MODEL CHECKING

BST228

Speaker notes

No notes on this slide.

ELECTIONS CAN BE STRESSFUL

https://www.hsph.harvard.edu/student-affairs/

Speaker notes

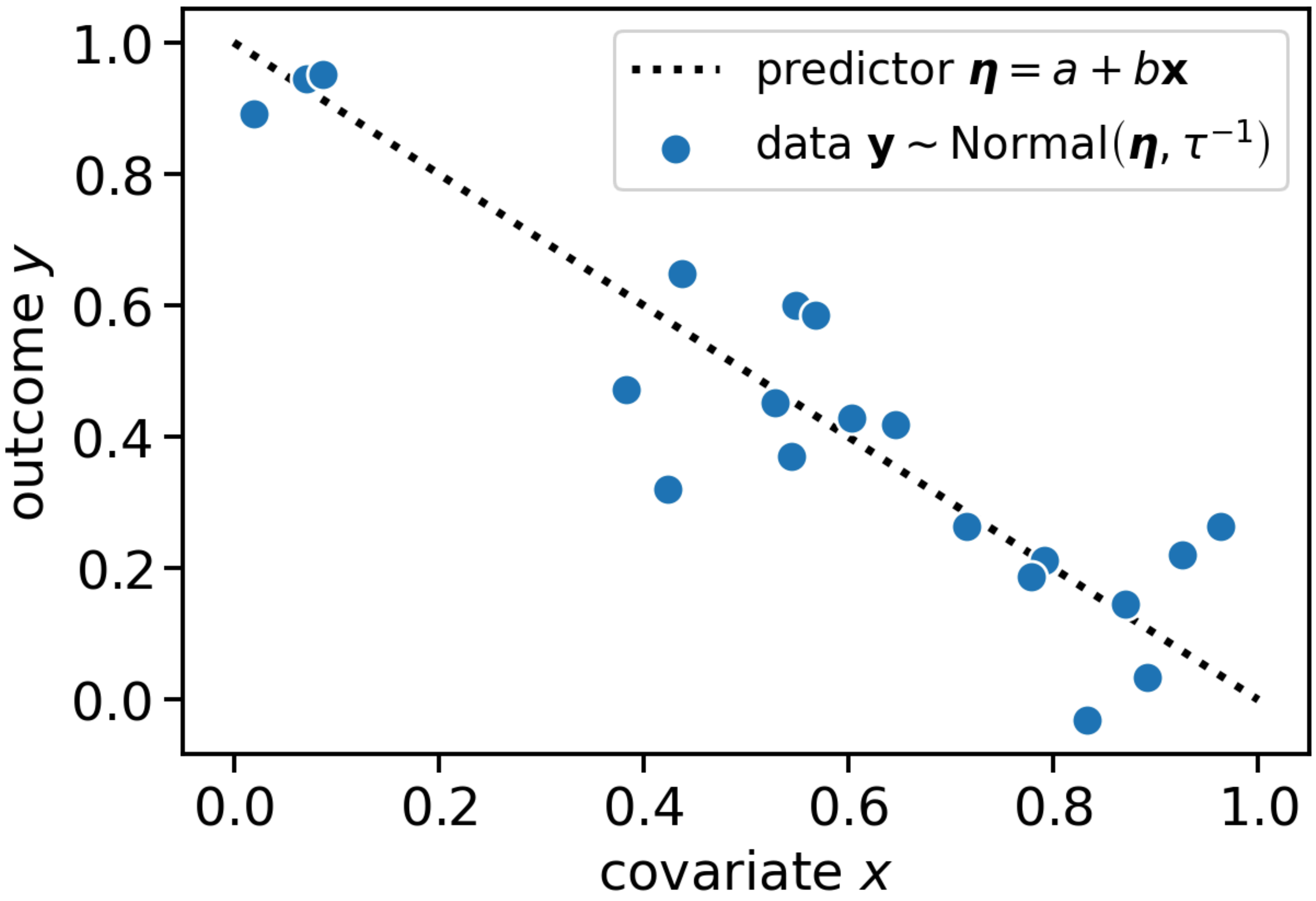
- Elections and the accompanying uncertainty can be stressful and at times scary.
- The office for student affairs has resources if you're finding this time difficult.

TODAY

- A bit more on Bayesian model averaging.
- Checking Bayesian models.
- In-class exercise on Bayesian model averaging.

Speaker notes

- Bayesian model averaging is very appealing in theory, but there can be some practical challenges.
- Before we can draw conclusions using our analysis, we need to check our models and assess if they capture the main aspects of the data.
- There will be time to work on the Bayesian model averaging exercise for the lab.



Speaker notes

- Consider the simplest of regression models. In this example, this is the *true* model with intercept a, slope b, and observation precision τ.
- We will compare it with a competing model that does not include the slope b, i.e., a model that just estimates the mean outcome.

MODEL 1

$$\mathbf{y} \mid \mathbf{x}, a, b, m_1 \sim \mathsf{Normal}\left(a, \tau^{-1}\right)$$
$$a \mid m_1 \sim \mathsf{Normal}\left(0, \kappa_a^{-1}\right)$$

MODEL 2

$$\mathbf{y} \mid \mathbf{x}, a, b, m_2 \sim \mathsf{Normal}\left(a + b\mathbf{x}, \tau^{-1}\right)$$
$$a \mid m_2 \sim \mathsf{Normal}\left(0, \kappa_a^{-1}\right)$$
$$b \mid m_2 \sim \mathsf{Normal}\left(0, \kappa_b^{-1}\right)$$

Speaker notes

- We consider two different models denoted by conditioning on model index m_1 without a slope or model index m_2 with a slope b.
- For m_1, we only have a single prior for the intercept.
- For m_2, we have a second prior for the slope.

We want to compare the marginal likelihood $p\left(\mathbf{y} \mid \mathbf{x}, m_1\right)$ for the first model with the marginal likelihood $p\left(\mathbf{y} \mid \mathbf{x}, m_2\right)$ for the second.

Speaker notes

- The marginal likelihood is called “marginal” because it integrates out the parameters $\boldsymbol{\theta}$ of the model, i.e., $p\left(\mathbf{y} \mid \mathbf{x}, m\right) = \int d\boldsymbol{\theta}\, p\left(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{x}, m\right) p\left(\boldsymbol{\theta} \mid m\right)$.
- Higher marginal likelihoods indicate that the observed data are plausible under the model.

To avoid repeating cumbersome algebra for the two models, we consider the general model

$$\begin{aligned} \mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta} &\sim \text{Normal}\left(\mathbf{X}\boldsymbol{\theta}, \tau^{-1}\right) \\ \theta &\sim \text{Normal}\left(\mathbf{0}, \boldsymbol{\kappa}_0^{-1}\right) \end{aligned}$$

for design matrix $\mathbf{X}$, regression coefficients $\boldsymbol{\theta}$, and prior precision matrix $\boldsymbol{\kappa}_0$.

Speaker notes

- Both models can be formulated as a standard linear regression model using different features.
- We will use this general formulation to evaluate the marginal likelihood of the data $\boldsymbol{y}$ under the two competing models.

MODEL m_1

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\boldsymbol{\theta} = \begin{pmatrix} a \end{pmatrix}^\intercal$$

MODEL m_2

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$\boldsymbol{\theta} = \begin{pmatrix} a & b \end{pmatrix}^\intercal$$

Speaker notes

- Both models can be formulated as a general linear regression problem.
- For the first, we use a design matrix of ones such that $\mathbf{X}\boldsymbol{\theta} = \begin{pmatrix} a & \dots & a \end{pmatrix}^\intercal$.
- For the second, we include the covariates $\mathbf{x}$ such that $\mathbf{X}\boldsymbol{\theta} = \begin{pmatrix} a + bx_1 & \dots & a + bx_n \end{pmatrix}^\intercal$.

The marginal likelihood is

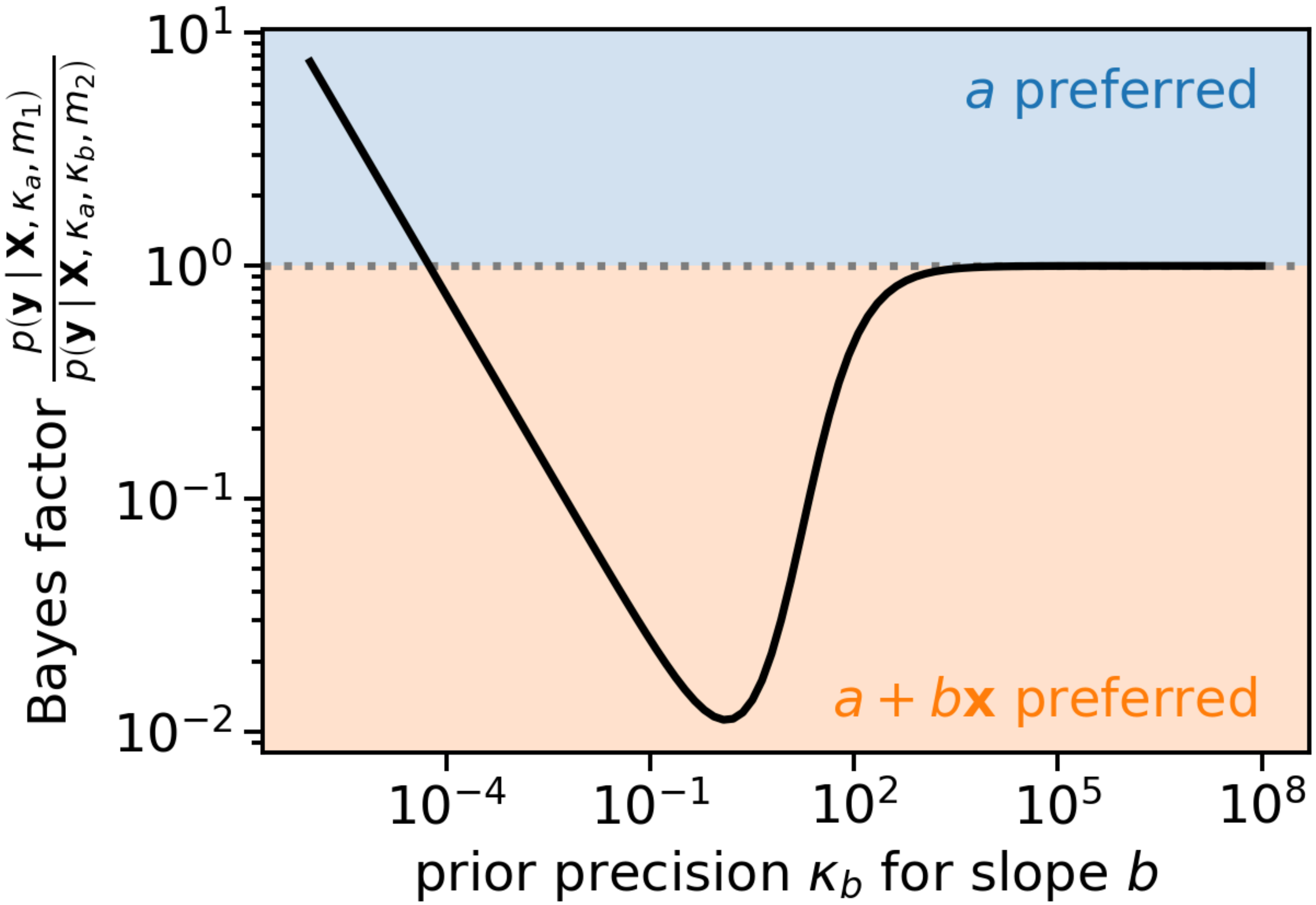
$$p\left(\mathbf{y}\mid\mathbf{X},\boldsymbol{\kappa}_0\right)=\left(\frac{\tau}{2\pi}\right)^{n/2}\sqrt{\frac{\left|\boldsymbol{\kappa}_0\right|}{\left|\boldsymbol{\kappa}_n\right|}}$$

$$\times\exp\left(\frac{1}{2}\left[\boldsymbol{\nu}_n^\top\boldsymbol{\kappa}_n\boldsymbol{\nu}_n-\tau\mathbf{y}^\top\mathbf{y}\right]\right)$$

for posterior precision $\boldsymbol{\kappa}_n=\boldsymbol{\kappa}_0+\tau\mathbf{X}^\top\mathbf{X}$ and posterior mean $\boldsymbol{\nu}_n=\tau\boldsymbol{\kappa}_n^{-1}\boldsymbol{X}^\top\mathbf{y}$.

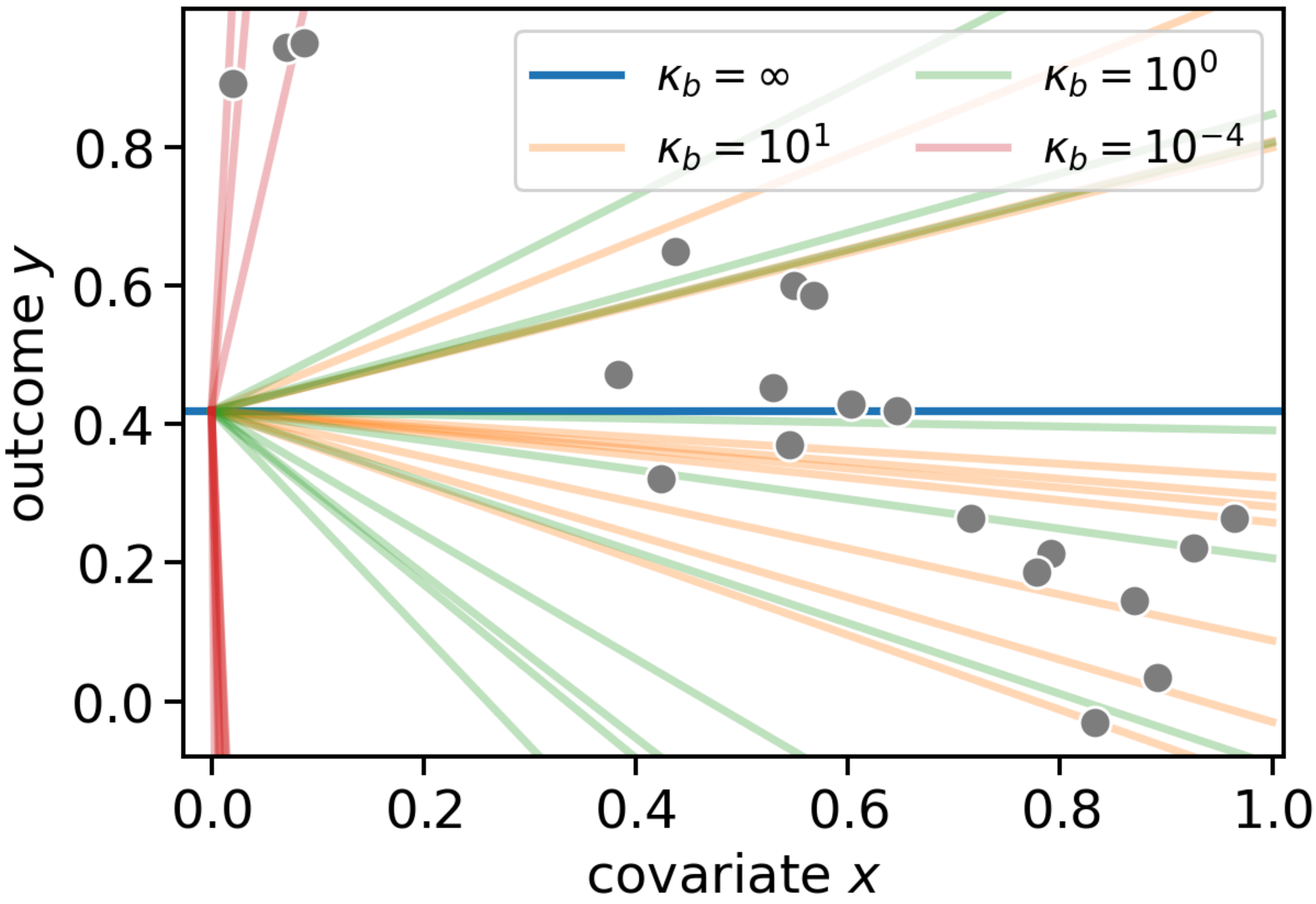
Speaker notes

- Deriving the marginal likelihood requires integrating out the regression coefficients $\boldsymbol{\theta}$ and is tedious. The details are omitted, but it is instructive to derive $p\left(\mathbf{y}\mid\mathbf{X},\boldsymbol{\kappa}_0\right)$ yourself.
- The first factor is due to n i.i.d. observations in the likelihood.
- The second is the ratio of how precisely we know the regression parameters a priori vs a posteriori. Remember that the determinant is the product of eigenvalues, and it expresses an overall notion of precision for multivariate parameters.
- The second term in the exponential captures overall variance in the data and is not relevant for comparison because it does not depend on the model.
- The first in the exponential is somewhat difficult to interpret, however.



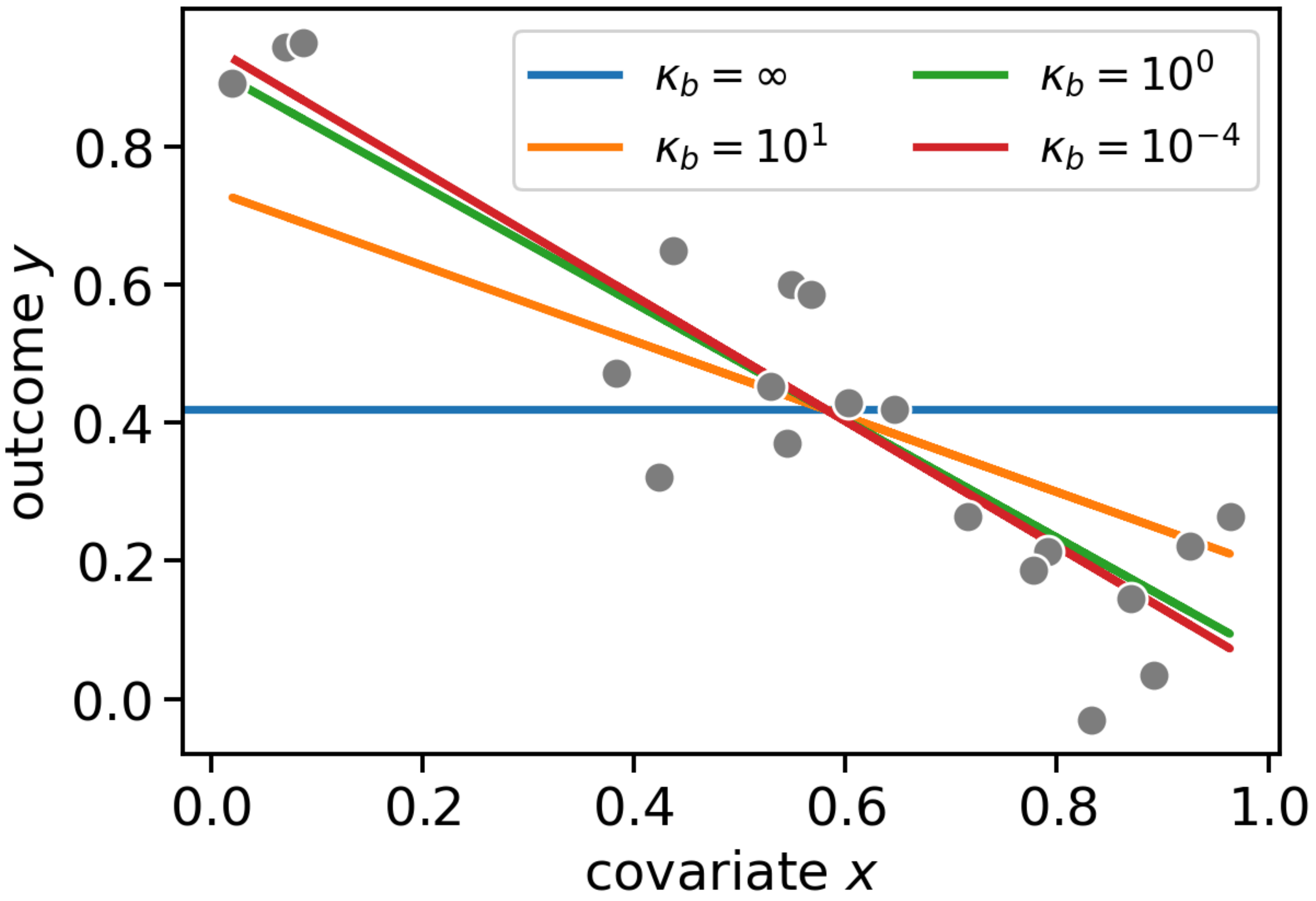
Speaker notes

- Figure shows the Bayes factor with values greater than 1 favoring the simpler model plotted against the prior precision κ_b for the slope b. We set $\kappa_a = 10^{-2}$.
- For very large κ_b, our prior belief is that $b \approx 0$, and there is no difference between the two models. The Bayes factor is 1.
- For very small κ_b, our prior belief is that b varies wildly. *A priori*, the predictions are terrible because a slope of 100 is not implausible.
- Then there is a “just right” value for the prior precision where the more complex model is better. The prior is neither too restrictive nor too permissive.



Speaker notes

- The figure shows a few samples of linear predictors for different priors.
- For $\kappa_b \to \infty$, we recover the first model with only the overall mean a. This model is not flexible enough to capture the data well.
- For small prior precision κ_b, the a priori magnitude of regression coefficients is very large. This model is flexible enough but gives terrible predictions a priori, e.g., $a + bx > 100$ at $x = 1$ for some samples for $\kappa_b = 10^{-4}$.



Speaker notes

- The posterior tells a different story, however, because the fit improves for smaller κ_b and asymptotes to the MLE in the limit $\kappa_b \to 0$.
- Nevertheless, the marginal likelihood approaches zero, and the Bayes factor diverges, favoring the simpler model.
- Why does that happen? The marginal likelihood quantifies how well the model fits *a priori*, i.e., before actually having been fit to data.
- The prior really matters for the marginal likelihood, and non-informative priors will always aggressively favor simple models.
- So is Bayesian model comparison using marginal likelihoods or Bayes factors useful? Maybe; it depends on if you can formulate reasonable priors.

The posterior predictive distribution of *any* quantity ϕ in light of data $\mathbf{y}$ is

$$\begin{aligned} p\left(\phi \mid \mathbf{y}\right) &= \sum_m p\left(\phi \mid \mathbf{y}, m\right) p\left(m \mid \mathbf{y}\right) \\ &= \sum_m p\left(\phi \mid \mathbf{y}, m\right) \frac{p\left(\mathbf{y} \mid m\right) p\left(m\right)}{p\left(\mathbf{y}\right)} \\ &= \frac{1}{p\left(\mathbf{y}\right)} \sum_m p\left(\phi \mid \mathbf{y}, m\right) p\left(\mathbf{y} \mid m\right) p\left(m\right). \end{aligned}$$

Speaker notes

- This matters greatly for Bayesian model averaging because the weight assigned to each model is proportional to the marginal likelihood $p\left(\mathbf{y} \mid m\right)$ of the model m.
- If we take the common approach of choosing non-informative priors for a model, it will have zero weight.
- Bayesian model averaging only works for proper priors; it breaks for improper priors.
- But even if priors are proper, the weights are heavily influenced by prior choice.
- In the equations to the left, the first equality follows by the law of total probability, the second by Bayes theorem, and the third because $p\left(\mathbf{y}\right)$ is independent of the model index m.

MODEL CHECKING

1. Collect data.
2. Formulate model.
3. Fit model.
4. **Check model.**
5. Draw conclusions.

Speaker notes

- With that aside, how would you go about checking, extending, or comparing models?
- Apart from Bayesian model averaging, we have largely brushed 4. under the carpet.

OPTIONS

- Posterior predictive replication.
- Posterior prediction for other data and cross-validation.
- …

Speaker notes

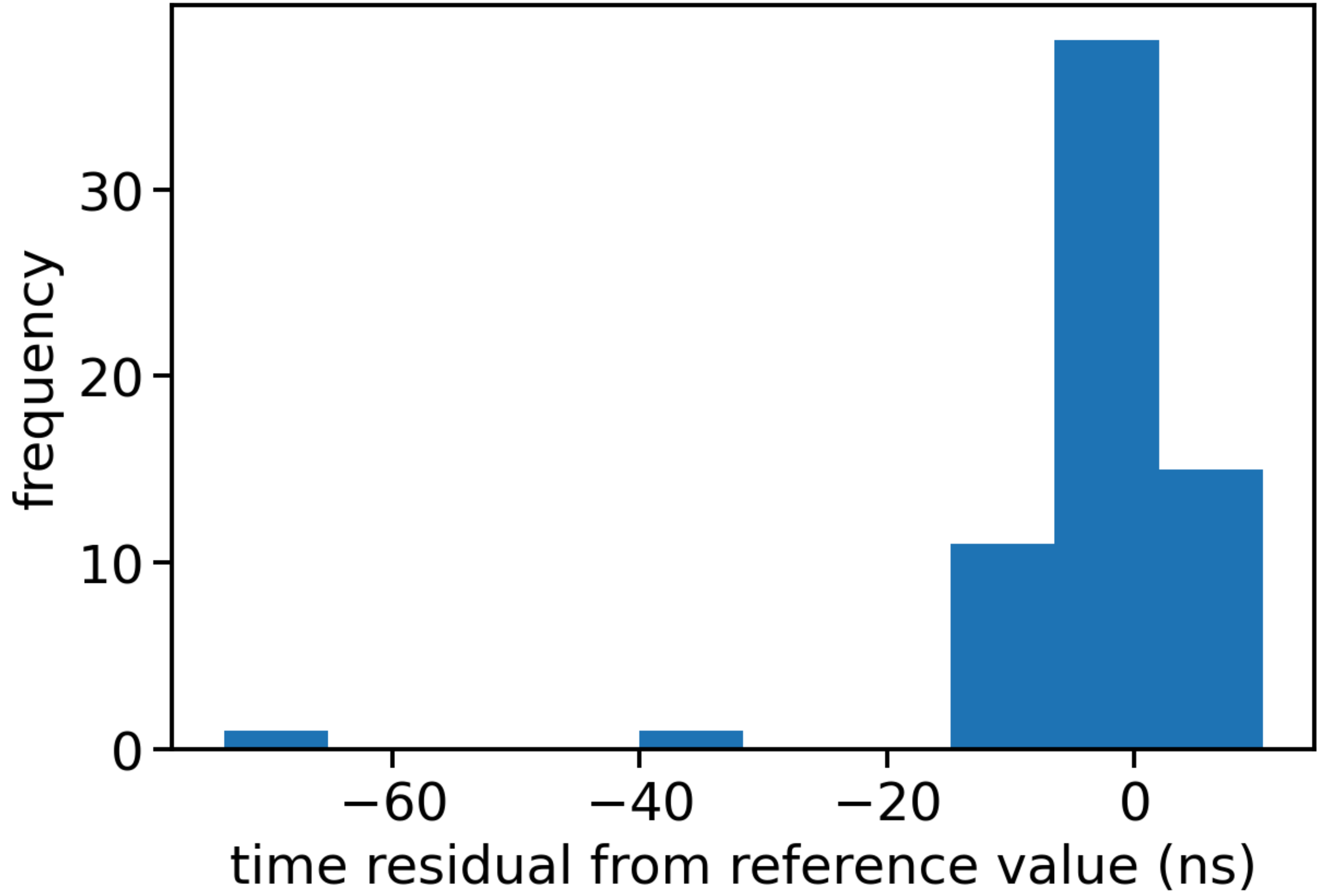
- In light of the difficulties with the marginal likelihood, how can we compare different models in a way consistent with our intuition? E.g., we would like comparison to be robust to the prior in the limit of large data.
- Replication generates hypothetical data *after* having fit the model. As a minimum, the model should be able to reproduce the data.
- Posterior prediction can, for example, reproduce other studies in silico, and we can compare summary statistics.
- Here, we focus on posterior predictive replication.

EXAMPLE: NEWCOMB’S LIGHT SPEED MEASUREMENTS

Data $\mathbf{y}$ are $n = 66$ measurements of the time of flight of light between mirrors taken atop Mount Washington.

Speaker notes

- Measuring the speed of light was all the rage in the late 19th, early 20th century.
- Given the time of flight and distance between the mirrors, we can estimate the speed of light.



Speaker notes

- The figure shows a histogram of 66 measurements as deviations from the reference value we know today in nanoseconds.
- They did amazingly well in terms of the measurement. But there are some outliers suggesting faster-than-light travel between the mirrors.

```stan
data {
    int n;
    vector [n] y;
}

parameters {
    real mu;
    real<lower=0> sigma;
}

model {
    y ~ normal(mu, sigma);
    mu ~ normal(0, 100);
    sigma ~ exponential(1e-2);
}
```

Speaker notes

- To run replication, we first need to build a model. We use a standard normal model $\mathbf{y} \sim \text{Normal}\left(0, \sigma^2\right)$ for n observations $\mathbf{y}$ with unknown μ and σ.
- We use a vaguely informative prior for μ (deviations much larger than 100ns are unlikely) and σ (variation much larger than 100ns are unlikely; remember that the argument of `exponential` is the decay rate, hence `1e-2`).

POSTERIOR PREDICTIVE REPLICATION

Having fitted the model, we can replicate the data by sampling from the posterior predictive distribution

$$p\left(\mathbf{y}^{\mathrm{repl}} \mid \mathbf{y}\right) = \int d\boldsymbol{\theta}\, p\left(\mathbf{y}^{\mathrm{repl}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right).$$

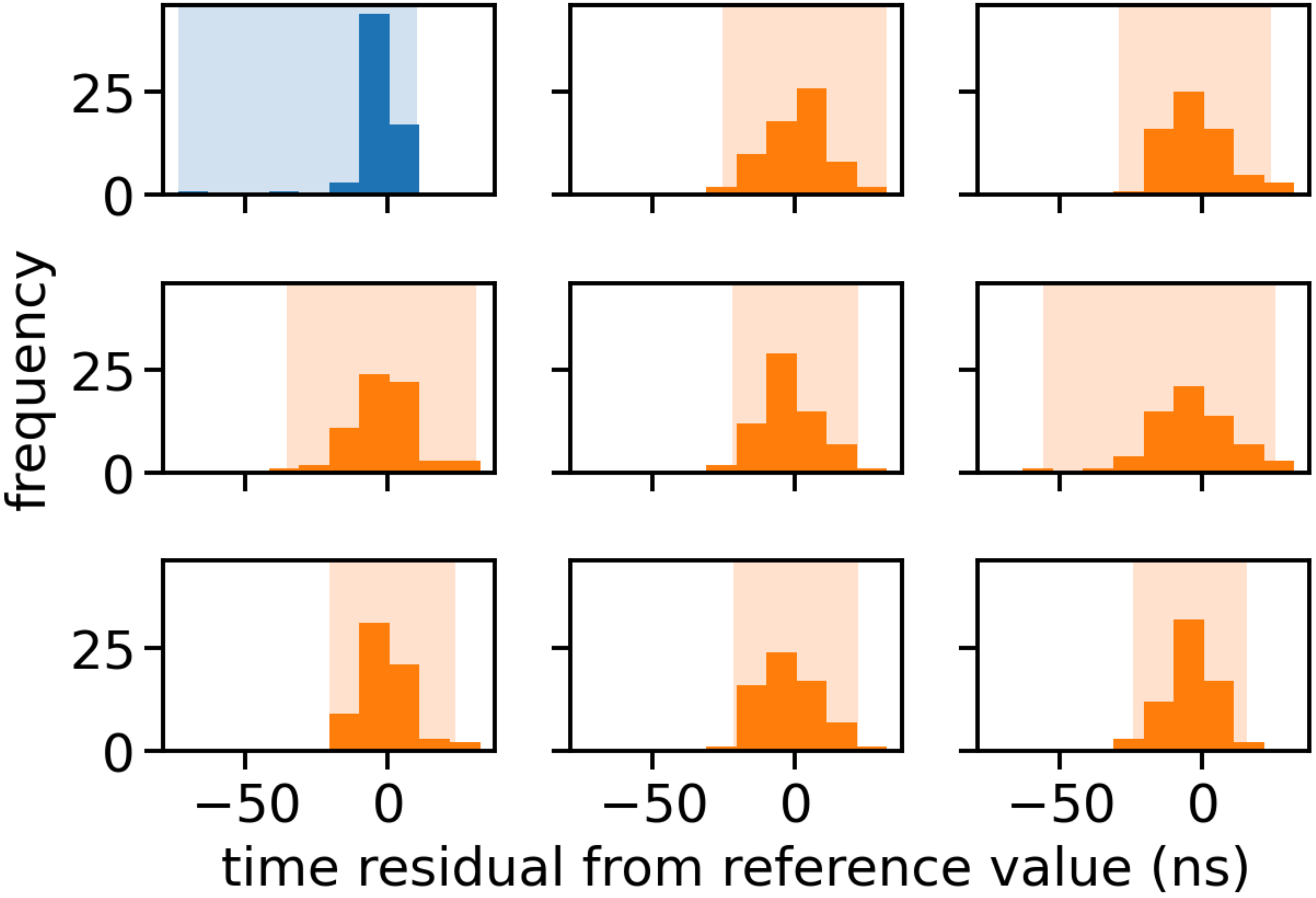
Speaker notes

- How does replication work in practice? We first fit he model to get posterior samples and, second, sample from the likelihood for each of the posterior samples.
- The main idea is that if we fit the model and generate predictions, then those predictions should look a lot like the data we fitted to.
- This is in some sense double dipping. We assess the model using the same data we used to fit it.
- That's why we can only ever use replication to test if the model does *not* fit rather than being able to confirm that it is the “right” model.

```stan
generated quantities {
    vector [n] y_repl;
    for (i in 1:n) {
        y_repl[i] = normal_rng(mu, sigma);
    }
}
```

Speaker notes

- Implementing replication has the same structure as the model. However, we need to explicitly loop over the n observations to sample them using `normal_rng`.
- We could also run the replication outside the Stan program in R or Python, but it is often nice to have everything in the same place.
- Aside: Other probabilistic programming languages like `numpyro` let you re-use the model definition to run posterior replications and posterior predictions without having to implement a `generated quantities` block.



Speaker notes

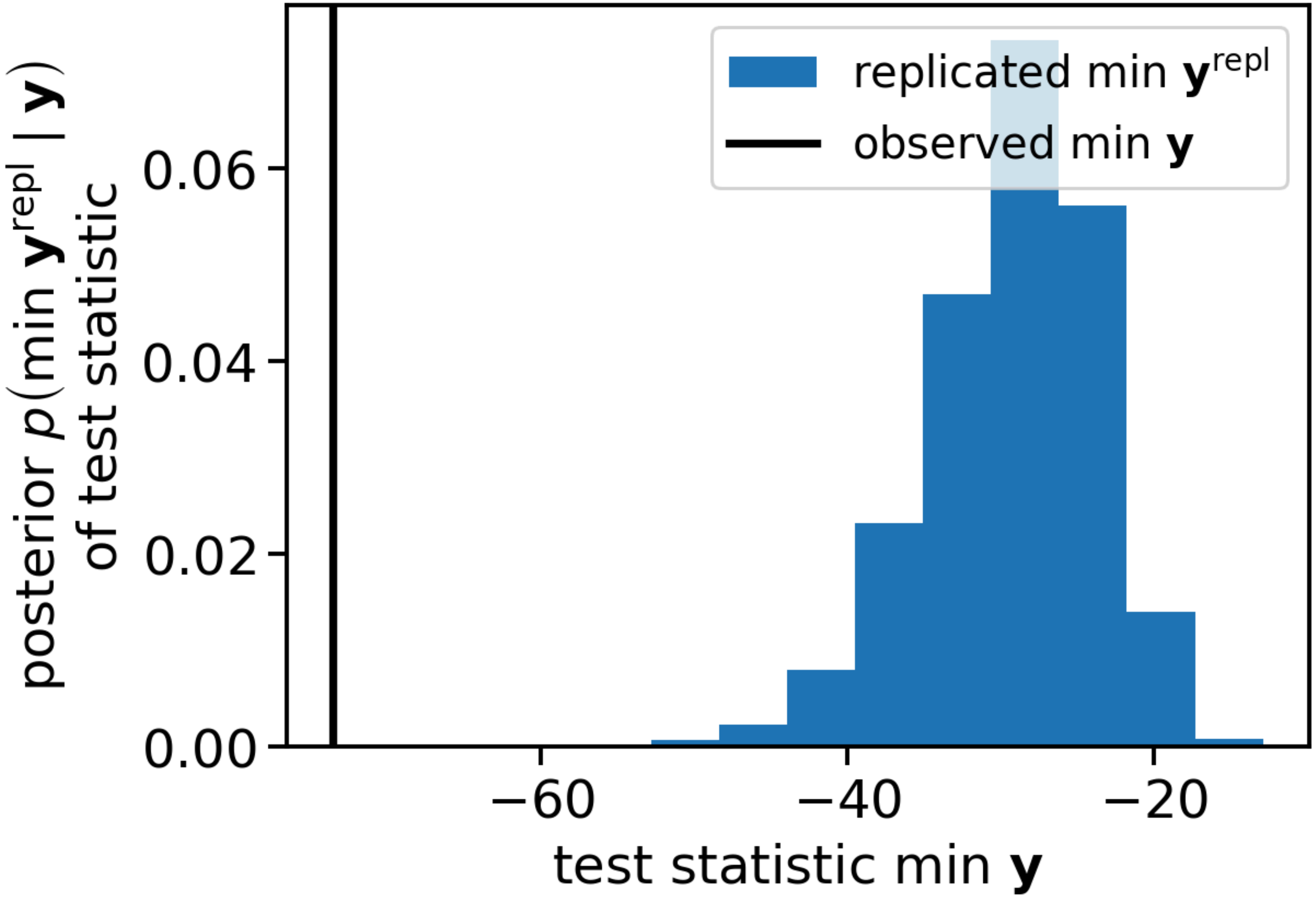
- The first panel shows a histogram of the measured data. The shaded region is the range, i.e., min to max.
- Subsequent panels show histograms of replicated data. We observe smaller ranges and less skew than in the measured data.
- The simple normal model cannot replicate some features of the data.

Visually assessing similarity between observed and replicated data is challenging. We define a test statistic $t\left(\mathbf{y}\right)$ to summarize salient aspects of the data.

For Newcomb’s measurements, we consider $t\left(\mathbf{y}\right)=\mathbf{min}\,\mathbf{y}$ because the smallest value seems like an outlier.

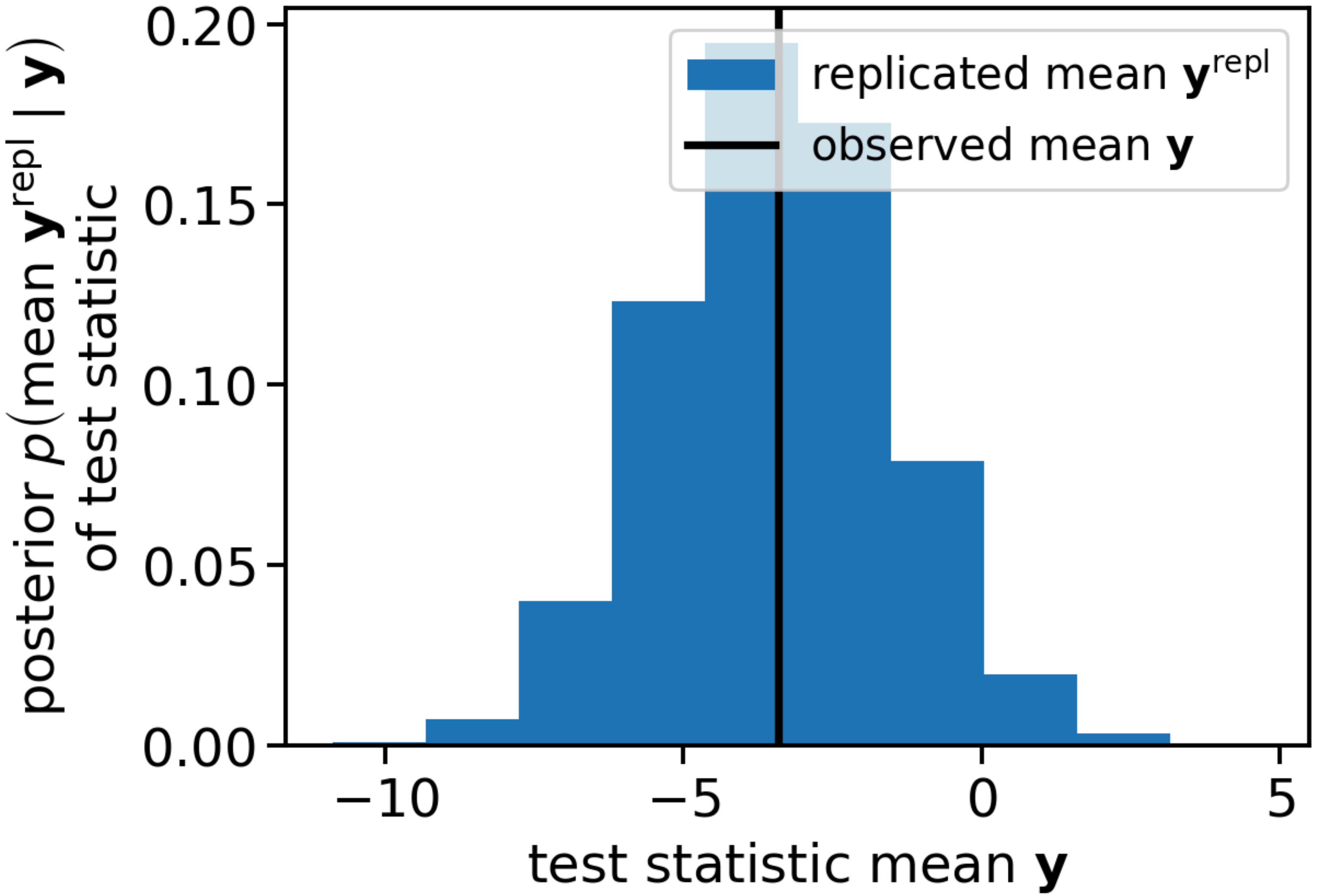
Speaker notes

No notes on this slide.



Speaker notes

- The figure shows a histogram of the minimum value of replicated data. Each sample contributing to the histogram is an independent replication of the data. The black vertical line is the minimum of the observed data.
- Replicated data here are inconsistent with the observed data, indicating that our model cannot capture *this particular aspect* of the data.



Speaker notes

- Consider instead replicating the mean of the data. This analysis does not suggest that there is anything wrong with the model.
- Why? Because fitting a normal model to data will faithfully capture the mean. Test statistics can only inform how well the model captures a specific aspect of the data.
- Ideally, these test statistics are something that is not directly captured by the model. E.g., the sample mean for a normal model or the fraction of successes for a Bernoulli model.
- This is a great opportunity to get your collaborators involved. They likely have a hunch for important aspects of the data that your model has to be able to capture. Together, you can translate these aspects into test statistics for model checking.

EXAMPLE: COIN FLIPS

$$\mathbf{y} = (1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0)^{\mathsf{T}}$$

Speaker notes

- Assuming independent coin flips (heads = 1, tails = 0), we can use a binomial model for the number of heads or independently consider each flip using a Bernoulli model. The resulting inference for the bias of the coin θ will be the same.
- However, if the data correspond to a sequence of observations, our hypothesis of independent samples may not be appropriate. There are long runs of 0s and 1s.

INDEPENDENT COIN FLIPS

$$\begin{aligned} y_i \mid \theta &\sim \text{Bernoulli}\left(\theta\right) \\ \theta &\sim \text{Uniform}\left(0, 1\right) \end{aligned}$$

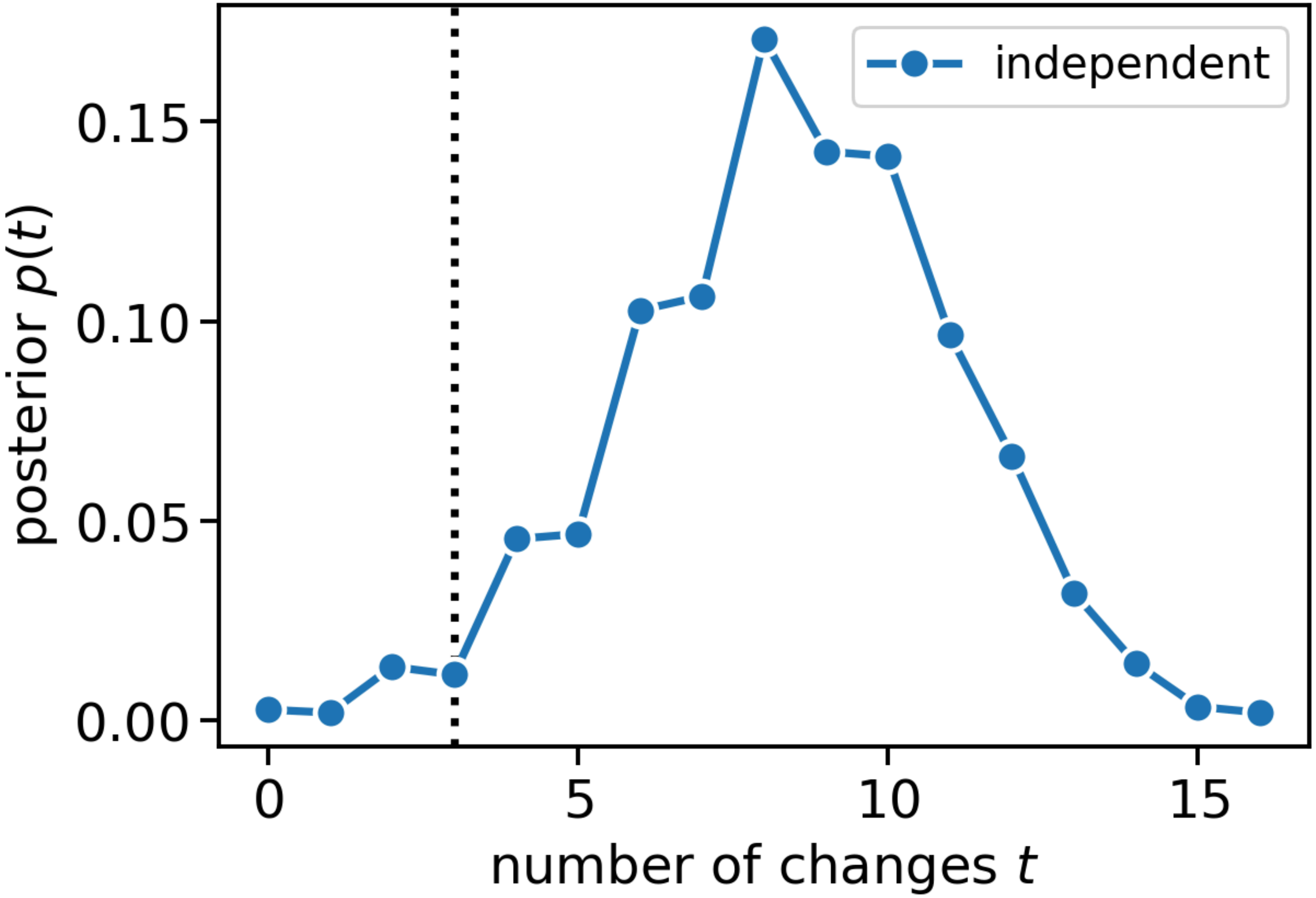
Speaker notes

- We first consider the standard independent coin flip model using a Bernoulli likelihood.

```
data {
    int n;
    array [n] int y;
}

parameters {
    real<lower=0, upper=1> theta;
}

model {
    y ~ bernoulli(theta);
}

generated quantities {
    array [n] int y_repl;
    for (i in 1:n) {
        y_repl[i] = bernoulli_rng(theta);
    }
}
```

Speaker notes

- We use a standard coin-flip model. We used an implicit uniform prior for `theta`.
- Just like in the light speed measurement example, we use `generated quantities` to replicate the data.



Speaker notes

- We use the number of changes from heads to tails or tails to heads as the test statistic and compare with the replicated results.
- The number of replicated changes far exceed the number of observed changes, suggesting that our model cannot capture this aspect.

SEQUENTIAL MODEL

$$
\begin{aligned}
y_1 \mid \theta &\sim \text{Bernoulli}\left(\theta\right) \\
y_{i>1} \mid y_{i-1}, \rho &\sim \text{Bernoulli}\left(\rho_{y_{i-1}}\right) \\
\{\theta, \rho_0, \rho_1\} &\sim \text{Uniform}\left(0, 1\right)
\end{aligned}
$$

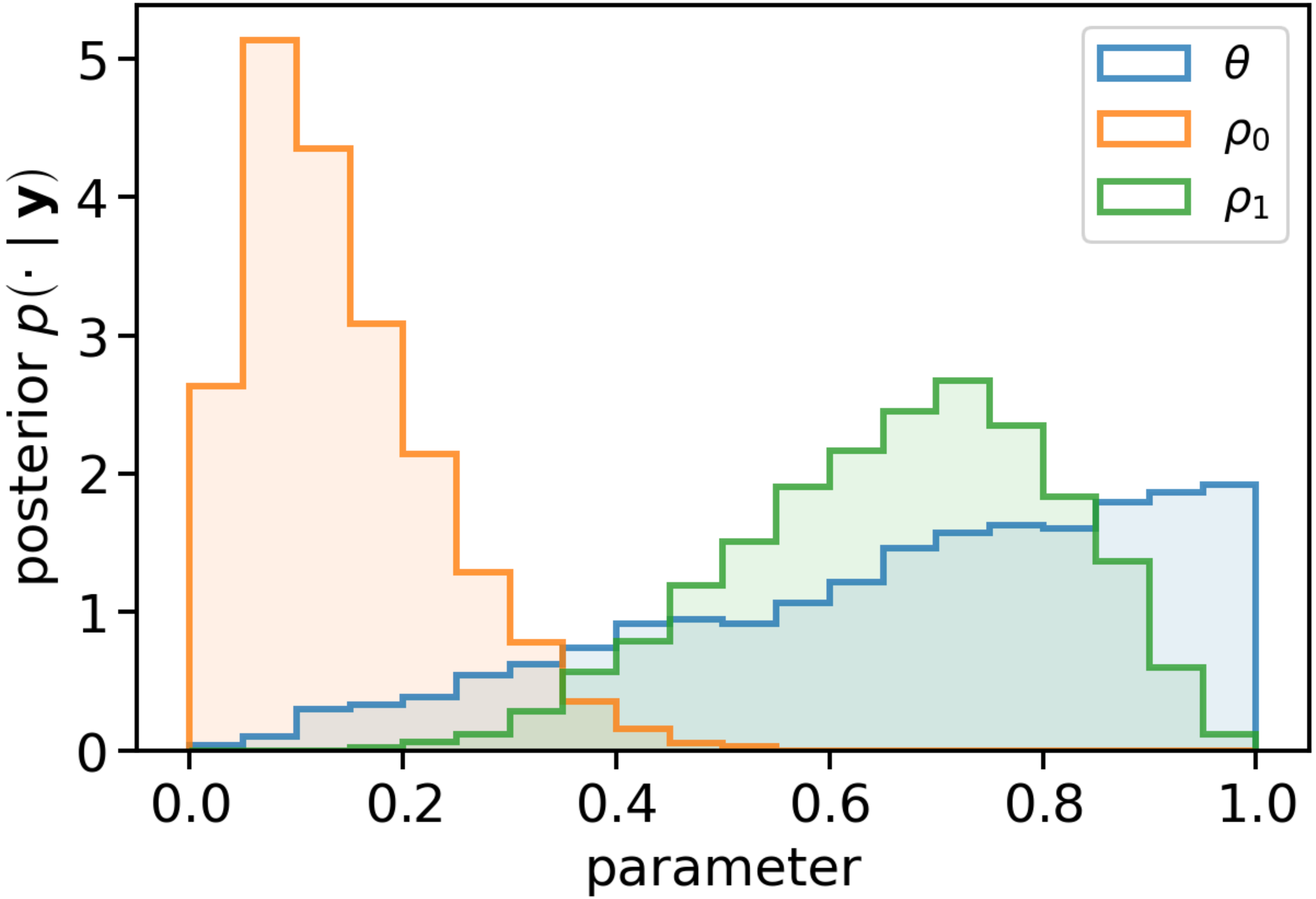
Speaker notes

- We explicitly consider the sequential nature of the data to build an improved model.
- The first flip is heads with probability θ. For all subsequent flips, the outcome is heads with probability ρ_1 if the previous flip was heads and with probability ρ_0 if the previous flip was tails.

```stan
parameters {
    real<lower=0, upper=1> theta, rho0, rho1;
}

model {
    y[1] ~ bernoulli(theta);
    for (i in 2:n) {
        y[i] ~ bernoulli(y[i - 1] ? rho1 : rho0);
    }
}

generated quantities {
    array [n] int y_repl;
    y_repl[1] = bernoulli_rng(theta);
    for (i in 2:n) {
        y_repl[i] = bernoulli_rng(
            y_repl[i - 1] ? rho1 : rho0);
    }
}
```

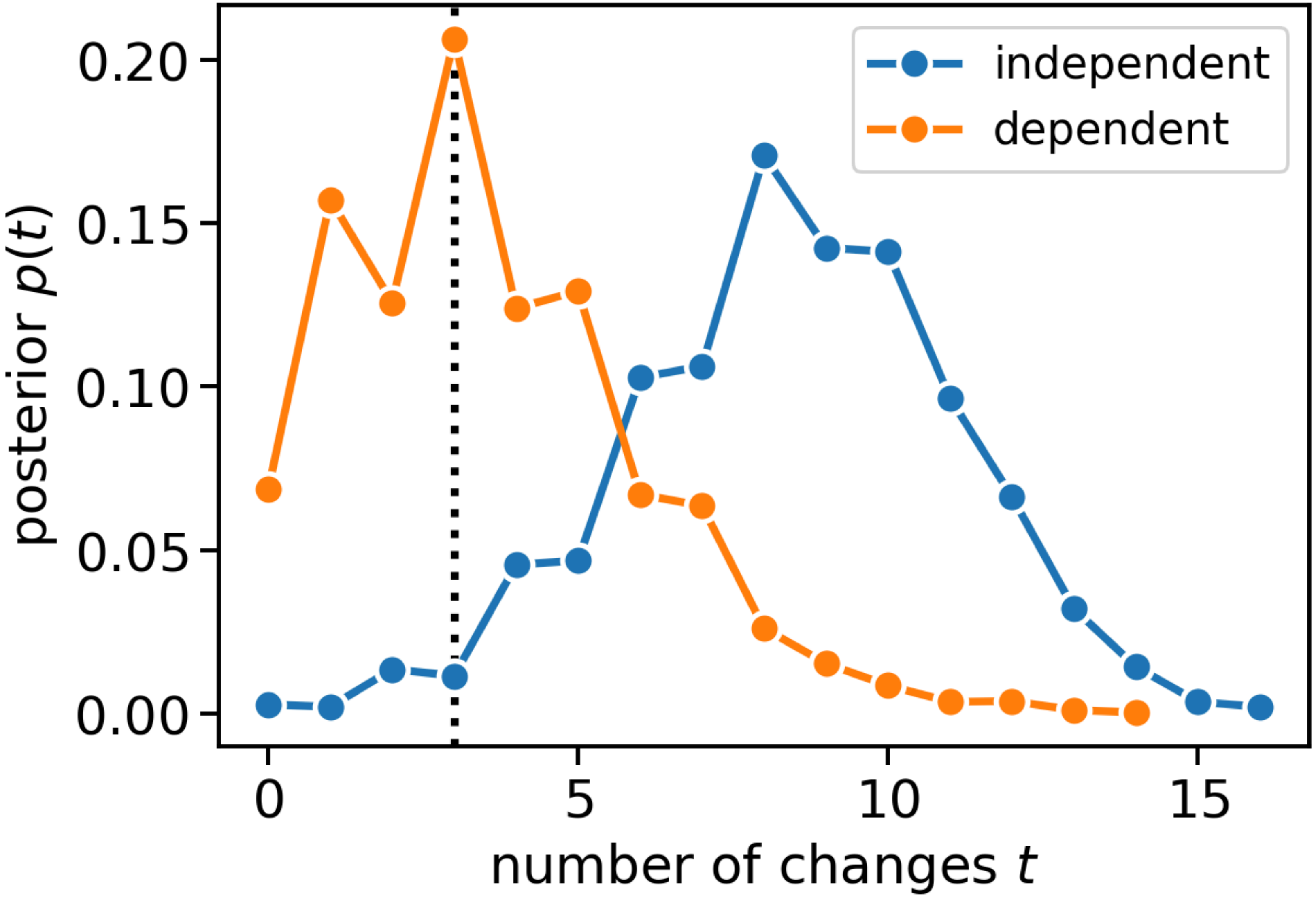
Speaker notes

- The model is slightly more complex, and we need to iterate over the `n` observations to both declare the model and replicate the data.
- We use the ternary operator with syntax borrowed from C. In short, `predicate ? a : b` evaluates to `a` if the predicate satisfied and `b` otherwise.



Speaker notes

- Before considering the replicated data, we investigate the parameters.
- θ is slightly biased towards larger values because the first element in the sequence was heads. However, we only had a single observation such that the uncertainty remains large.
- θ_0 is small, i.e., the probability of heads given tails at the previous step is small. θ_1 is large, and the probability of heads given heads at the previous step is large. These parameters can reproduce the long runs of 0s and 1s.



Speaker notes

- Replicating the number of changes between heads and tails again, we find that our sequential model is much better at capturing this statistic.
- If we had instead replicated the number or proportion of heads, we probably would not have been able to distinguish between the two models. Which test statistic you use to investigate the model really matters.

FREQUENTIST p-VALUES

Frequentist p-value is

$$p_f\left(\hat{\boldsymbol{\theta}}\right) = p\left(t\left(\mathbf{y}\right) \leq t\left(\mathbf{y}^{\mathrm{repl}}\right) \mid \hat{\boldsymbol{\theta}}\right).$$

Speaker notes

- We have still been investigating replications and test statistics visually. This is not feasible if you want to assess the model along many different dimensions.
- Recap: In a frequentist setting, we often consider the tail probability of a test statistic to assess model fit.
- Given an estimate of parameters $\hat{\boldsymbol{\theta}}$ (usually the MLE), we consider many different (hypothetical) realizations of the data–replications–and consider the rank of the observed test statistic among the population of replicated statistics.

POSTERIOR PREDICTIVE p-VALUES

Posterior predictive p-value is

$$\begin{aligned} p_B &= p\left(t\left(\mathbf{y}\right) \leq t\left(\mathbf{y}^{\text{repl}}\right) \mid \mathbf{y}\right) \\ &= \int d\boldsymbol{\theta}\, p\left(t\left(\mathbf{y}\right) \leq t\left(\mathbf{y}^{\text{repl}}\right) \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \\ &= \int d\boldsymbol{\theta}\, p_f\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \end{aligned}$$

Speaker notes

- For posterior predictive p-values, we condition on the data $\mathbf{y}$ rather than a parameter estimate $. This approach explicitly accounts for the residual uncertainty in the parameters $\boldsymbol{\theta}$.
- In a Bayesian setting, we typically use p-values to check models instead of doing hypothesis tests.
- On the left, the first equality follows by the law of total probability and the second by noting that the first conditional distribution is the frequentist p-value.
- What is the practical difference? The sampling distribution of $t\left(\mathbf{y}^{\text{repl}}\right) \mid \mathbf{y}$ tends to have heavier tails than $t\left(\mathbf{y}^{\text{repl}}\right) \mid \hat{\boldsymbol{\theta}}$ because we consider all possible parameter values that are consistent with the data. Hence, Bayesian p-values tend to be more conservative and closer to 0.5 than frequentist p-values.

Model	p_f	p_B
independent	0.010	0.029
dependent	0.244	0.438

Speaker notes

- The table shows p-values based on 4,000 replications for the independent coin flip and sequential model. The first column contains frequentist p-values and the second Bayesian p-values.
- As we expect, the independent coin flip model is rejected based on the number of changes between heads and tails.
- For both models, $p_f < p_B$ because the distribution of the test statistic conditioning on the data has heavier tails than conditioning on an estimate of the parameter values.

RECAP

- Prior sensitivity for the prior predictive distribution.
- Model checking and replication.
- Posterior predictive p-values for model checking.

Speaker notes

No notes on this slide.