$ consist of structural assumptions S, such as the link function in a GLM, and parameters θ whose meaning depends on S

> of S. It is common in statistical theory and practice to acknowledge parametric uncertainty about θ given a particular assumed structure S; it is less common to acknowledge structural uncertainty about S itself. A widely used approach involves enlisting the aid of x to specify a plausible single 'best' choice S^* for S, and then proceeding as if S^* were known to be correct. In general this approach fails to assess and propagate structural uncertainty fully and may lead to miscalibrated uncertainty assessments about y given x. When miscalibration

Draper (1995), "Assessment and propagation of model uncertainty"

Bayesian model averaging

model selection
with AIC, BIC, etc.

$$P(\xi \mid \mathscr{D}) = \frac{P(\mathscr{D} \mid \xi)\, P(\xi)}{P(\mathscr{D})} \quad \longleftrightarrow \quad P(\xi \mid \mathscr{D}, \mathscr{M}^*) = \frac{P(\mathscr{D} \mid \xi, \mathscr{M}^*)\, P(\xi \mid \mathscr{M}^*)\, P(\mathscr{M}^*)}{P(\mathscr{D} \mid \mathscr{M}^*)}$$

$$P(\xi \mid \mathscr{D}) = \sum_{k=1}^{K} P(\xi \mid \mathscr{M}_k, \mathscr{D})\, P(\mathscr{M}_k \mid \mathscr{D})$$

model weighting

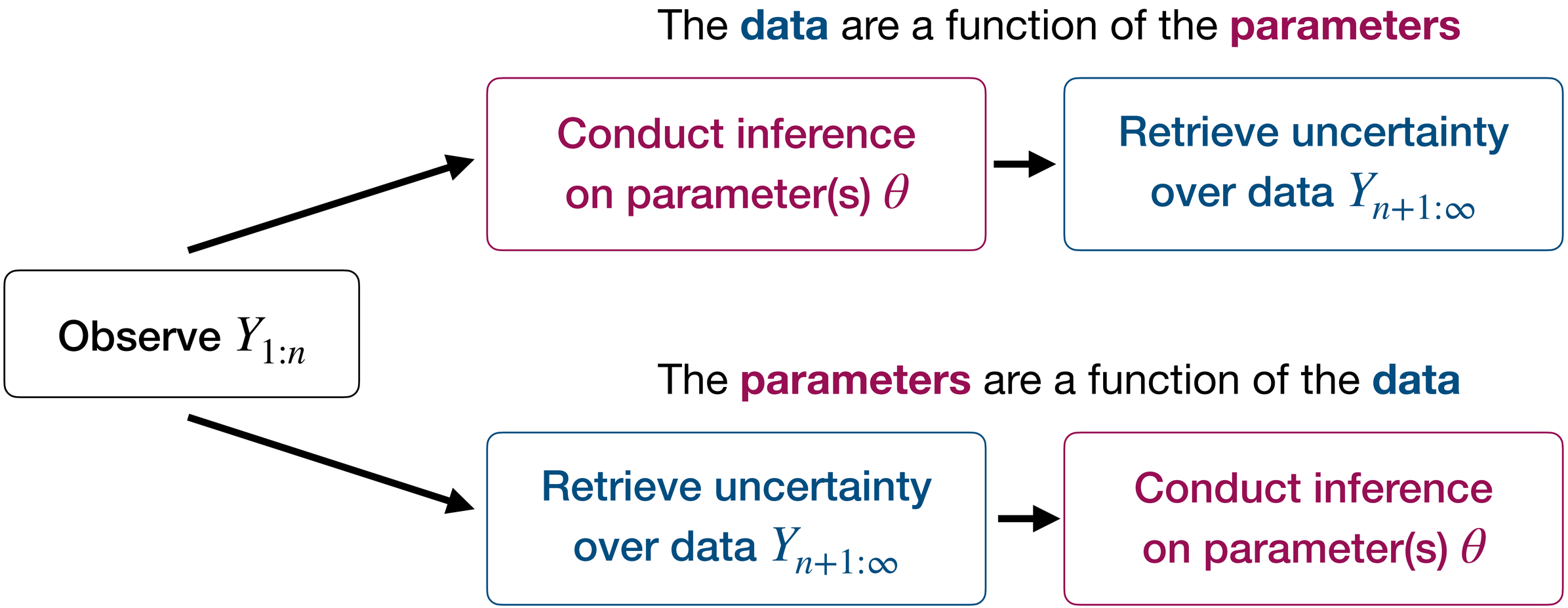
Where do we get the weights?

Existing approaches share some common concerns

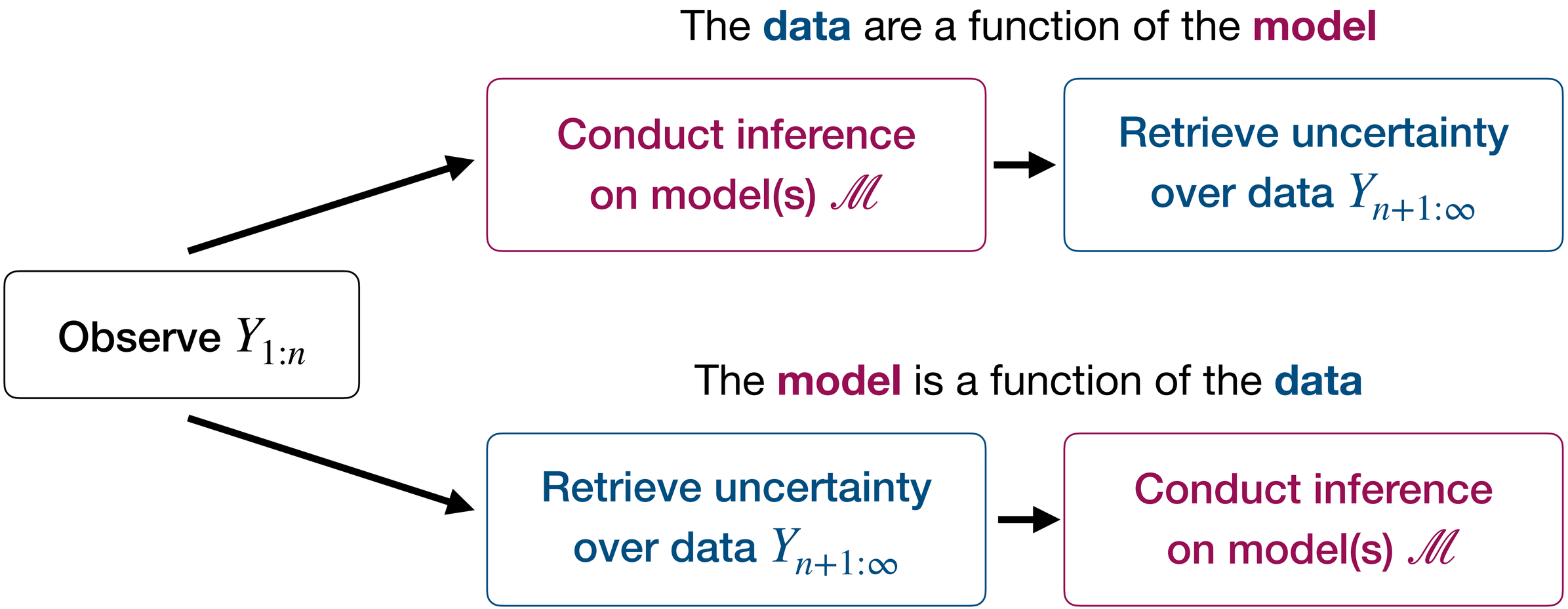
- Specification of **priors**
- Complex **integration** over parameter spaces
- **Inexact** or approximate results

So we return to our guiding principle from earlier…

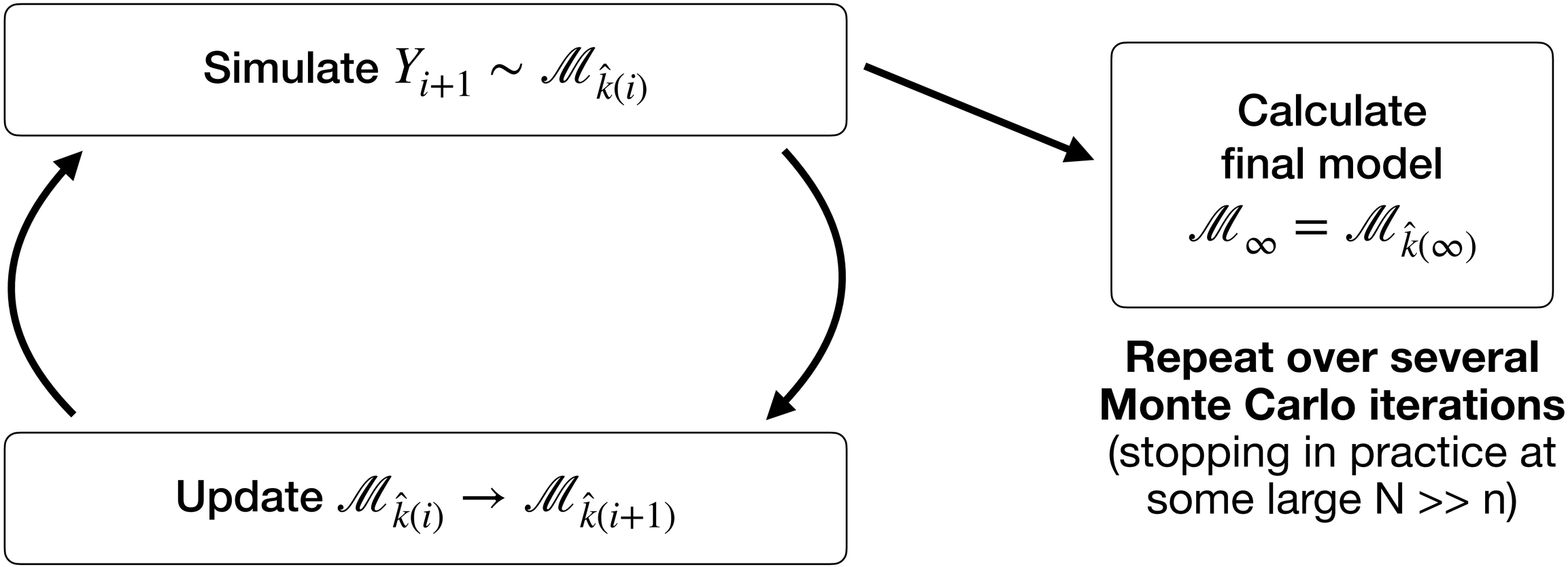
Uncertainty comes from missing data



Uncertainty comes from missing data



The predictive resampling pipeline



Repeat over several Monte Carlo iterations (stopping in practice at some large N >> n)

The predictive resampling pipeline

Key requirement: specification of a **consistent** one-step model selection criterion, which is then converted directly into a measure of uncertainty

BIC AIC LOO-CV

Convergence again comes from martingales

Calculate final model

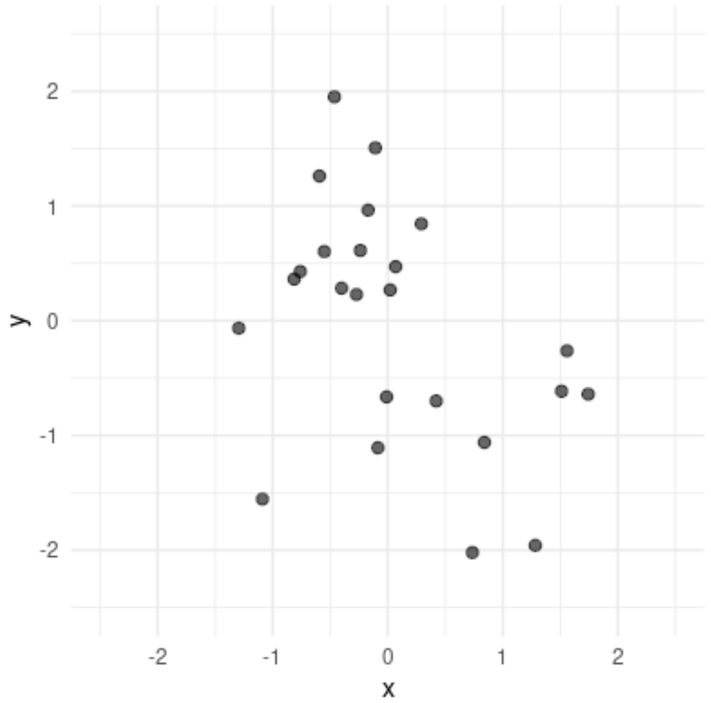
$$\mathscr{M}_\infty = \mathscr{M}_{\hat{k}(\infty)}$$

Repeat over several Monte Carlo iterations (stopping in practice at some large N >> n)

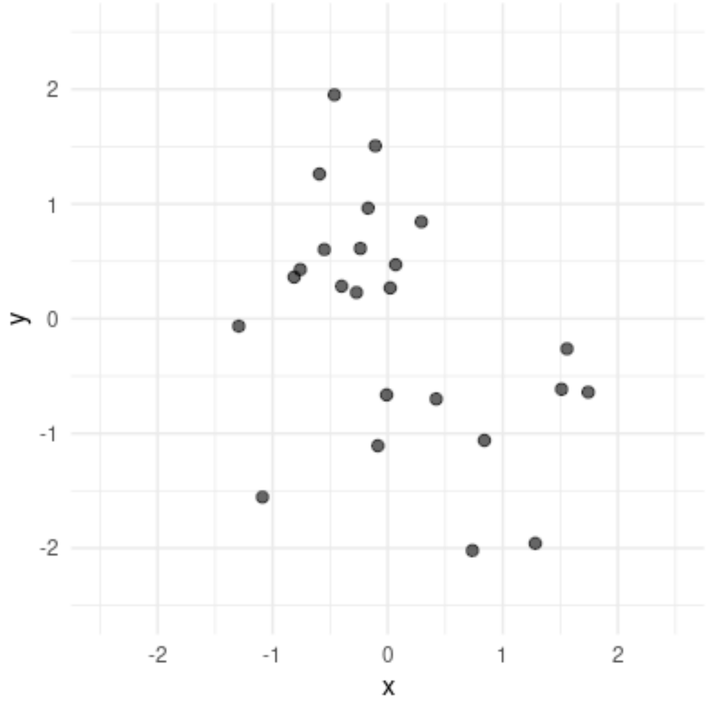
Outline

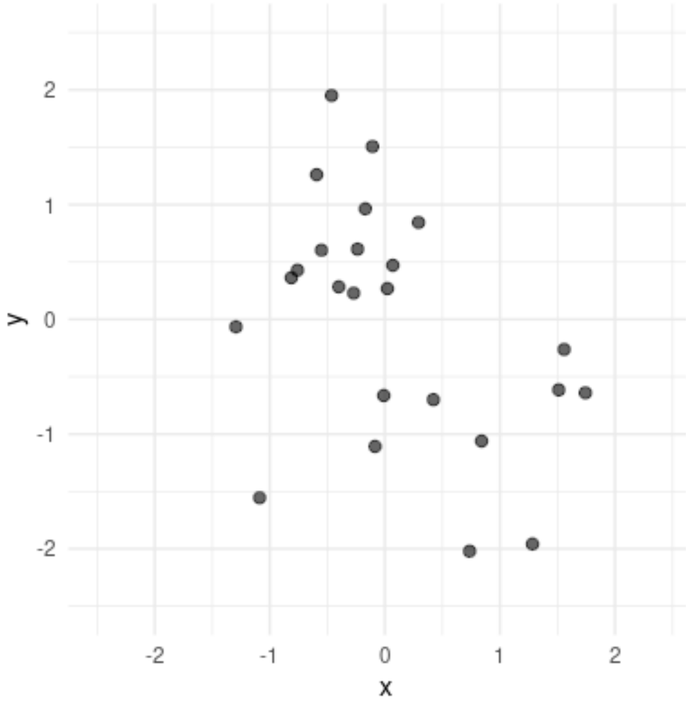
1. Prequential forecasting and proper scoring
2. Martingale posterior distributions (Fong et al., 2023)
3. Model uncertainty via predictive resampling
    A. Illustration: density estimation
    B. Illustration: hypothesis testing

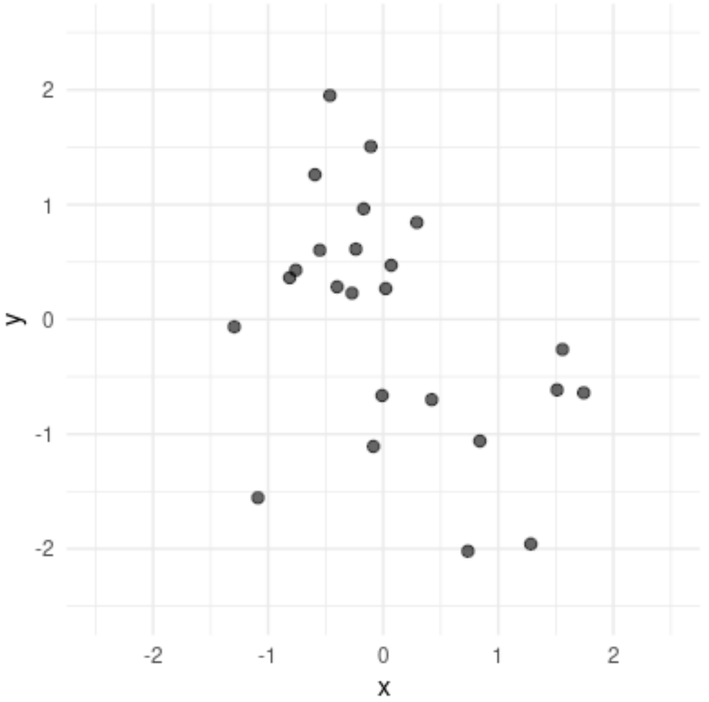
Problem: Find the number of components in a Gaussian mixture model (GMM)

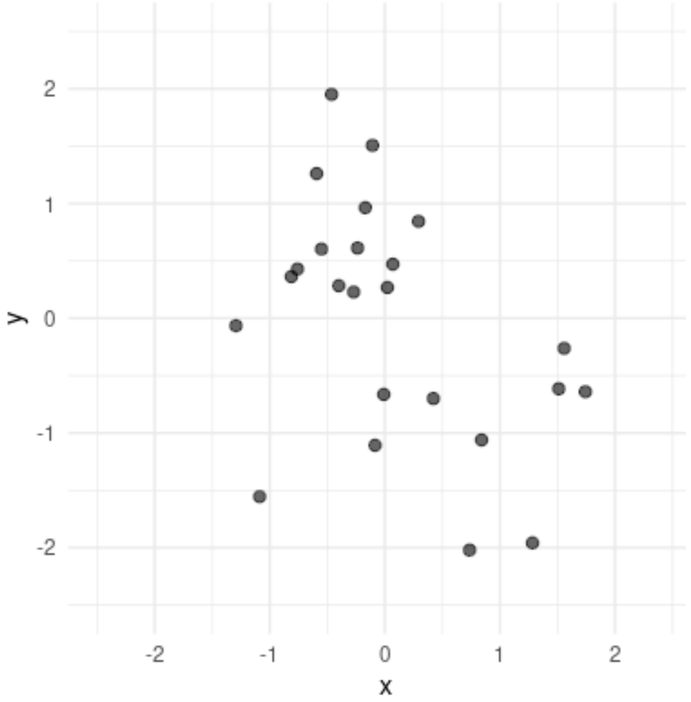


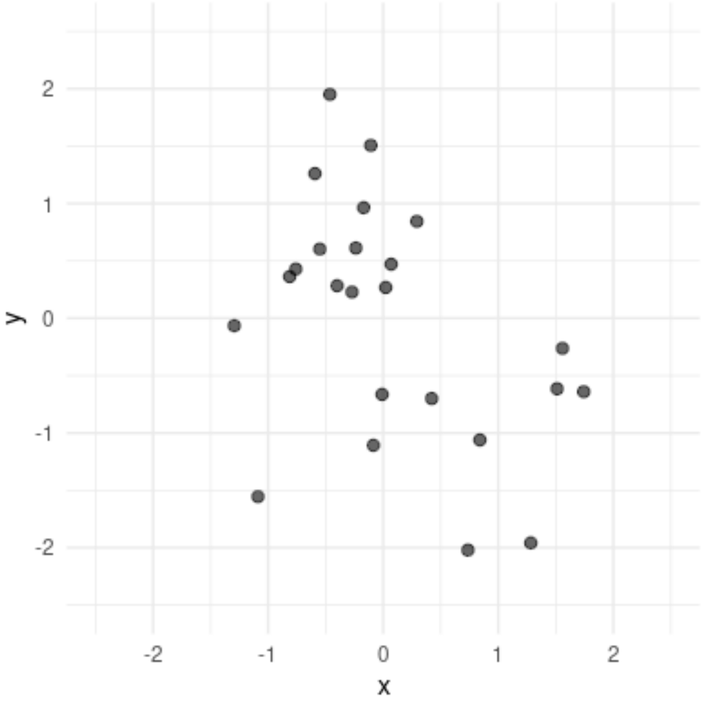
Problem: Find the number of components in a Gaussian mixture model (GMM)



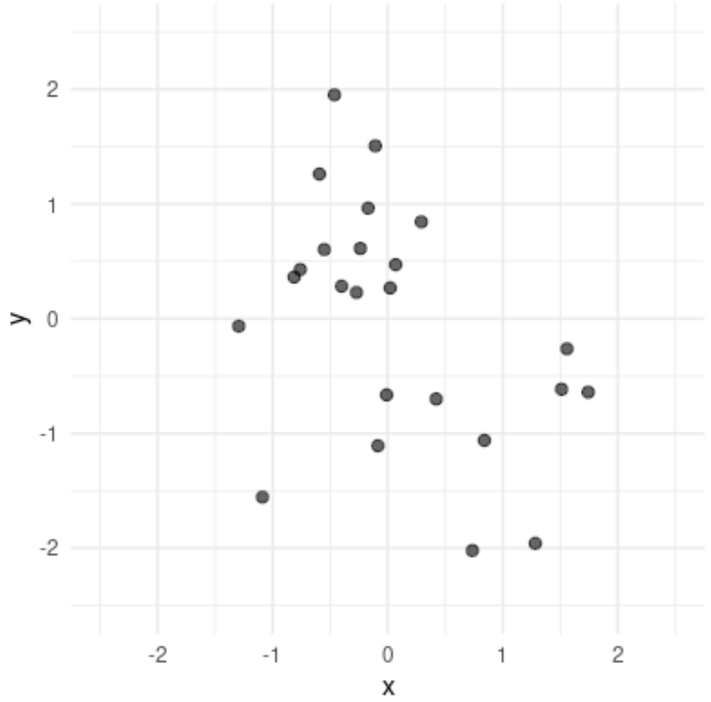


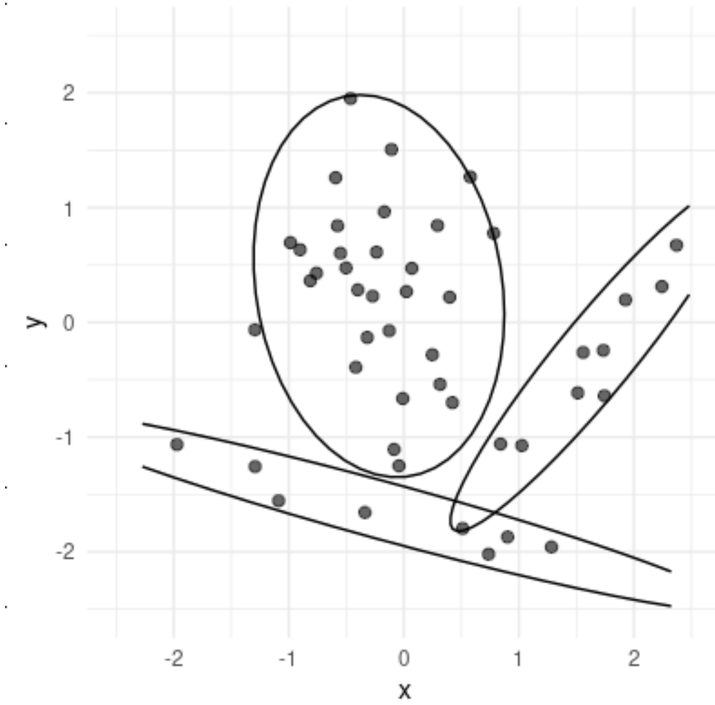


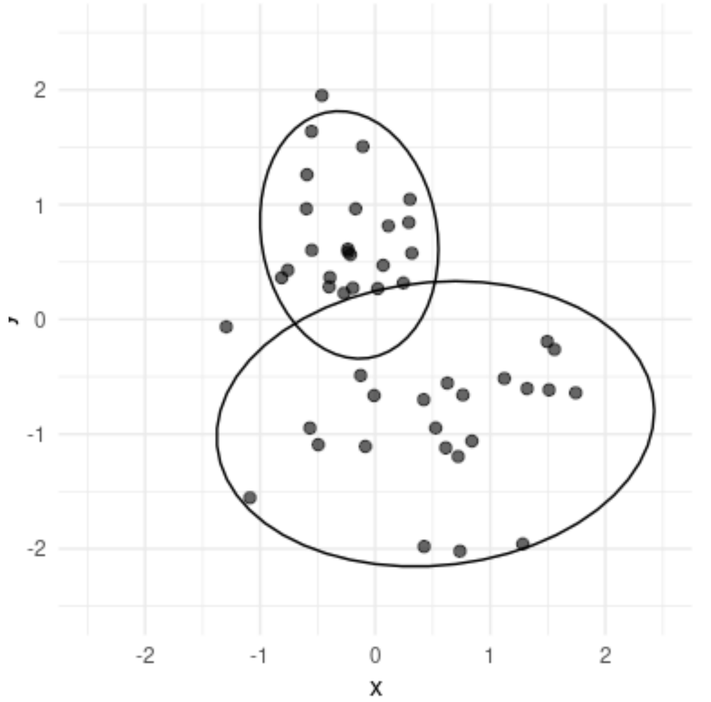


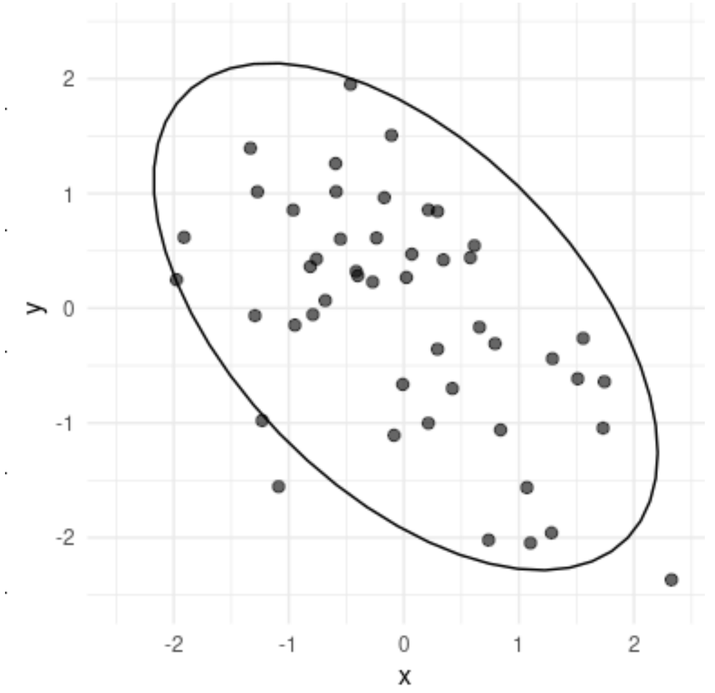


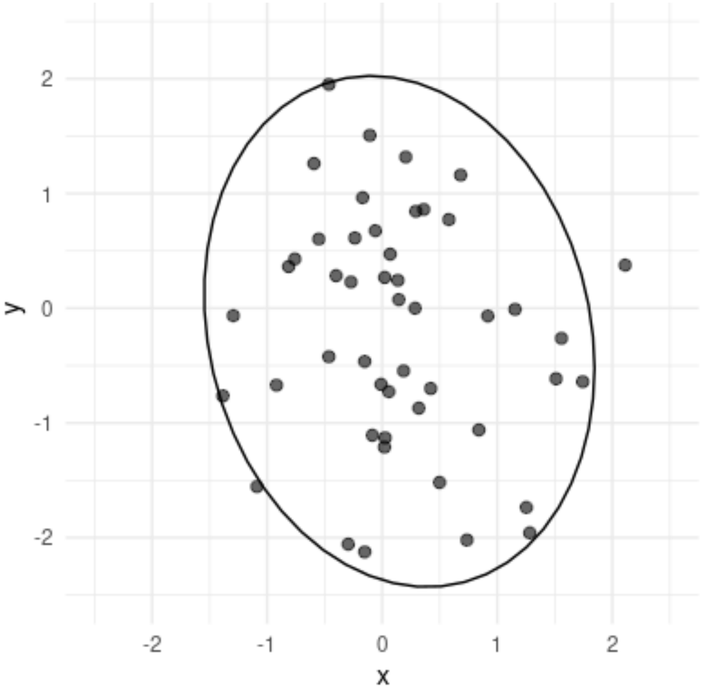
Problem: Find the number of components in a Gaussian mixture model (GMM)











Outline

1. Prequential forecasting and proper scoring
2. Martingale posterior distributions (Fong et al., 2023)
3. Model uncertainty via predictive resampling
    A. Illustration: density estimation
    B. Illustration: hypothesis testing

How do we test a hypothesis?

Frequentists

The parameter is **fixed** and the data is **random**

"If I ran this experiment several times, I'd expect to reject the null hypothesis only 5% of the time when it is actually true."

Bayesians

The data is **fixed** and the parameter is **random**

"Given the observed data and my prior, there is a 95% posterior probability that the alternative hypothesis is true over the null."

Simple point hypothesis demonstration

Unknown mean parameter for $\mathcal{N}(\theta, 1)$

Observed data $x_1, \ldots, x_n$

$$H_0 : \theta = 0$$
$$H_1 : \theta = 0.1$$

$$\max_{\theta_k} \mathscr{L} = \sum_{i=1}^{n} \log\, p(x_i \mid \theta_k) \quad \longrightarrow \quad x_{n+1} \sim \mathcal{N}(\theta_{\hat{k}(n)}, 1)$$

Simple point hypothesis demonstration

Two-sided hypothesis demonstration

Unknown binomial proportion parameter

Observed data $x_1, \ldots, x_n$

$$H_0 : \theta = 0.5$$
$$H_1 : \theta \neq 0.5$$

Suppose I run a survey and observe **168 🍊 and 132 🍏 out of 300** total (56%)

Null hypothesis (H_0)

🍊 $=$ 🍏

$\theta = 0.5$

Alternative hypothesis (H_1)

🍊 $\neq$ 🍏

$\theta \neq 0.5$

Suppose I run a survey and observe **168 🍊 and 132 🍏 out of 300** total (56%)

Null hypothesis (H_0)

🍊 $=$ 🍏

$\theta = 0.5$

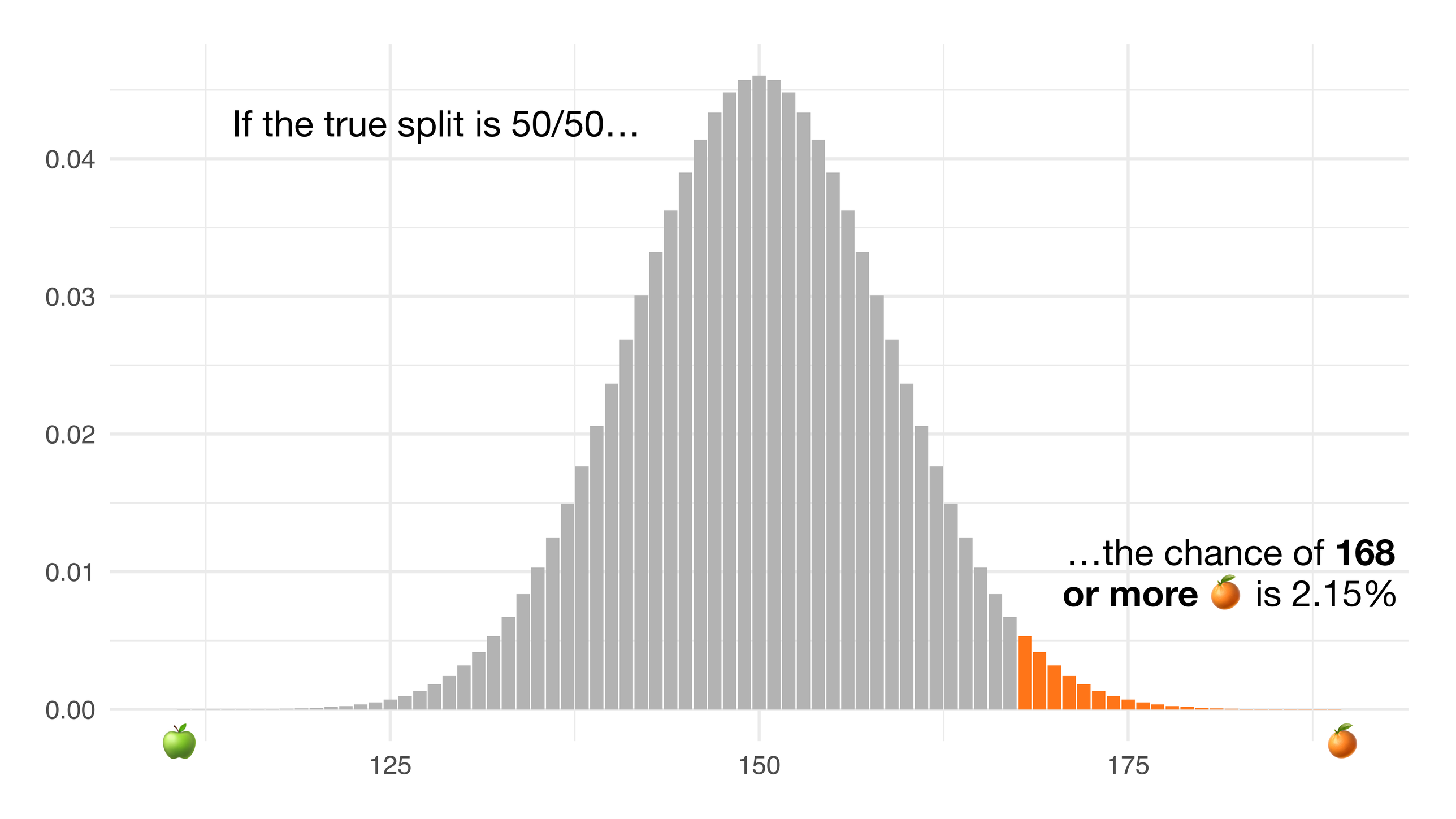
Alternative hypothesis (H_1)

🍊 $\neq$ 🍏

$\theta \neq 0.5$

Frequentist

If the null hypothesis is true, how extreme would my result be if I repeated this experiment several times?


If the true split is 50/50…
0.04
0.03
0.02
0.01
0.00
125
150
175
…the chance of **168 or more** 🍊 is 2.15%

Suppose I run a survey and observe **168 🍊 and 132 🍏 out of 300** total (56%)

Null hypothesis (H_0)

$$🍊 = 🍏$$

$$\theta = 0.5$$

Alternative hypothesis (H_1)

$$🍊 \neq 🍏$$

$$\theta \neq 0.5$$

Frequentist

If the null hypothesis is true, how extreme would my result be if I repeated this experiment several times?

The chance of **168 or more** 🍊 is 2.15%

If it's less than 2.5%/5%, **reject the null hypothesis**

Suppose I run a survey and observe **168 🍊 and 132 🍏 out of 300** total (56%)

Null hypothesis (H_0)

$$🍊 = 🍏$$

$$\theta = 0.5$$

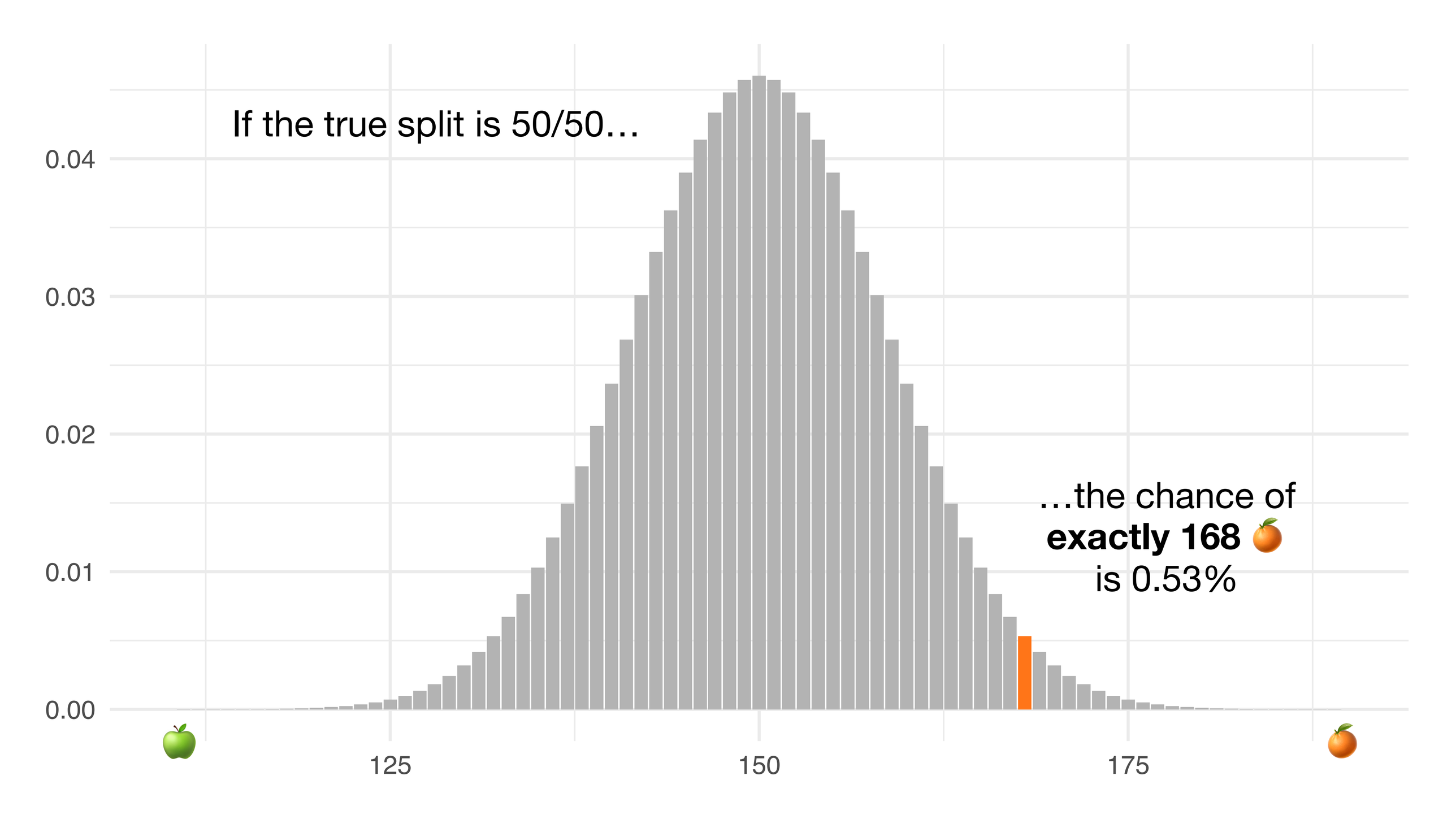
Alternative hypothesis (H_1)

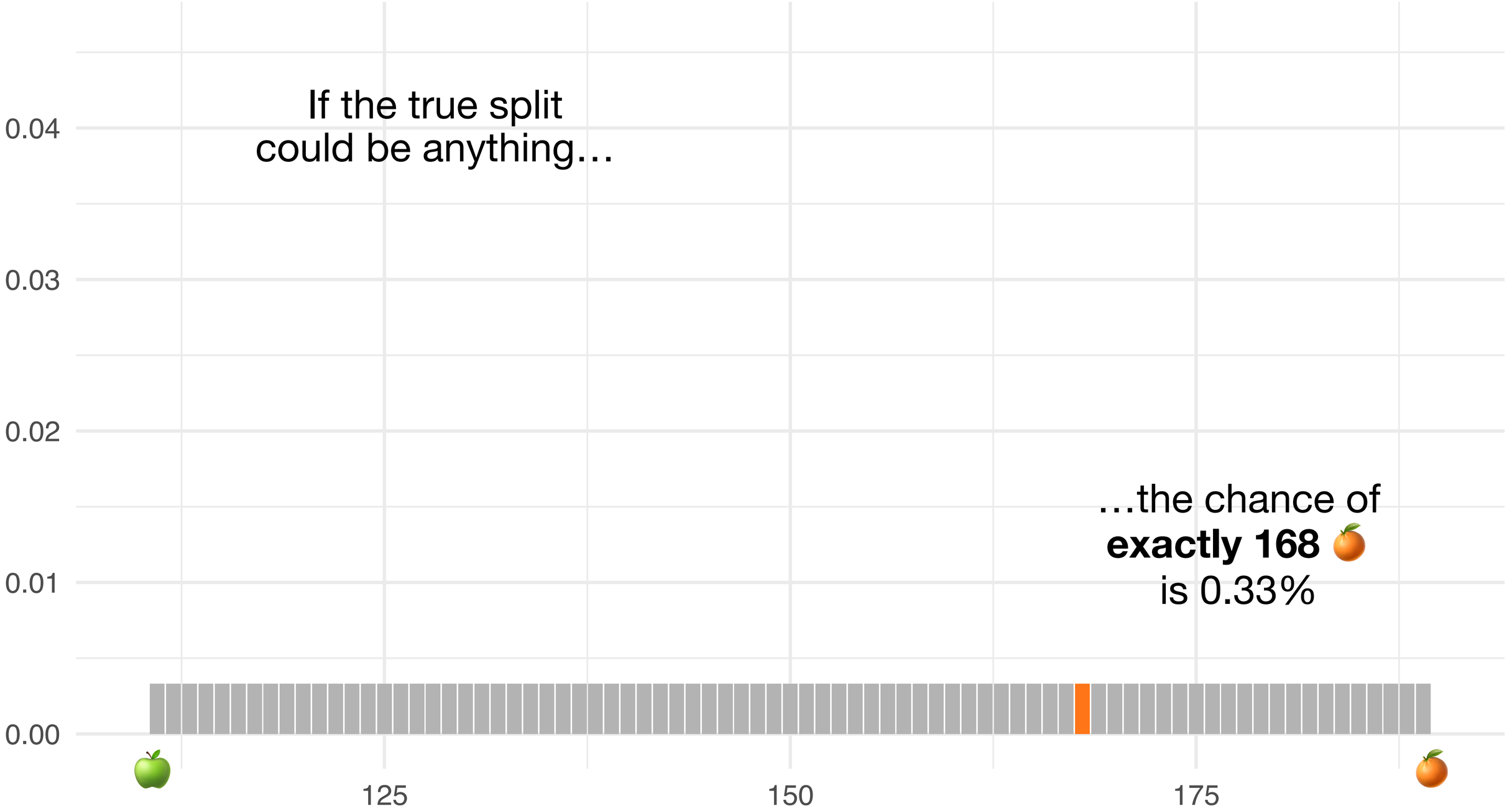
$$🍊 \neq 🍏$$

$$\theta \neq 0.5$$

Bayesian

What is the evidence
for each hypothesis
given the observed data?


If the true split is 50/50…
0.04
0.03
0.02
0.01
0.00
125
150
175
…the chance of
exactly 168 🍊
is 0.53%


If the true split
could be anything…
…the chance of
exactly 168 🍊
is 0.33%
0.04
0.03
0.02
0.01
0.00
125
150
175
🍏
🍊

Suppose I run a survey and observe **168 🍊 and 132 🍏 out of 300** total (56%)

Null hypothesis (H_0)

$$🍊 = 🍏$$

$$\theta = 0.5$$

Alternative hypothesis (H_1)

$$🍊 \neq 🍏$$

$$\theta \neq 0.5$$

The ratio of these values is 61.6%, so we find that there is **greater evidence for the null hypothesis**

What is the evidence for each hypothesis given the observed data?

Suppose I run a survey and observe **168 🍊 and 132 🍏 out of 300** total (56%)

Null hypothesis (H_0)

$$🍊 = 🍏$$

$$\theta = 0.5$$

Alternative hypothesis (H_1)

$$🍊 \neq 🍏$$

$$\theta \neq 0.5$$

The “paradox” comes from the **prior** — we’ve implicitly assumed that a single value ($\theta = 0.5$) is just as likely as any value between 0 and 1

What is the evidence for each hypothesis given the observed data?

Two-sided hypothesis demonstration

Unknown binomial proportion parameter

Observed data $x_1, \ldots, x_n$

$$H_0 : \theta = 0.5$$
$$H_1 : \theta \neq 0.5$$

$$\min_{H_k} \text{BIC} = -2\widehat{\mathscr{L}}(x_{1:n} \mid H_k) + d_k \log(n)$$

Two-sided hypothesis demonstration

Unknown binomial proportion parameter

Observed data $x_1, \ldots, x_n$

$$H_0 : \theta = 0.5$$
$$H_1 : \theta \neq 0.5$$

$$\min_{H_k} \text{BIC} = -2\widehat{\mathscr{L}}(x_{1:n} \mid H_k) + d_k \log(n)$$

$\widehat{\mathscr{L}}$ at $\theta = 0$ and $d_k = 0$ $\longrightarrow$ $x_{n+1} \sim \text{Bernoulli}(0.5)$

Two-sided hypothesis demonstration

Unknown binomial proportion parameter

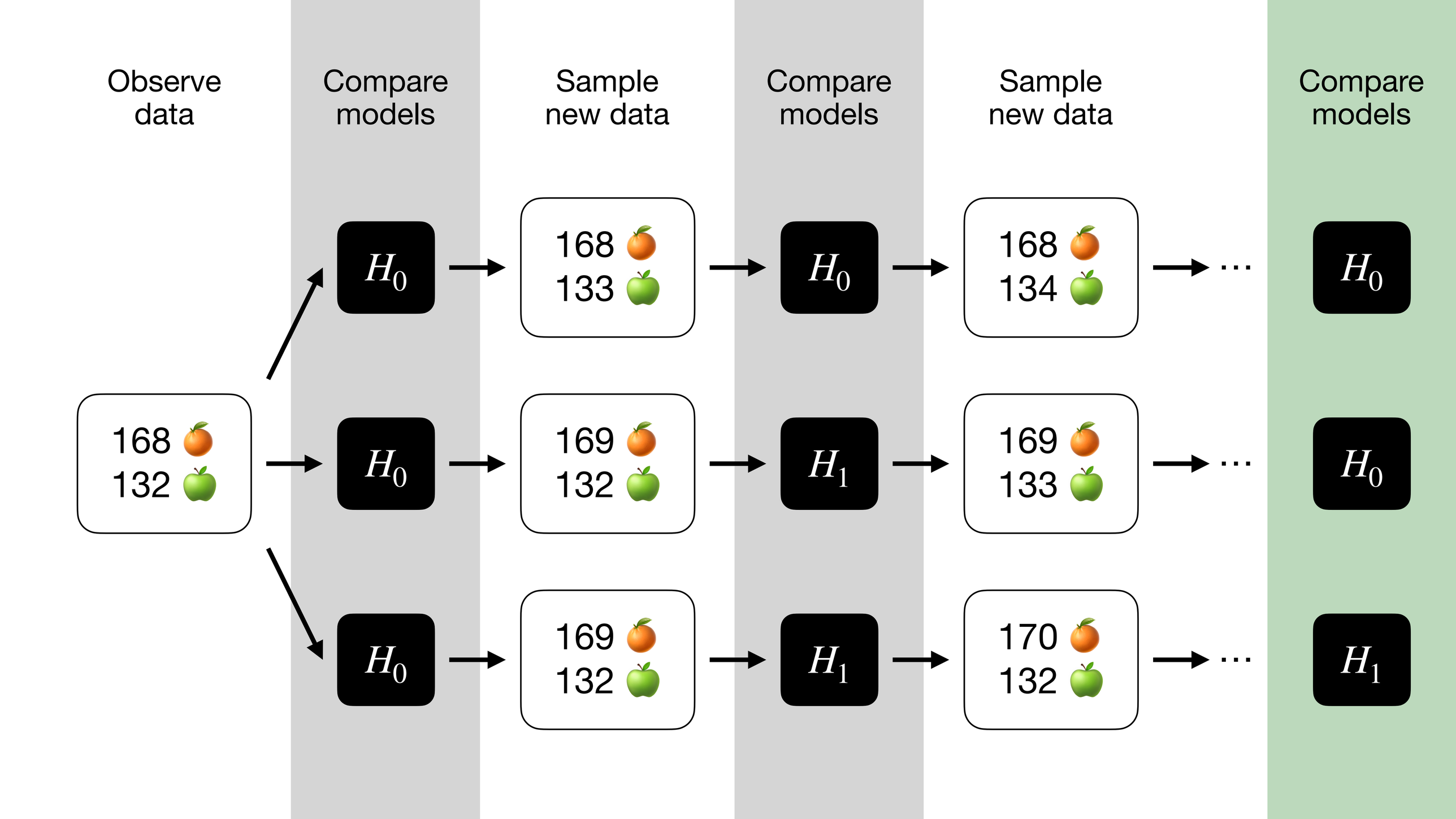
Observed data $x_1, \ldots, x_n$

$$H_0 : \theta = 0.5$$

$$H_1 : \theta \neq 0.5$$

$$\min_{H_k} \text{BIC} = -2\widehat{\mathscr{L}}(x_{1:n} \mid H_k) + d_k \log(n)$$

$\widehat{\mathscr{L}}$ at $\theta = \bar{x}_n$ and $d_k = 1$ $\longrightarrow$ $x_{n+1} \sim \text{Bernoulli}(\bar{x}_n)$


Observe data
Compare models
Sample new data
Compare models
Sample new data
Compare models
168 🍊
132 🍏
H_0
168 🍊
133 🍏
H_0
168 🍊
134 🍏
…
H_0
H_0
169 🍊
132 🍏
H_1
169 🍊
133 🍏
…
H_0
H_0
169 🍊
132 🍏
H_1
170 🍊
132 🍏
…
H_1

Two-sided hypothesis demonstration

The final probabilities are **76.9% for the null** and **23.1% for the alternative**

Results have a prior-free interpretation in terms of posterior uncertainty

Evidence can be accumulated in favor of the null

Thank you!

“A general framework for probabilistic model uncertainty” (Shirvaikar, Walker, and Holmes)

https://arxiv.org/abs/2410.17108