Model Checking

Overview: to check the adequacy of the fit of the model to the data and the plausibility of the model for the purposes for which the model will be used.

We assume $p(y|\theta)$ and $p(\theta)$, so it would be prudent to determine if these assumptions are reasonable.

- (Prior) sensitivity analysis
- Posterior predictive checks
  - Graphical checks
  - Posterior predictive pvalues

Sensitivity analysis

- How much do 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

Prior sensitivity analysis

Since a prior specifies our prior belief, we may want to check to determine whether our conclusions would change if we held different prior beliefs. Suppose a particular scientific question can be boiled down to

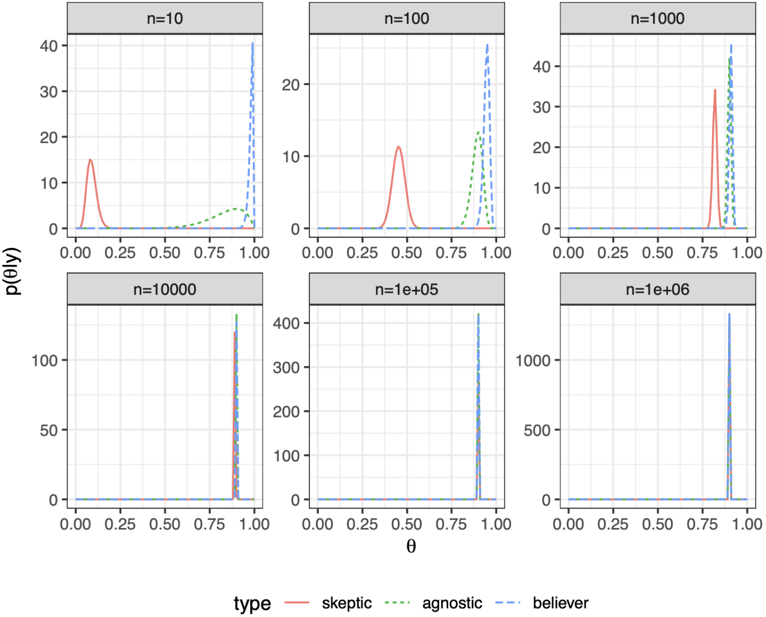
$$Y_i \overset{ind}{\sim} Ber(\theta)$$

and that there is wide disagreement about the value for θ such that the following might reasonably characterize different individual beliefs before the experiment is run:

- Skeptic: $\theta \sim Beta(1, 100)$
- Agnostic: $\theta \sim Beta(1, 1)$
- Believer: $\theta \sim Beta(100, 1)$

An experiment is run and the posterior under these different priors are compared.

Posterior distribution



Hierarchical variance prior (Review Gelman 2006)

Recall the normal hierarchical model

$$y_i \stackrel{ind}{\sim} N(\theta_i, s_i^2), \quad \theta_i \stackrel{ind}{\sim} N(\mu, \tau^2)$$

which results in the posterior distribution for τ of

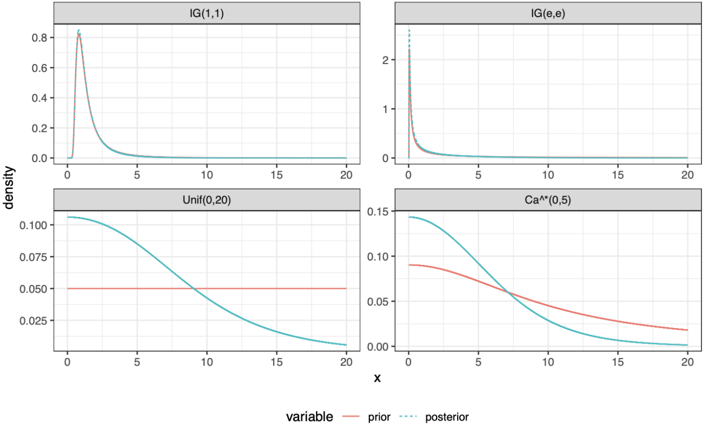
$$p(\tau|y) \propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp(-\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)})$$

School	A	B	C	D	E	F	G	H
y_j	28	8	-3	7	-1	1	18	12
σ_j	15	10	16	11	9	22	20	28

As an attempt to be non-informative, consider an $IG(\epsilon, \epsilon)$ prior for τ^2. As an alternative, consider $\tau \sim Unif(0, C)$ or $\tau \sim Ca^+(0, C)$ where C is problem specific, but is chosen to be relatively large for the particular problem.

Posterior distribution - 8 schools example

Reproduction of Gelman 2006:



Summary

For a default prior on a variance (σ^2) or standard deviation (σ), use

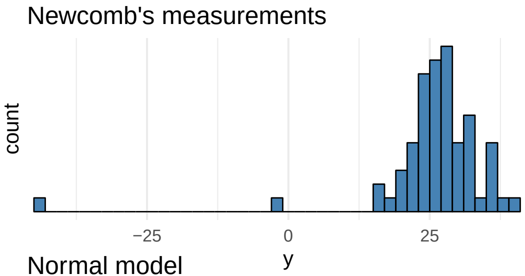
1. Easy
    - Half-Cauchy on the standard deviation ($\sigma \sim Ca^+(0, C)$).
2. Complex
    - Data-level variance
        - Use default prior ($p(\sigma^2) \propto 1/\sigma^2$)
    - Hierarchical standard deviation
        - Use uniform prior (Unif$(0, C)$) if there are enough reps (5 or more) of that parameter.
        - Use half-Cauchy prior ($Ca^+(0, C)$) otherwise.

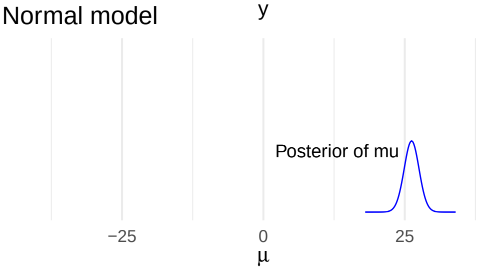
When assigning the values for C

- For a uniform prior (Unif$(0, C)$) make sure C is large enough to capture any reasonable value for the standard deviation.
- For a half-Cauchy prior ($Ca^+(0, C)$), a value of C that is too small will fail to capture the tail, implying that standard deviation needs to be larger whereas a value of C that is too large will put too much weight toward large values of the standard deviation and make the prior more informative.

Simon Newcomb’s light of speed experiment in 1882

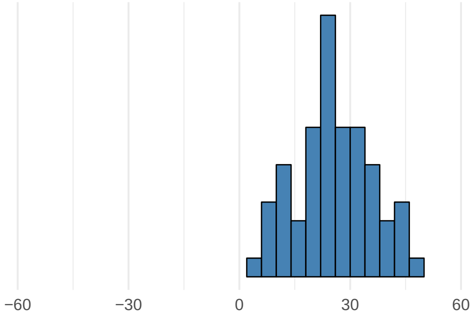
Newcomb measured ($n = 66$) the time required for light to travel from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of 7422 meters.





Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{N}(\mu, \sigma)$ 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 $\mathrm{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates

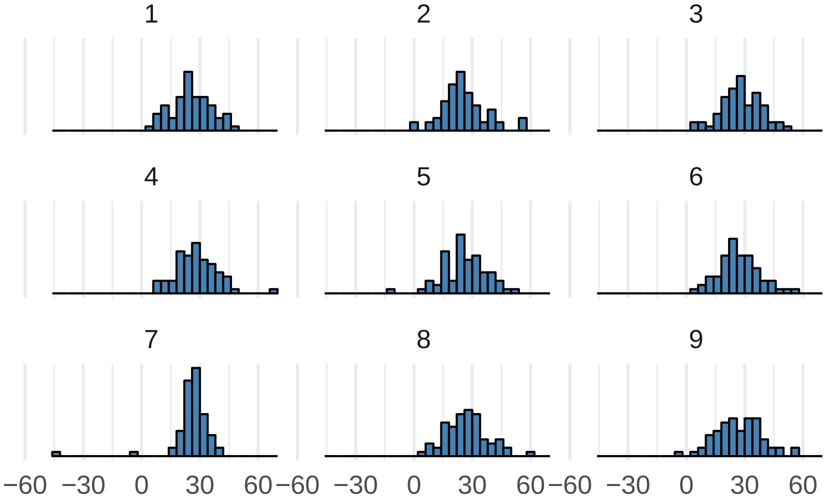


Replicates vs. future observation

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

Posterior predictive checking – example

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

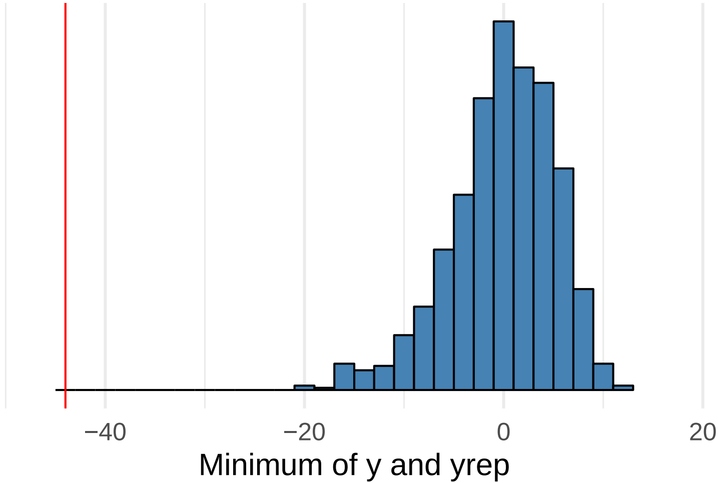


Posterior predictive checking with test statistic

- 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 than data sets

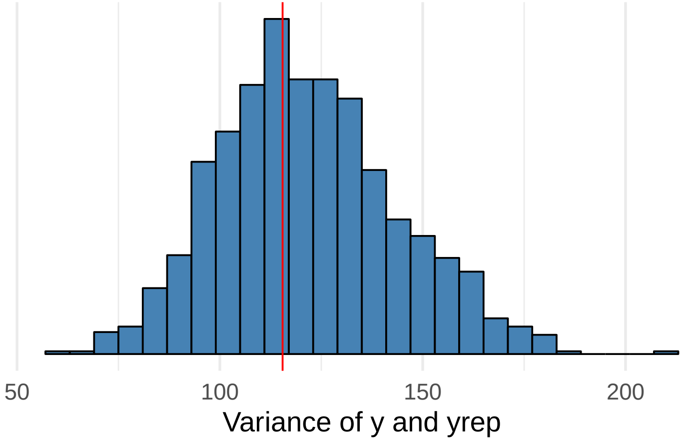
Posterior predictive checking – example

- 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

- Good test statistic is ancillary (or almost)
  - ancillary if it depends only on observed data and if its distribution is independent of the parameters of the model
- Bad test statistic is highly dependent of the parameters
  - e.g. variance for normal model



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, easy to compute

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

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance
- Not commonly used, since the distribution of test statistic has more information

Marginal 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 $T(y_i) = y_i$

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

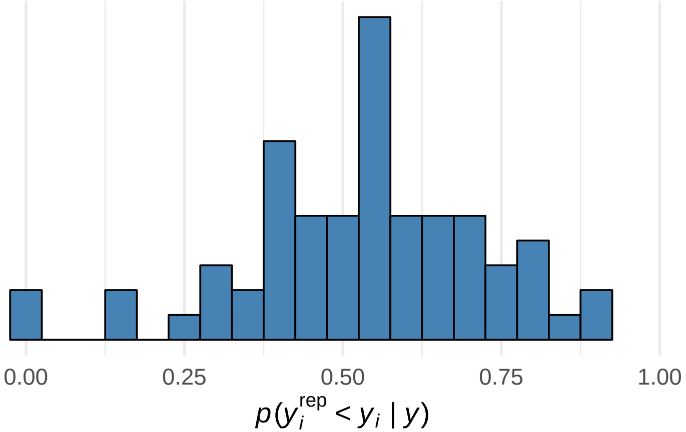
- if $Pr(\tilde{y}_i|y)$ well calibrated, distribution of p_i would be uniform between 0 and 1

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

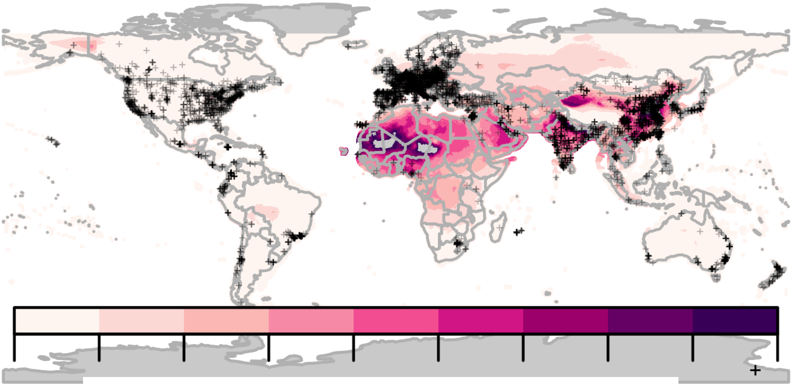


Example: Exposure to air pollution

- 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.

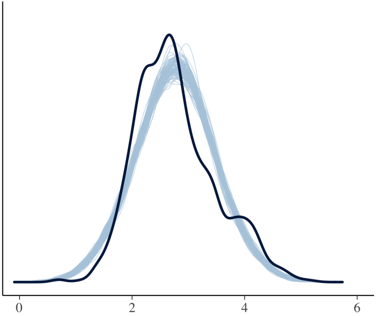
Example: Exposure to air pollution

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

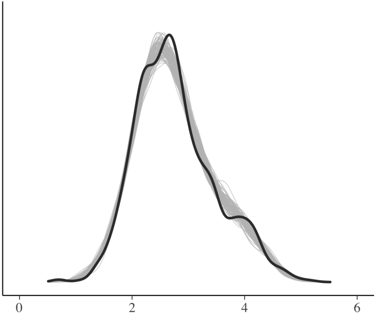


Example: Exposure to air pollution

Posterior predictive checking – marginal predictive distributions



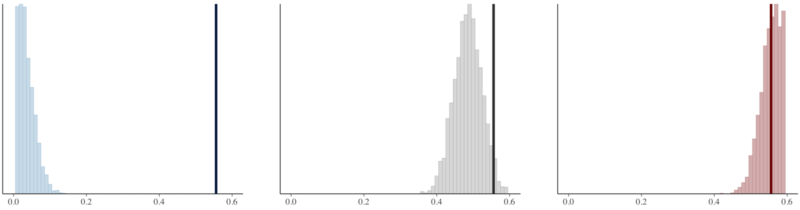
(a) Model 1



(b) Model 2

Example: Exposure to air pollution

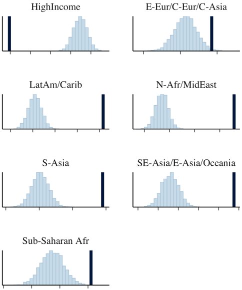
Posterior predictive checking – test statistic (skewness)



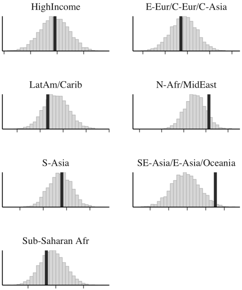
(a) Model 1 (b) Model 2 (c) Model 3

Example: Exposure to air pollution

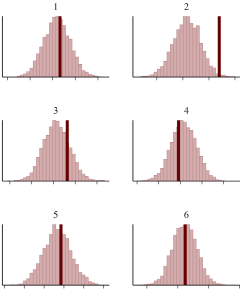
Posterior predictive checking – test statistic (median for groups)



(a) Model 1



(b) Model 2



(c) Model 3