UPPSALA
UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Bayesian Statistics and Data Analysis
Lecture 8a

Måns Magnusson
Department of Statistics, Uppsala University
Thanks to Aki Vehtari, Aalto University

Section 1

Model Checking and Assessment

The Box process: Probabilistic modeling

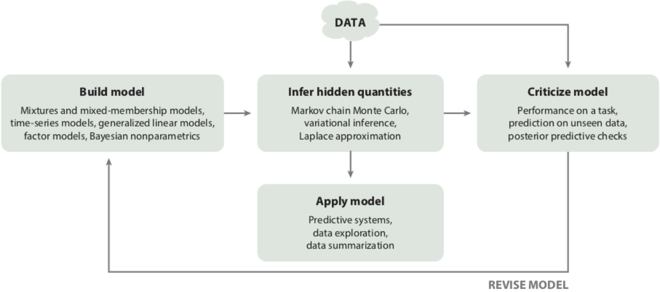


Figure: The Box approach (Box, 1976, Blei, 2014)

Model assessment

- Sensibility with respect to additional information not used in model
  - e.g., if posterior would claim that hazardous chemical decreases probability of death

Model assessment

- Sensibility with respect to additional information not used in model
  - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
  - compare predictions to completely new observations

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Model assessment

- Sensibility with respect to additional information not used in model
  - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
  - compare predictions to completely new observations
- Internal validation
  - posterior predictive checking
  - cross-validation predictive checking

Section 2

Posterior predictive checking

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Newcombs speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Newcombs speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Newcombs speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$

Posterior predictive checking – example

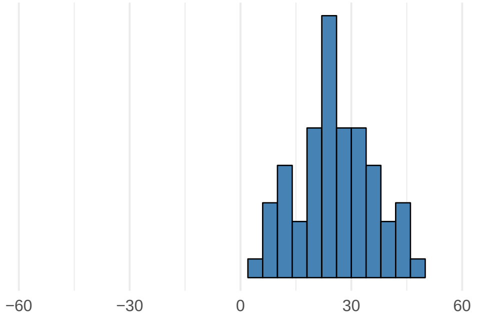
- Newcombs speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$

Posterior predictive checking – example

- Newcombs speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log\sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates

Posterior predictive checking – example

- Newcombs speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates



Replicates vs. future observation

- Predictive $\tilde{y}$ is the next not yet observed possible observation.

Replicates vs. future observation

- Predictive $\tilde{y}$ is the next not yet observed possible observation.
- y^{rep} refers to replicating the whole experiment (potentially with same values of x)
  i.e. obtaining as many replicated observations as in the original data.

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Generate replicated datasets y^{rep}

UPPSALA UNIVERSITET

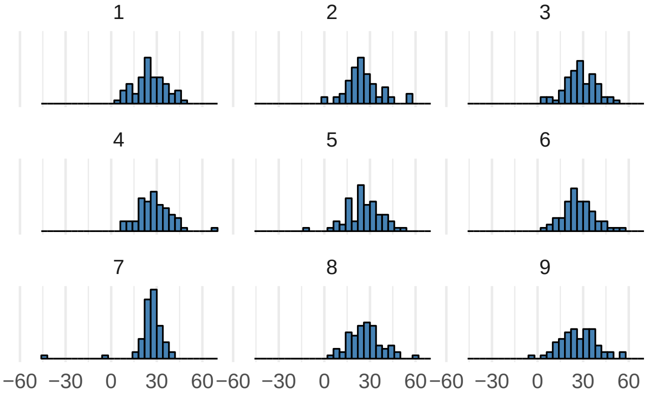
- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Generate replicated datasets y^{rep}
- Compare to the original dataset

Posterior predictive checking – example

- Generate replicated datasets y^{rep}
- Compare to the original dataset



Posterior predictive checking with test statistic

- Replicated data sets y^{rep}
- Test quantity (or discrepancy measure) $T(y, \theta)$
  - summary quantity for the observed data $T(y, \theta)$
  - summary quantity for a replicated data $T(y^{\mathrm{rep}}, \theta)$
  - can be easier to compare summary quantities (y^{rep} statistics) than data sets

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Compute test statistic for data $T(y, \theta) = \min(y)$

UPPSALA UNIVERSITET

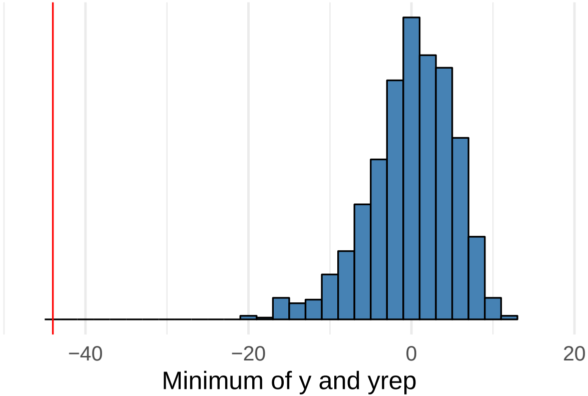
- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\mathrm{rep}})$ for many replicated datasets

Posterior predictive checking – example

- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\text{rep}})$ for many replicated datasets



Posterior predictive checking – example

UPPSALA UNIVERSITET

- Good test statistic is ancillary (or almost)
  - a statistic $T(X)$ that does not depend on the parameters of the model are ancillary
    e.g. in a normal model with known σ^2,

$$s^2 = \sum_i^n \frac{(x_i - \bar{x})^2}{n-1},$$

is ancillary (μ cancel out).

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example

- Good test statistic is ancillary (or almost)
  - a statistic $T(X)$ that does not depend on the parameters of the model are ancillary
    e.g. in a normal model with known σ^2,

    $$s^2 = \sum_i^n \frac{(x_i - \bar{x})^2}{n-1},$$

    is ancillary (μ cancel out).
- Bad test statistic is highly dependent of the parameters
  - e.g. variance (or mean) for normal model with unknown σ^2. If σ^2 changes so will $T(X)$.

Posterior predictive checking – example

- Good test statistic is ancillary (or almost)
  - a statistic $T(X)$ that does not depend on the parameters of the model are ancillary
    e.g. in a normal model with known σ^2,

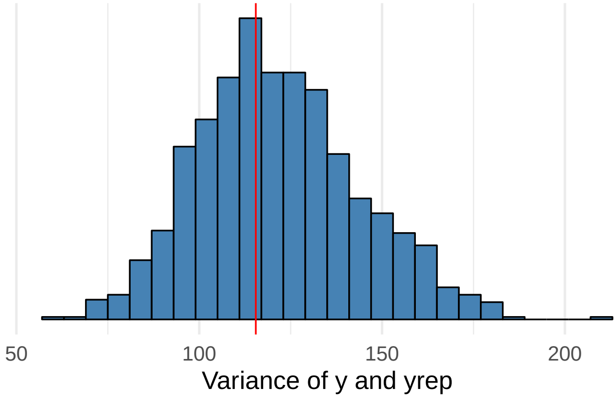
    $$s^2 = \sum_i^n \frac{(x_i - \bar{x})^2}{n-1},$$

    is ancillary (μ cancel out).
- Bad test statistic is highly dependent of the parameters
  - e.g. variance (or mean) for normal model with unknown σ^2. If σ^2 changes so will $T(X)$.
- We want to identify problems in data not captured by the model

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive checking – example



Posterior predictive p-value

- *Posterior predictive p-value*

$$
\begin{aligned}
p &= \Pr(T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta) | y) \\
&= \int \int I_{T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta)} p(y^{\mathrm{rep}} | \theta) p(\theta | y) dy^{\mathrm{rep}} d\theta
\end{aligned}
$$

where I is an indicator function

Posterior predictive p-value

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

- *Posterior predictive p-value*

$$
\begin{aligned}
p &= \Pr(T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta) | y) \\
&= \int\int I_{T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta)} p(y^{\mathrm{rep}} | \theta) p(\theta | y) dy^{\mathrm{rep}} d\theta
\end{aligned}
$$

where I is an indicator function

- having $(y^{\mathrm{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution (Monte Carlo):

$$
T(y^{\mathrm{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S
$$

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Posterior predictive p-value

- *Posterior predictive p-value*

$$
\begin{aligned}
p &= \Pr(T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta)|y) \\
&= \int\int I_{T(y^{\mathrm{rep}},\theta)\geq T(y,\theta)} p(y^{\mathrm{rep}}|\theta) p(\theta|y) dy^{\mathrm{rep}} d\theta
\end{aligned}
$$

where I is an indicator function

  - having $(y^{\mathrm{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution (Monte Carlo):

$$T(y^{\mathrm{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

- Posterior predictive p-value (ppp-value): could difference between the model and data arise by chance

Posterior predictive p-value

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

- *Posterior predictive p-value*

$$
\begin{aligned}
p &= \Pr(T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta) | y) \\
&= \int\int I_{T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta)} p(y^{\mathrm{rep}} | \theta) p(\theta | y) dy^{\mathrm{rep}} d\theta
\end{aligned}
$$

  where I is an indicator function
  - having $(y^{\mathrm{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution (Monte Carlo):

$$
T(y^{\mathrm{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S
$$

- Posterior predictive p-value (ppp-value): could difference between the model and data arise by chance
- Not commonly used, since the distribution of test statistic $T(y, \theta)$ has more information

Marginal and CV predictive checking

- Consider marginal predictive distributions $p(\tilde{y}_i|y)$ and each observation separately
  - marginal posterior p-values

$$p_i = \Pr(T(y_i^{\text{rep}}) \leq T(y_i)|y)$$

if $T(y_i) = y_i$

$$p_i = \Pr(y_i^{\text{rep}} \leq y_i|y)$$

Marginal and CV predictive checking

- Consider marginal predictive distributions $p(\tilde{y}_i|y)$ and each observation separately
  - marginal posterior p-values

$$p_i = \Pr(T(y_i^{\mathrm{rep}}) \leq T(y_i)|y)$$

if $T(y_i) = y_i$

$$p_i = \Pr(y_i^{\mathrm{rep}} \leq y_i|y)$$

- if $Pr(\tilde{y}_i|y)$ well calibrated, distribution of p_i would be uniform between 0 and 1
  - holds better for cross-validation predictive tests: $Pr(\tilde{y}_i|y_{-i})$ (cross-validation)

Marginal predictive checking - Example

- Marginal tail area or Probability integral transform (PIT)

$$p_i = p(y_i^{\mathrm{rep}} \leq y_i | y)$$

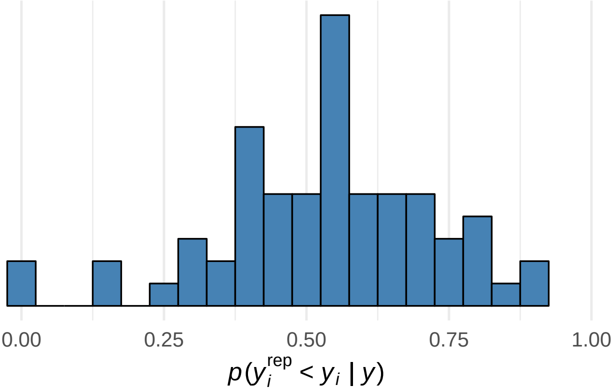
- if $p(\tilde{y}_i | y)$ is well calibrated, distribution of p_i's would be uniform between 0 and 1

Marginal predictive checking - Example

- Marginal tail area or Probability integral transform (PIT)

$$p_i = p(y_i^{\mathrm{rep}} \leq y_i | y)$$

- if $p(\tilde{y}_i|y)$ is well calibrated, distribution of p_i's would be uniform between 0 and 1



Sensitivity analysis

- How much different choices in model structure and priors affect the results

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. t distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to “outliers”
    - e.g. t instead of Gaussian

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. t distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to "outliers"
    - e.g. t instead of Gaussian
- Compare sensitivity of essential inference quantities
  - extreme quantiles are more sensitive than means and medians
  - extrapolation is more sensitive than interpolation

Example: Exposure to air pollution

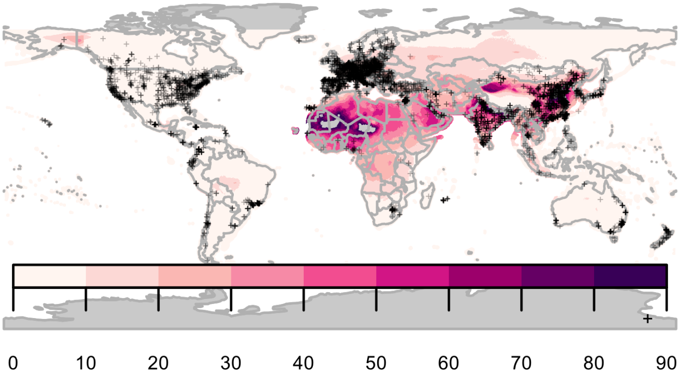
- Example from Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman (2019). Visualization in Bayesian workflow. https://doi.org/10.1111/rssa.12378
- Estimation of human exposure to air pollution from particulate matter measuring less than 2.5 microns in diameter ($\mathrm{PM}_{2.5}$)
  - Exposure to $\mathrm{PM}_{2.5}$ is linked to a number of poor health outcomes and a recent report estimated that $\mathrm{PM}_{2.5}$ is responsible for three million deaths worldwide each year (Shaddick et al., 2017)
  - In order to estimate the public health effect of ambient $\mathrm{PM}_{2.5}$, we need a good estimate of the $\mathrm{PM}_{2.5}$ concentration at the same spatial resolution as our population estimates.

UPPSALA
UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

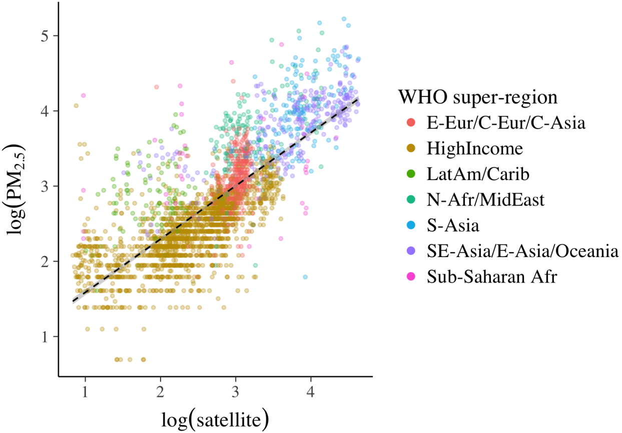
Example: Exposure to air pollution

- Direct measurements of PM 2.5 from ground monitors at 2980 locations
- High-resolution satellite data of aerosol optical depth



Example: Exposure to air pollution

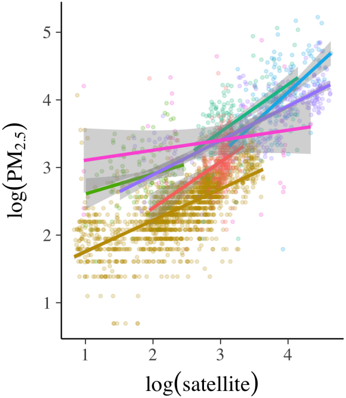
- Direct measurements of PM 2.5 from ground monitors at 2980 locations
- High-resolution satellite data of aerosol optical depth



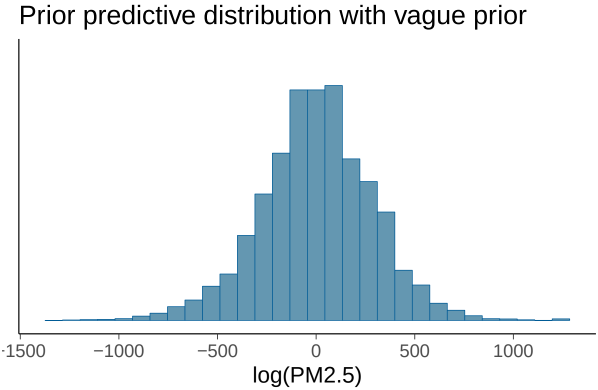
UPPSALA UNIVERSITET

Example: Exposure to air pollution

- Direct measurements of PM 2.5 from ground monitors at 2980 locations
- High-resolution satellite data of aerosol optical depth



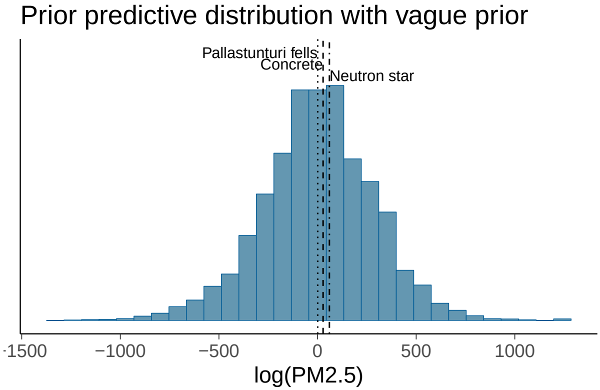
Example: Prior predictive checking



UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking

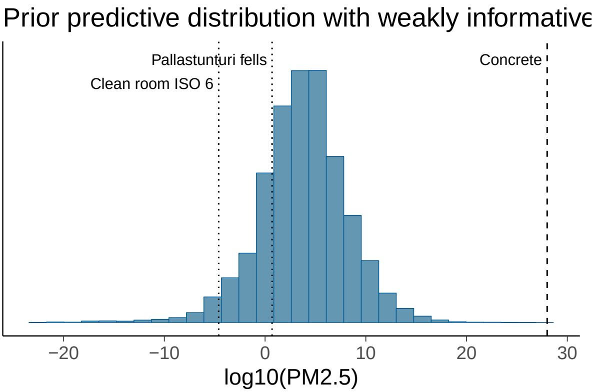
Example: Prior predictive checking



Example: Prior predictive checking

UPPSALA UNIVERSITET

- Model Checking and Assessment
- Posterior predictive checking



Example: Marginal predictive distributions

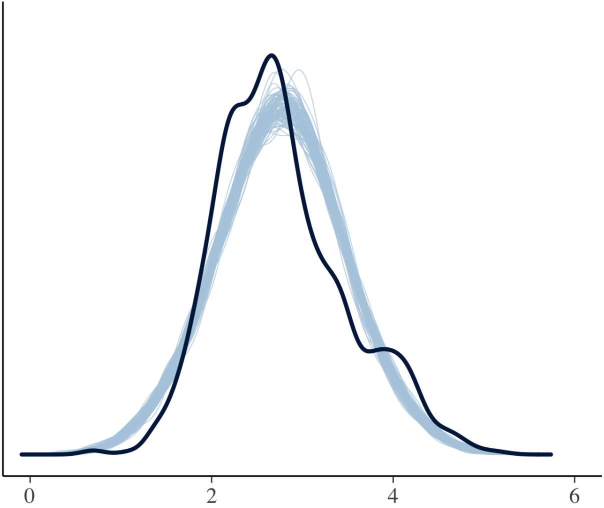


Figure: Model 1

Example: Marginal predictive distributions

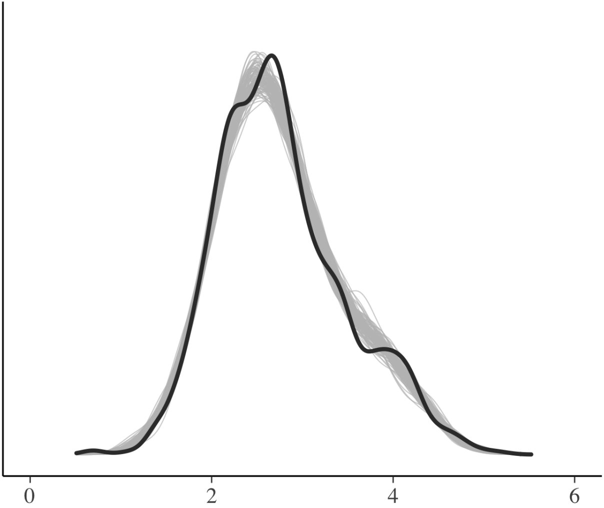


Figure: Model 2

Example: Marginal predictive distributions

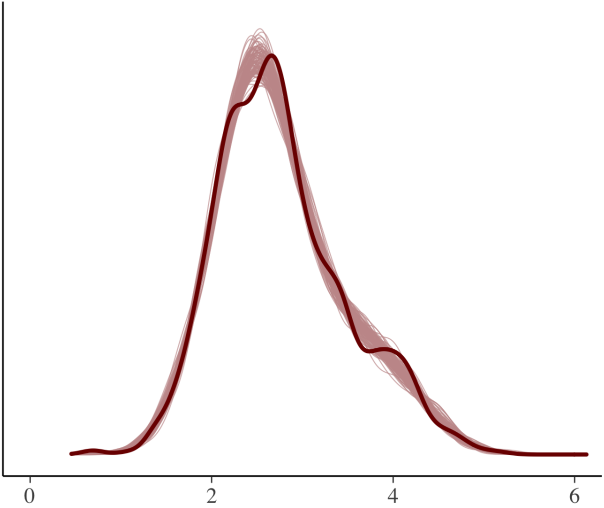


Figure: Model 3

Example: Test statistic (skewness)

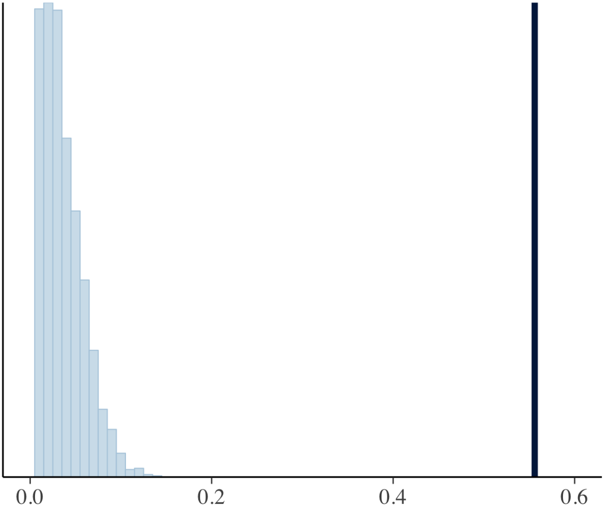


Figure: Model 1

Example: Test statistic (skewness)

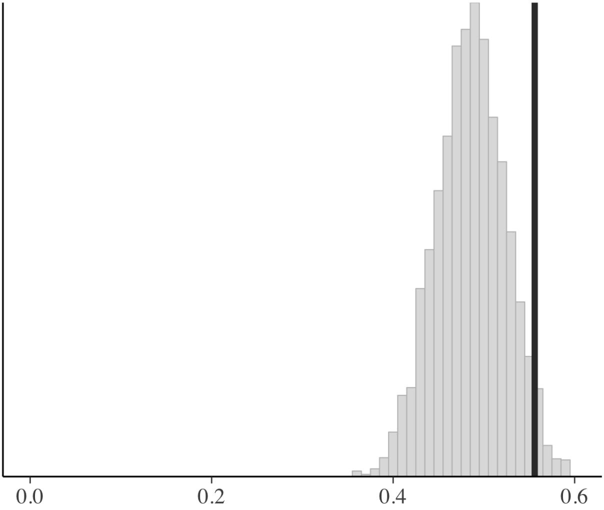


Figure: Model 2

Example: Test statistic (skewness)

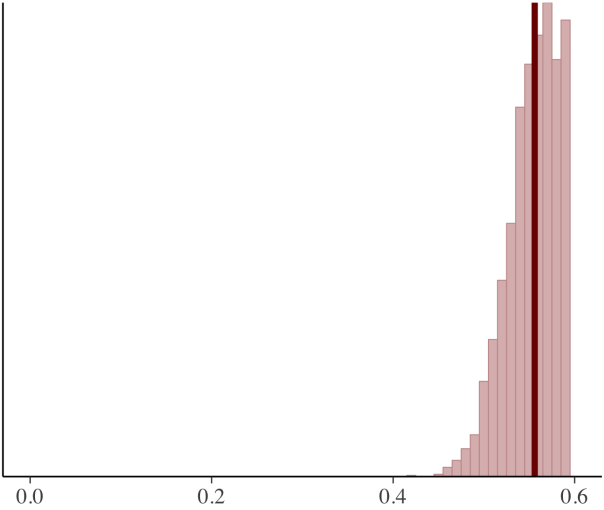


Figure: Model 3