Model checking – overview

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

Model checking – overview

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

Model checking – overview

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

Chapter 6

- 6.1 The place of model checking in applied Bayesian statistics
- 6.2 Do the inferences from the model make sense?
- 6.3 Posterior predictive checking
- 6.4 Graphical posterior predictive checks
  - this can be skimmed, see instead the paper
    Gabry et al. (2019). *Visualization in Bayesian workflow*
    https://doi.org/10.1111/rssa.12378
- 6.5 Model checking for the educational testing example

Model checking

- demo6_1: Posterior predictive checking - light speed
- demo6_2: Posterior predictive checking - sequential dependence
- demo6_3: Posterior predictive checking - poor test statistic
- https://avehtari.github.io/BDA_R_demos/demos_rstan/brms_demo.html

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 $y \sim \mathrm{normal}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{normal}(\mu, \sigma)$ with prior $(\mu, \log\sigma) \propto 1$
- Posterior predictive replicate y^{rep}

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{normal}(\mu, \sigma)$ 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

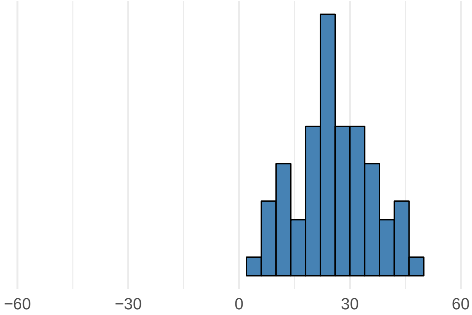
- Newcomb's speed of light measurements
  - model $y \sim \mathrm{normal}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma | y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathrm{normal}(\mu^{(s)},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

Posterior predictive checking – example

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

Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathrm{normal}(\mu, \sigma)$ 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 $\mathrm{normal}(\mu^{(s)}, \sigma^{(s)})$
  - 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

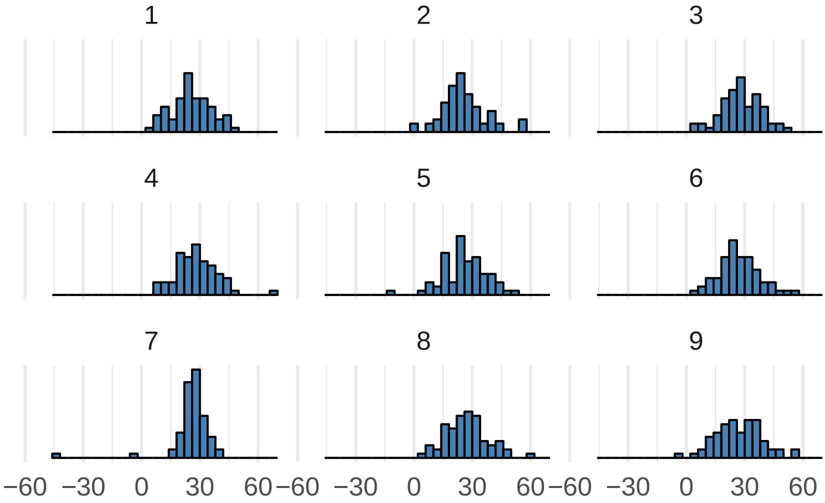
- Generate several replicated datasets y^{rep}

Posterior predictive checking – example

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

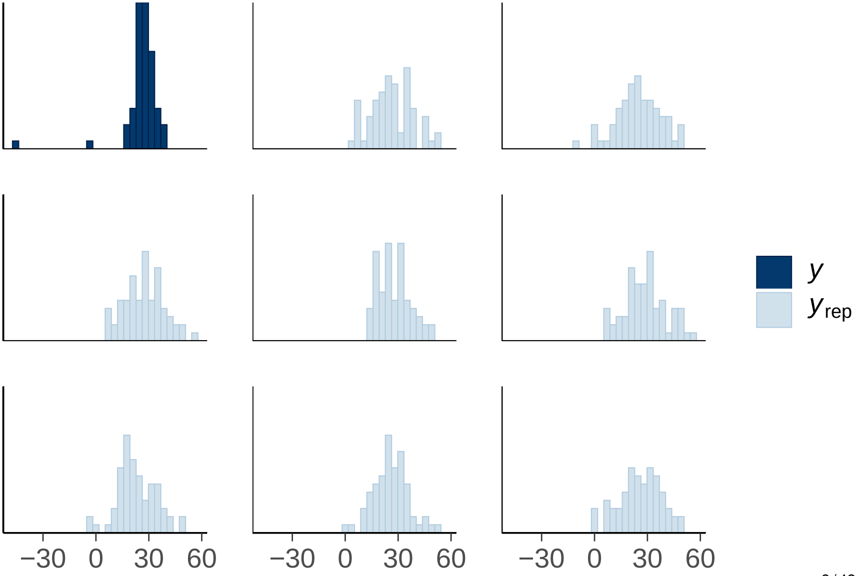
Posterior predictive checking – example

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



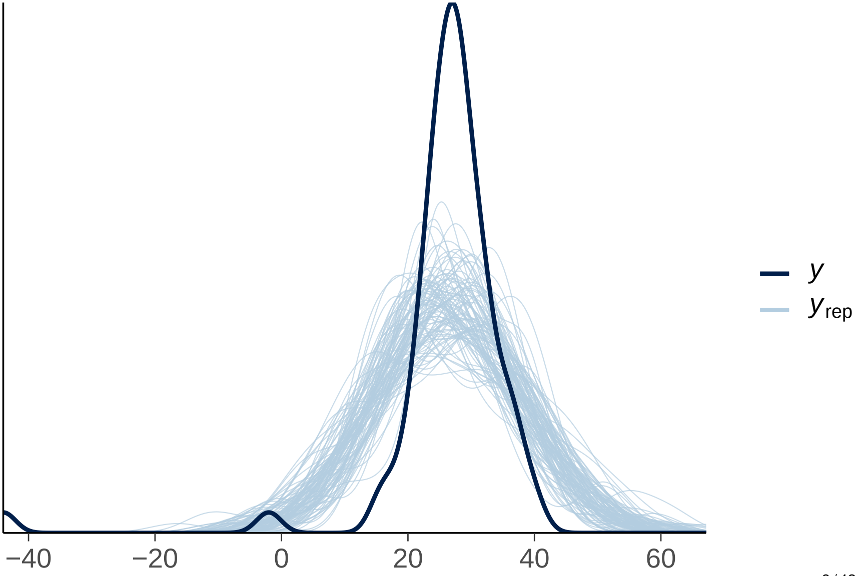
Posterior predictive checking – bayesplot

```
ppc_hist(y, yrep[1:8,])
```



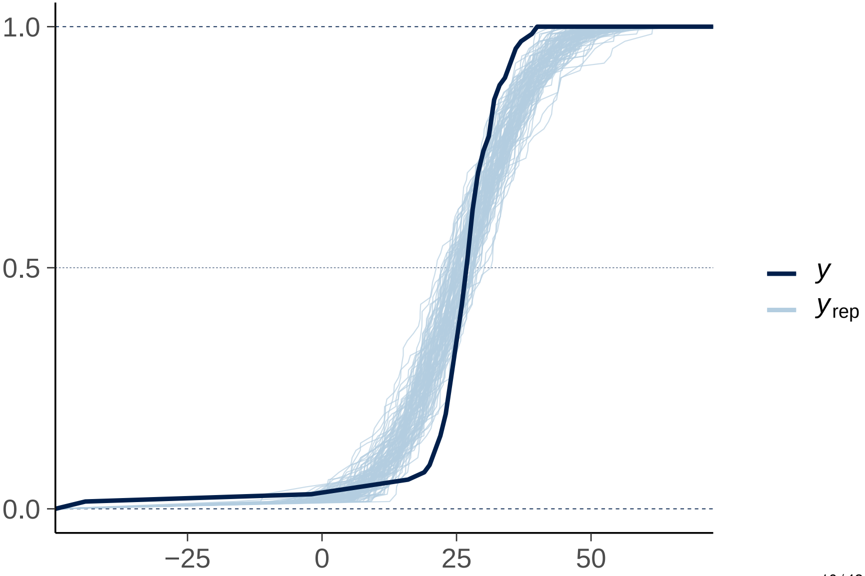
Posterior predictive checking – bayesplot

```
ppc_dens_overlay(y, yrep[1:100,])
```



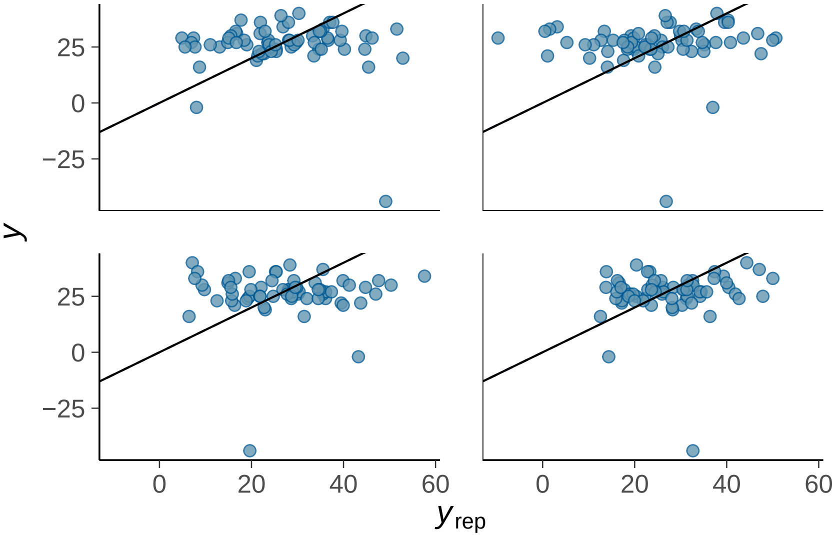
Posterior predictive checking – bayesplot

```
ppc_ecdf_overlay(y, yrep[1:100,])
```



Posterior predictive checking – bayesplot

```
ppc_scatter(y, yrep[1:4,]) + geom_abline()
```



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 $T(y^{\text{rep}}, \theta)$
  - can be easier to compare summary quantities than data sets

Posterior predictive checking – example

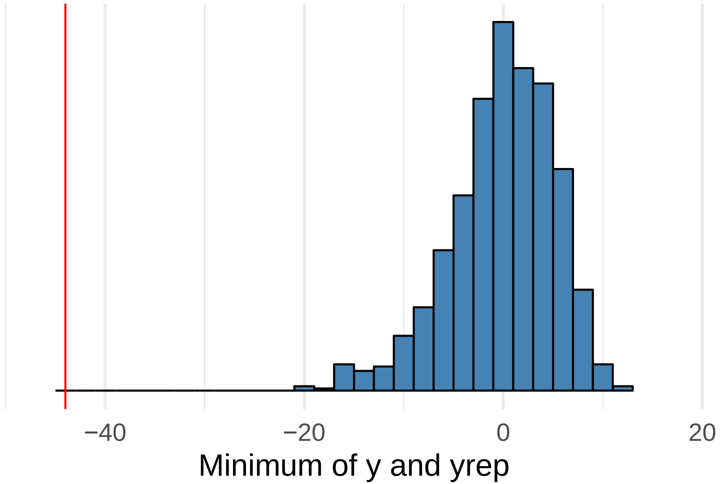
- Compute test statistic for data $T(y, \theta) = \min(y)$

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

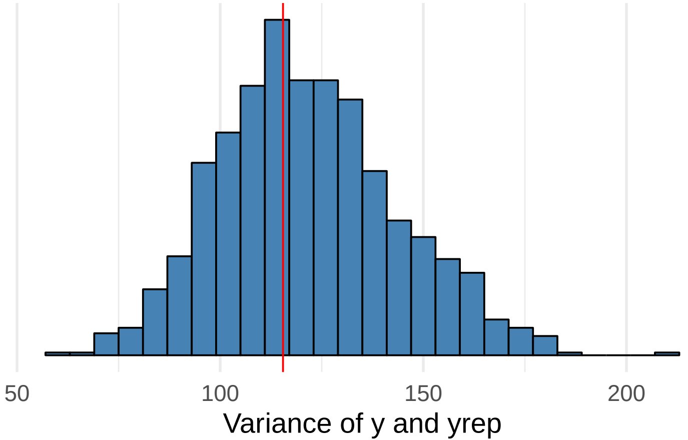
- Good test statistic is (almost) *ancillary* (fi: *aputunnusluku*)
  - ancillary if it depends only on observed data and if its distribution is independent of the parameters of the model

Posterior predictive checking – example

- Good test statistic is (almost) *ancillary* (fi: *aputunnusluku*)
  - 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 – example

- Good test statistic is (almost) *ancillary* (fi: *aputunnusluku*)
  - 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*

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

  where I is an indicator function

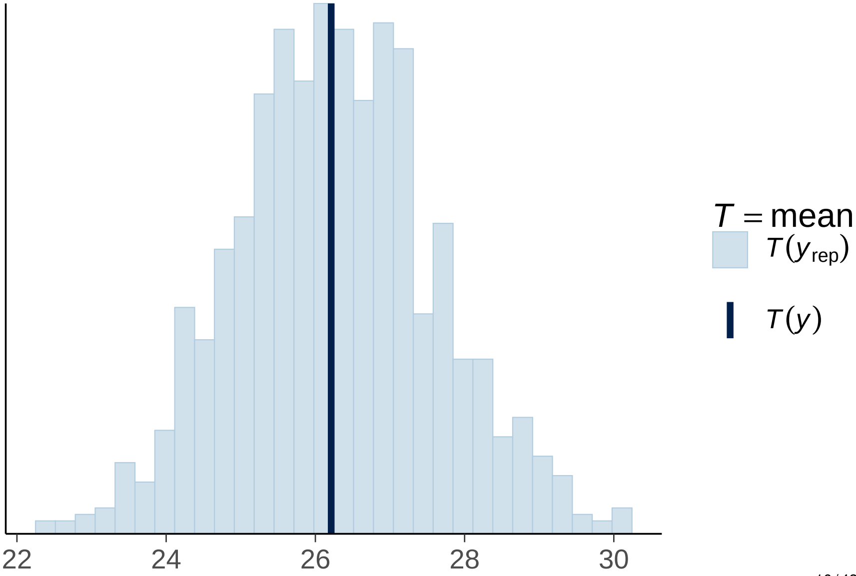
  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

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

- Posterior predictive p-value (ppp-value) estimates whether difference between the model and data could arise by chance
- Not commonly used, as
  - not calibrated in case of non-ancillary statistic
  - the distribution of test statistic has more information

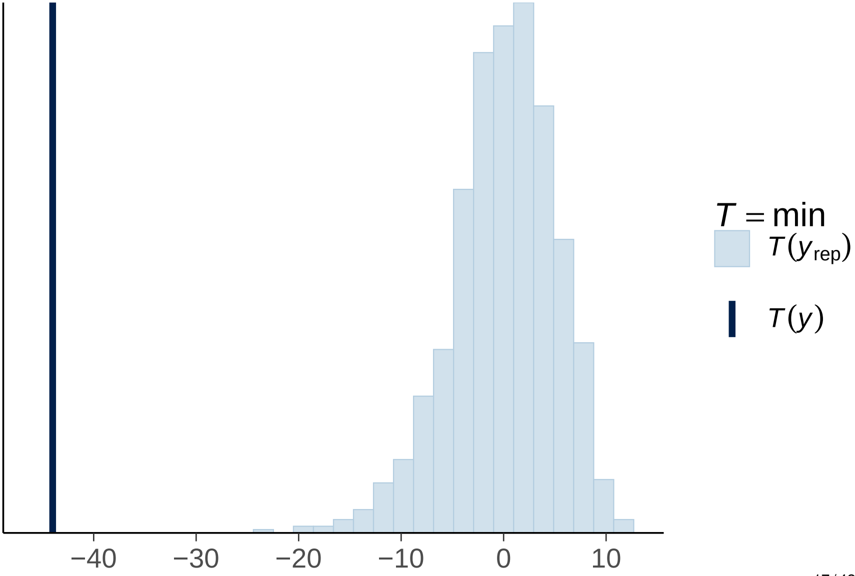
Posterior predictive checking – bayesplot

`ppc_stat(y, yrep)`, the default statistic "mean" is usually bad



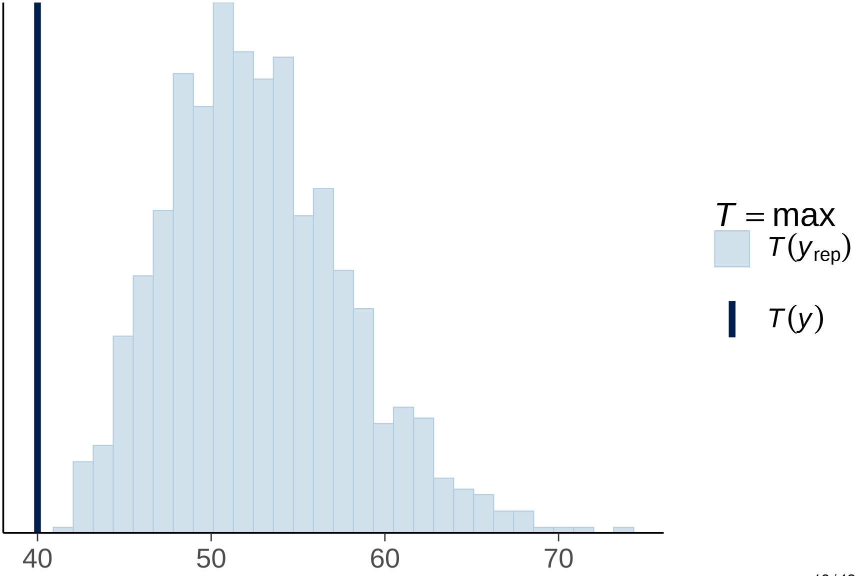
Posterior predictive checking – bayesplot

```
ppc_stat(y, yrep, stat="min")
```



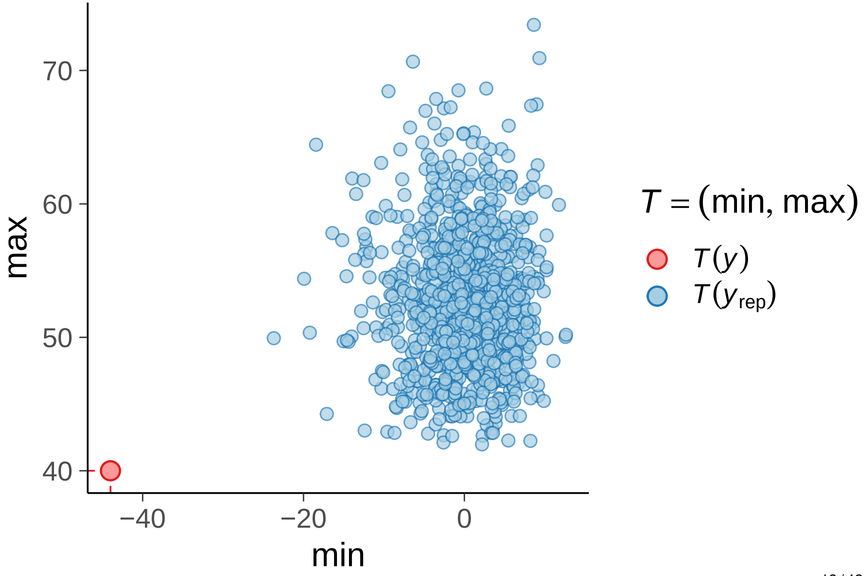
Posterior predictive checking – bayesplot

```
ppc_stat(y, yrep, stat="max")
```



Posterior predictive checking – bayesplot

```
ppc_stat2d(y, yrep, stat=c("min","max"))
```



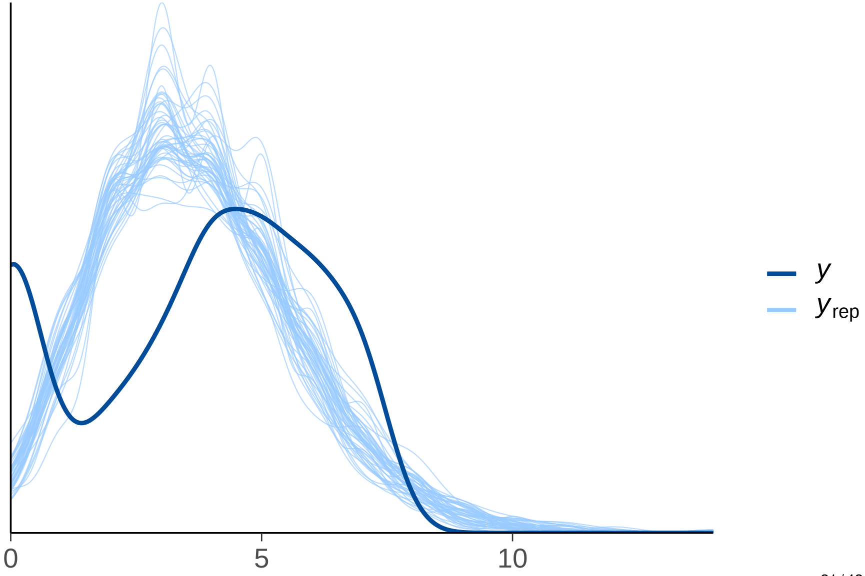
Posterior predictive checking – Stan code

- demo demos_rstan/ppc/poisson-ppc.Rmd

```
data {
  int<lower=1> N;
  array[N] int<lower=0> y;
}
parameters {
  real<lower=0> lambda;
}
model {
  lambda ~ exponential(0.2);
  y ~ poisson(lambda);
}
generated quantities {
  real log_lik[N];
  array[N] int y_rep;
  for (n in 1:N) {
    y_rep[n] = poisson_rng(lambda);
    log_lik[n] = poisson_lpmf(y[n] | lambda);
    }
}
```

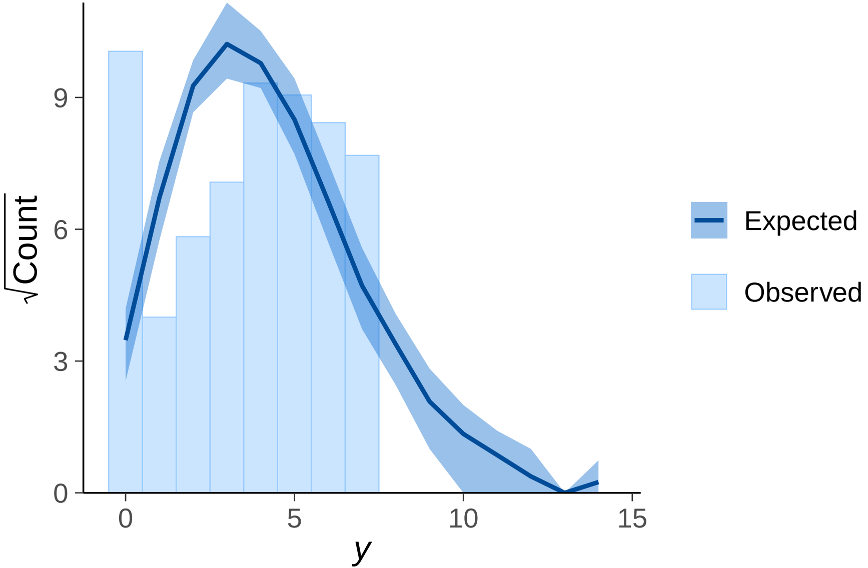
PPC for count data – Poisson model

```
ppc_dens_overlay(y, yrep[1:50,])
```



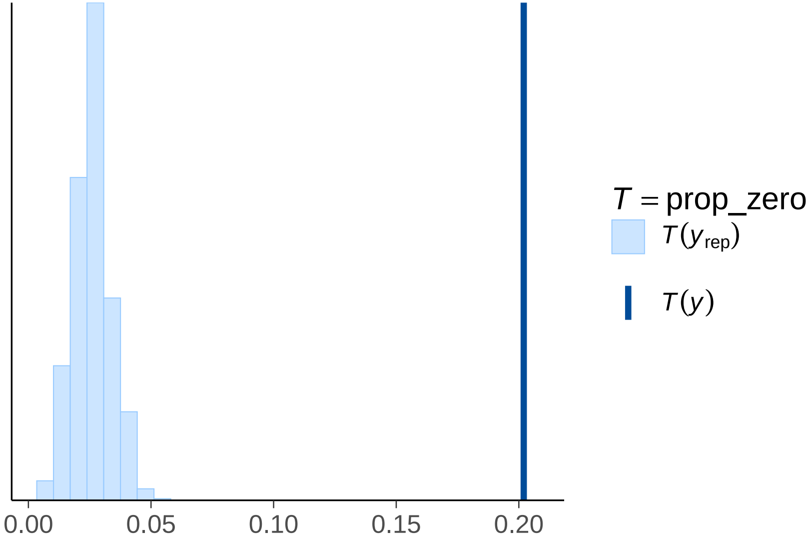
PPC for count data – Poisson model

```
ppc_rootogram(y, yrep)
```



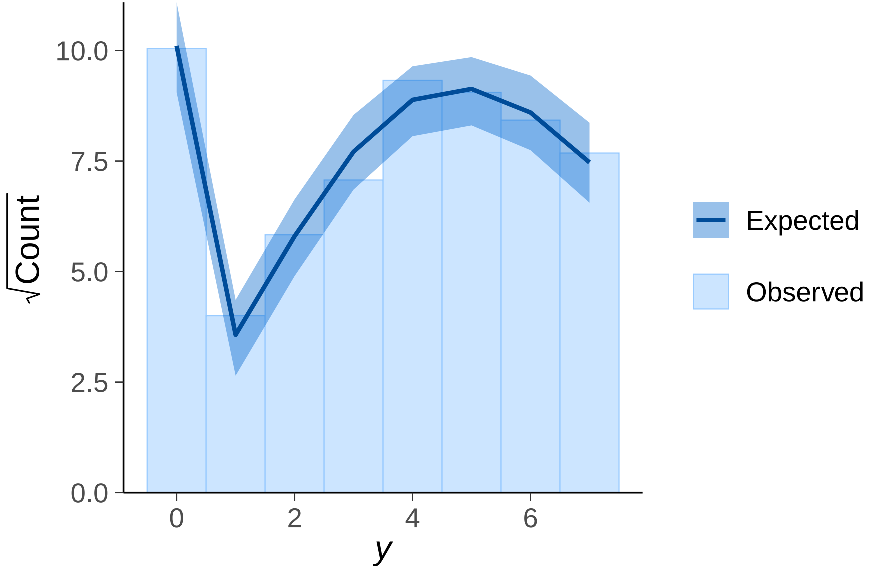
PPC for count data – Poisson model

```
prop_zero <- function(x) mean(x == 0)
ppc_stat(y, yrep, stat = "prop_zero")
```



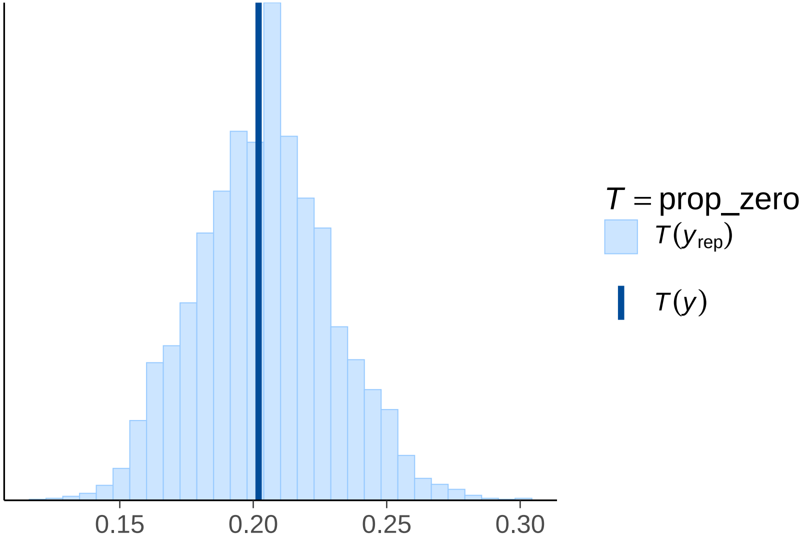
PPC for count data – hurdle truncated Poisson model

```
ppc_rootogram(y, yrep2)
```



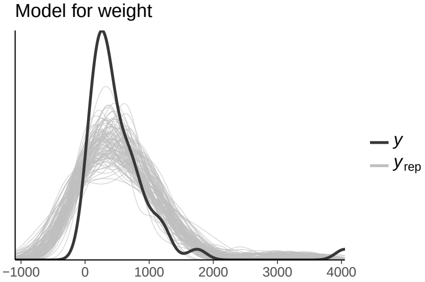
PPC for count data – hurdle truncated Poisson model

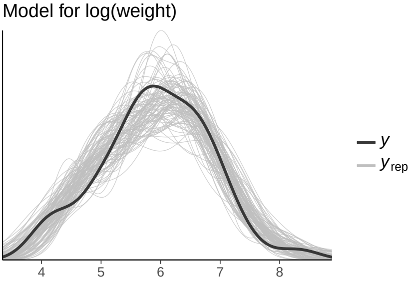
```r
prop_zero <- function(x) mean(x == 0)
ppc_stat(y, yrep2, stat = "prop_zero")
```



Posterior predictive checking: Mesquite bushes

Positive target: normal vs log-normal model



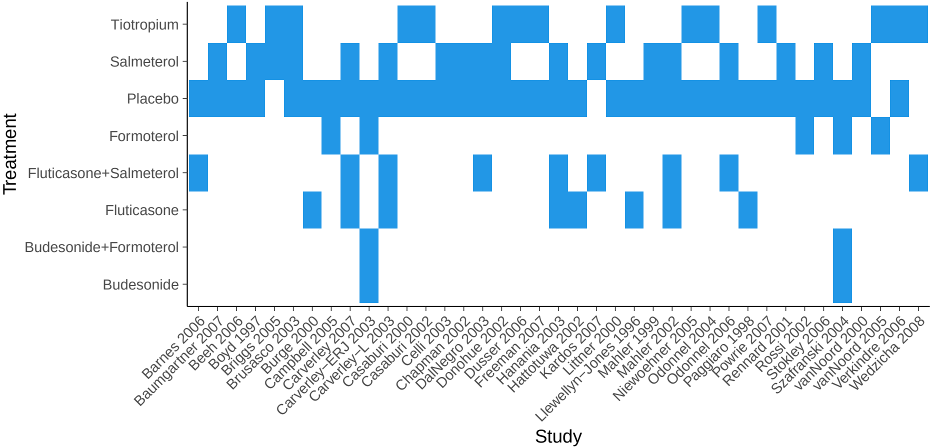


Predicting the yields of mesquite bushes.

Gelman, Hill & Vehtari (2020): Regression and Other Stories, Chapter 11.

Meta-analysis

Pharmacologic treatments for chronic obstructive pulmonary disease

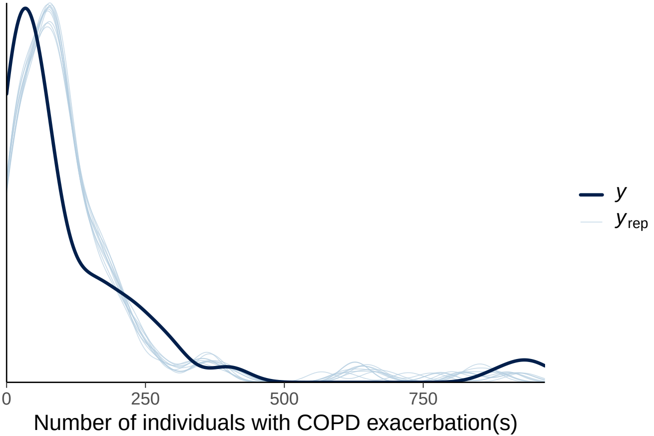


Posterior predictive checking

Pharmacologic treatments for chronic obstructive pulmonary disease

Pooled over studies, separate for treatments

```
fit_pooled <- brm(exac | trials(total) ~ 0 + treatment,
                  prior = prior(student_t(7, 0, 1.5), class='b'),
                  family=binomial(), data=dat.baker2009)
```

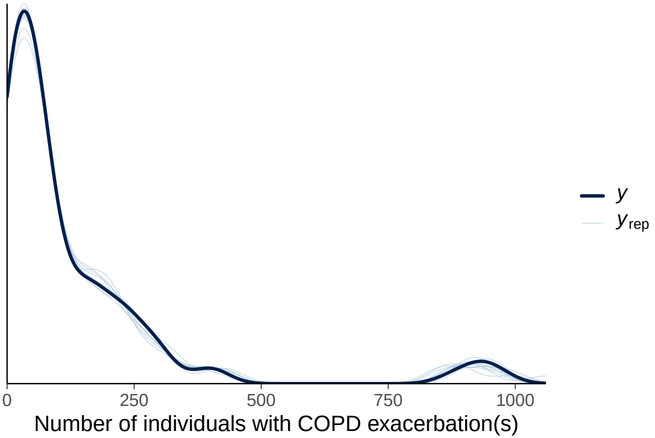


Posterior predictive checking

Pharmacologic treatments for chronic obstructive pulmonary disease

Hirerachical for studies, hierarchical for treatments

```
fit_hier <- brm(exac | trials(total) ~ (1 | treatment) + (1 | study),
                family=binomial(), data=dat.baker2009)
```

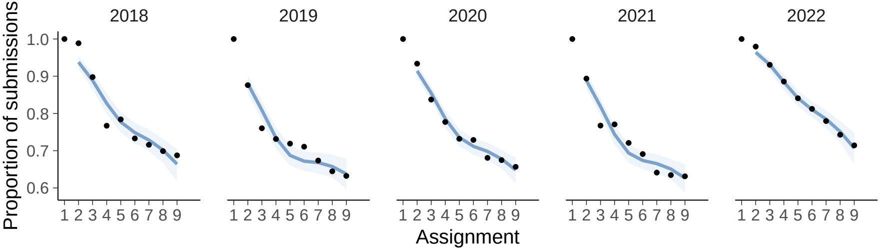


Student retention

Latent hierarchical linear + spline

```
nstudents | trials(nstudents1) ~
  s(assignment, k=4) + (assignment | year),
  family=binomial()
```

Latent functions + posterior uncertainty

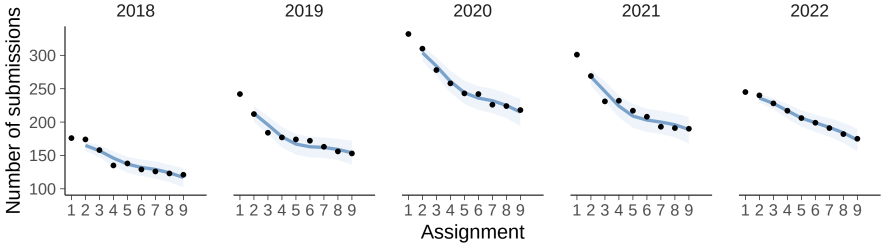


Student retention

Latent hierarchical linear + spline

```
nstudents | trials(nstudents1) ~
  s(assignment, k=4) + (assignment | year),
  family=binomial()
```

Latent functions + posterior uncertainty



Student retention

1. Latent hierarchical linear model

```
nstudents | trials(nstudents1) ~
  (assignment | year),
  family=binomial()
```

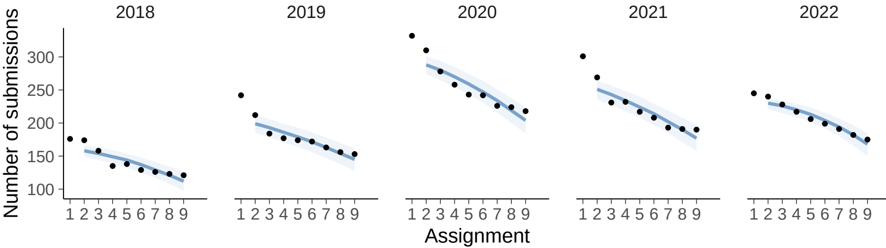
2. Latent spline + hierarchical linear model

```
nstudents | trials(nstudents1) ~
  s(assignment, k=4) + (assignment | year),
  family=binomial()
```

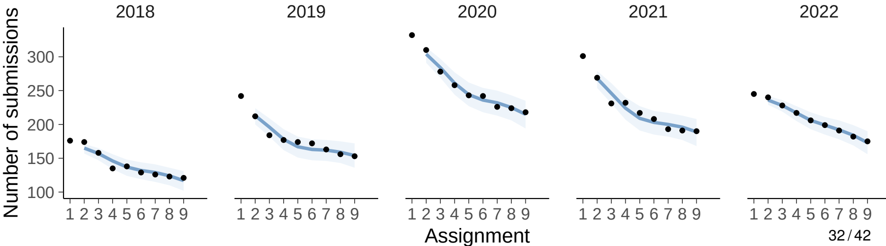
Student retention – Posterior predictive distributions

with `tidybayes`

Latent hierarchical linear model



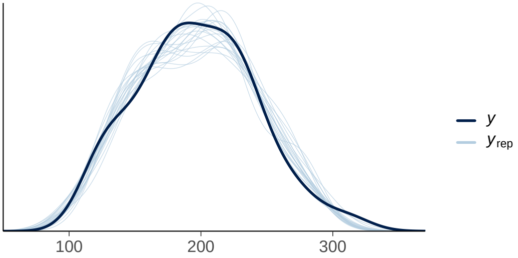
Latent hierarchical linear model + spline



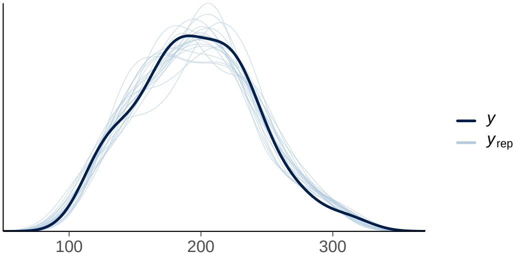
Student retention – Marginal PPC (brms)

```
pp_check(fit, ndraws=100)
```

Latent hierarchical linear model



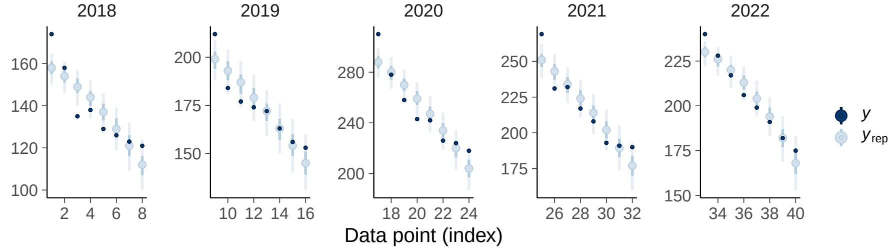
Latent hierarchical linear model + spline



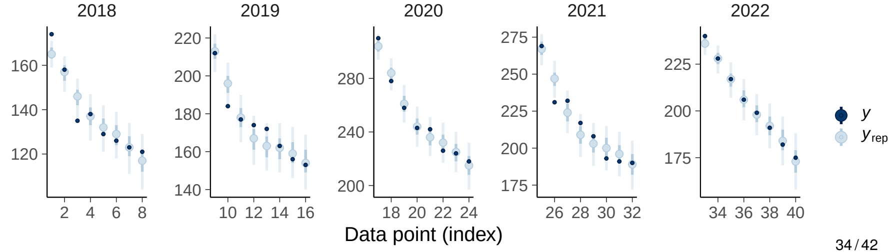
Student retention – Posterior predictive intervals (brms)

```
pp_check(fit, type = "intervals_grouped", group="year")
```

Latent hierarchical linear model



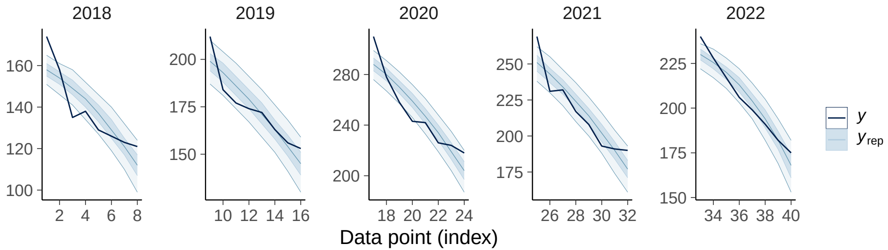
Latent hierarchical linear model + spline



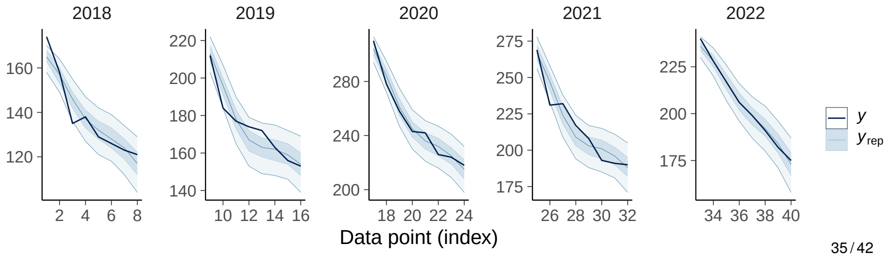
Student retention – Posterior predictive ribbon (brms)

```
pp_check(fit, type = "ribbon_grouped", group="year")
```

Latent hierarchical linear model

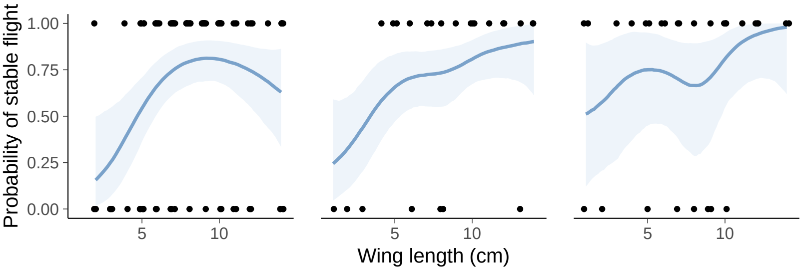


Latent hierarchical linear model + spline



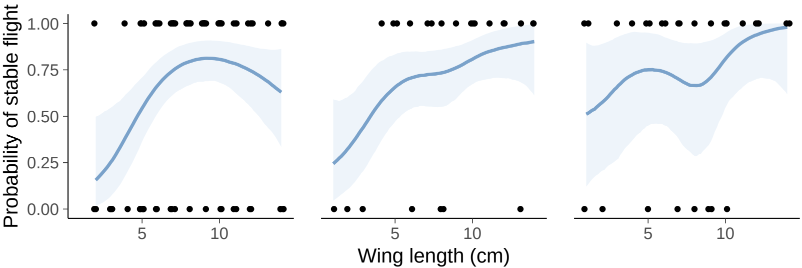
PPC for binary target – Helicopters (brms)

```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips),
family = bernoulli()
```

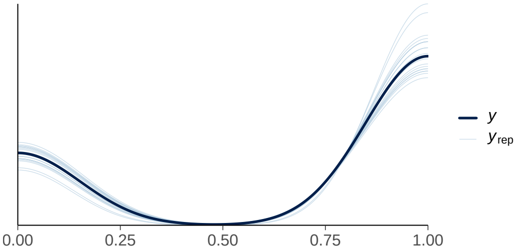


PPC for binary target – Helicopters (brms)

```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips),
family = bernoulli()
```



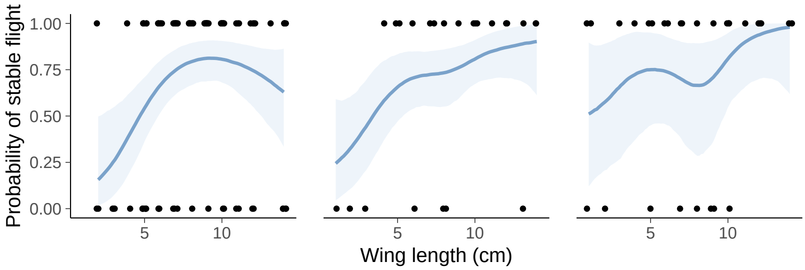

```
pp_check(fit, ndraws=20)
```



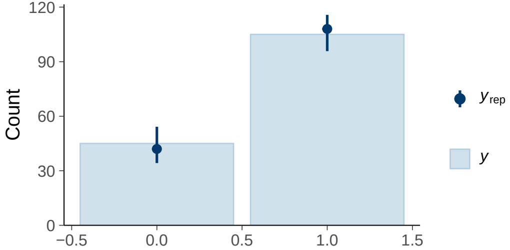
Continuous kernel density estimate for binary target doesn't make sense!

PPC for binary target – Helicopters (brms)

```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips),
family = bernoulli()
```



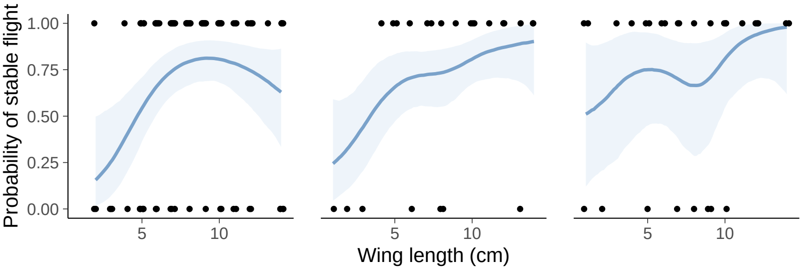

```
pp_check(fit, type="bars")
```



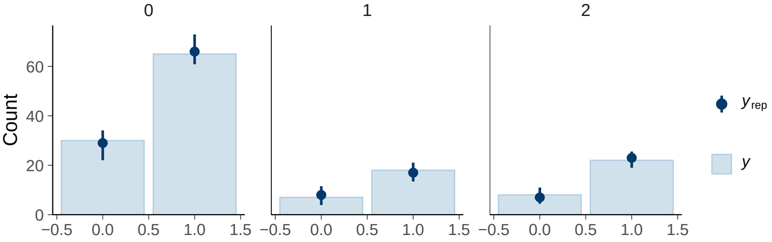
Bar plot showing marginal probabilities for binary data is almost always useless!

PPC for binary target – Helicopters (brms)

```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips),
family = bernoulli()
```

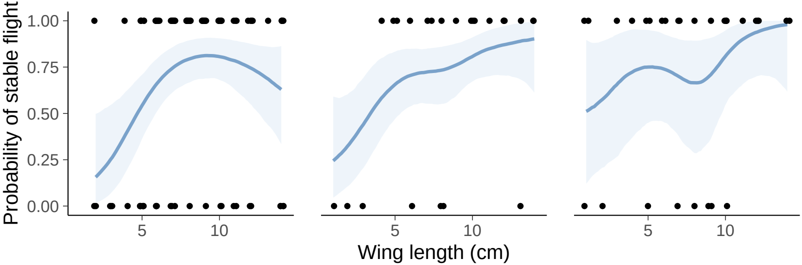



```
pp_check(fit, type="bars_grouped")
```

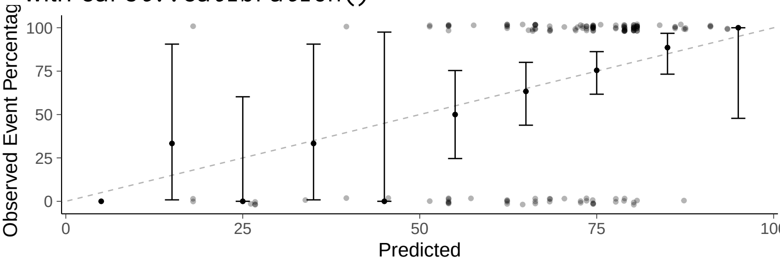


PPC for binary target – Helicopters (brms)

```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips),
family = bernoulli()
```

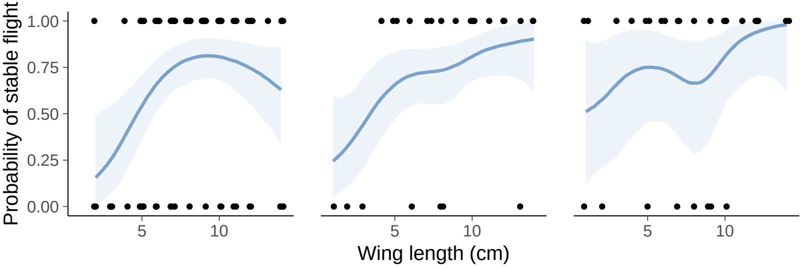


with `caret::calibration()`

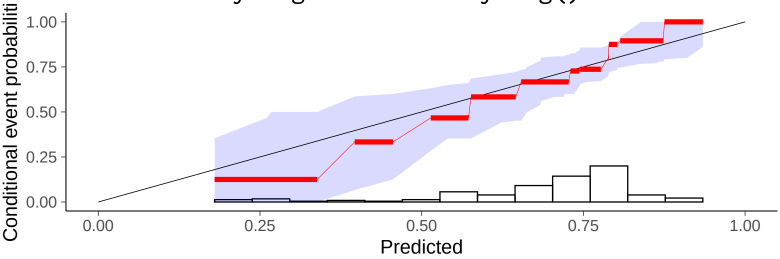


PPC for binary target – Helicopters (brms)

```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips),
family = bernoulli()
```



with `reliabilitydiag::reliabilitydiag()`



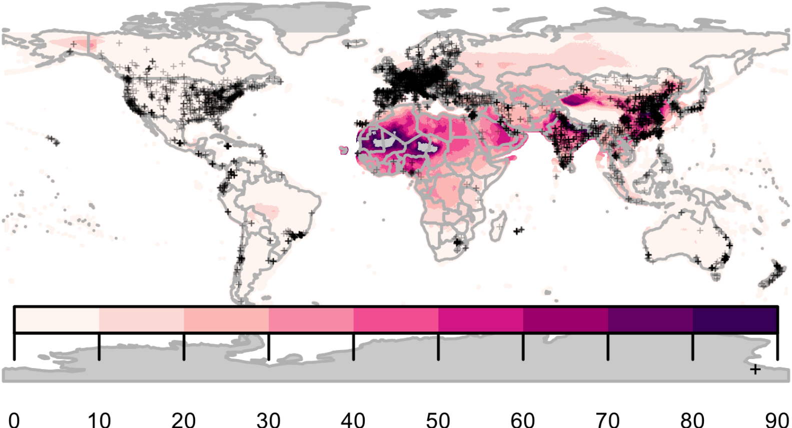
Use this for binary data!

Prior predictive checking: Exposure to air pollution

- Example from Gabry, Simpson, Vehtari, Betancourt, and 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.

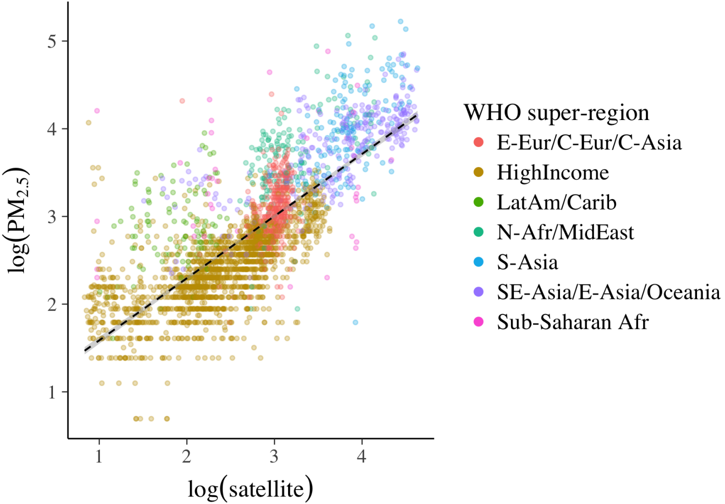
Prior predictive checking: Exposure to air pollution

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



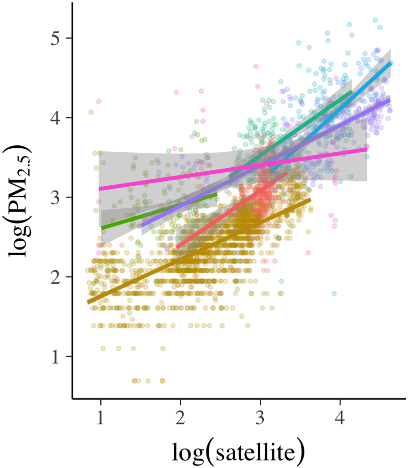
Prior predictive checking: Exposure to air pollution

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



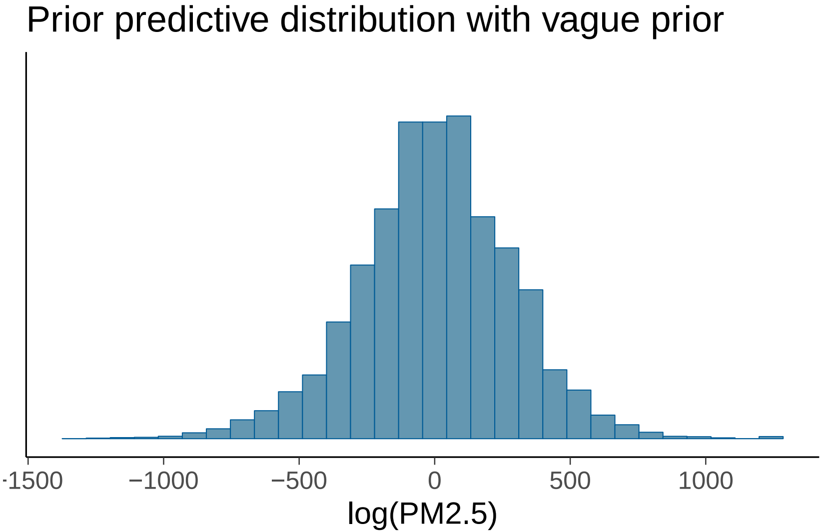
Prior predictive checking: Exposure to air pollution

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



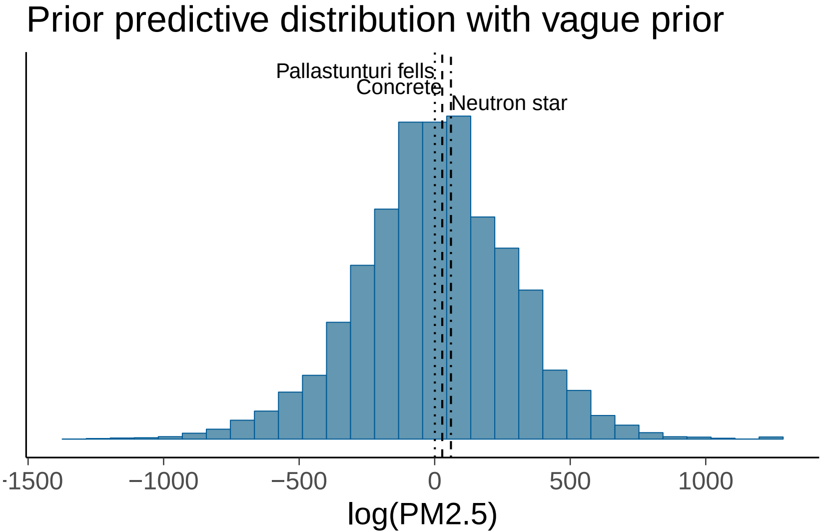
Prior predictive checking: Exposure to air pollution

Prior predictive checking



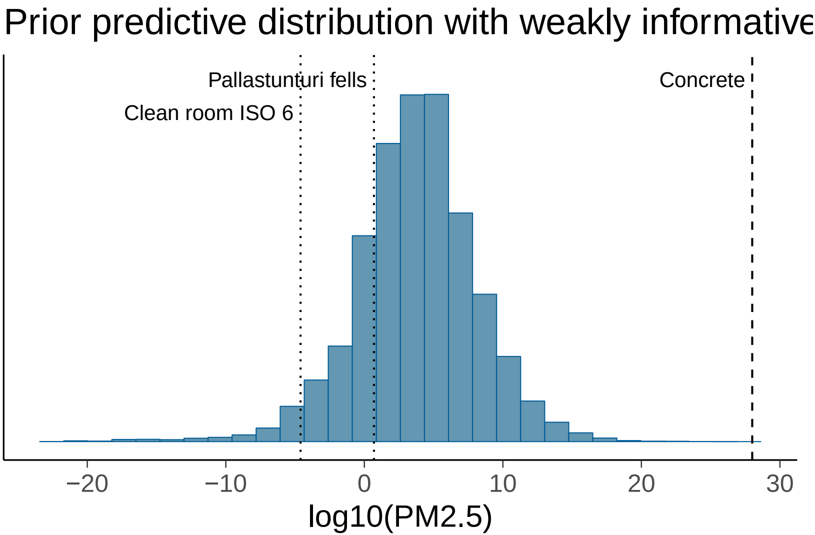
Prior predictive checking: Exposure to air pollution

Prior predictive checking



Prior predictive checking: Exposure to air pollution

Prior predictive checking



Further reading and examples

- Gabry, Simpson, Vehtari, Betancourt, and Gelman (2019). Visualization in Bayesian workflow. https://doi.org/10.1111/rssa.12378.
- Säilynoja, Johnson, Martin, and Vehtari (2025). Recommendations for visual predictive checks in Bayesian workflow. https://teemusailynoja.github.io/visual-predictive-checks/
- Graphical posterior predictive checks using the bayesplot package http://mc-stan.org/bayesplot/articles/graphical-ppcs.html
- brms demos https://avehtari.github.io/BDA_R_demos/demos_rstan/brms_demo.html

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
    `priorsense` and `adjustr` packages use importance sampling for faster prior sensitivity analysis

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
    `priorsense` and `adjustr` packages use importance sampling for faster prior sensitivity analysis
    - Kallioinen, Paananen, Bürkner, and Vehtari (2024). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. https://doi.org/10.1007/s11222-023-10366-5

Sensitivity analysis

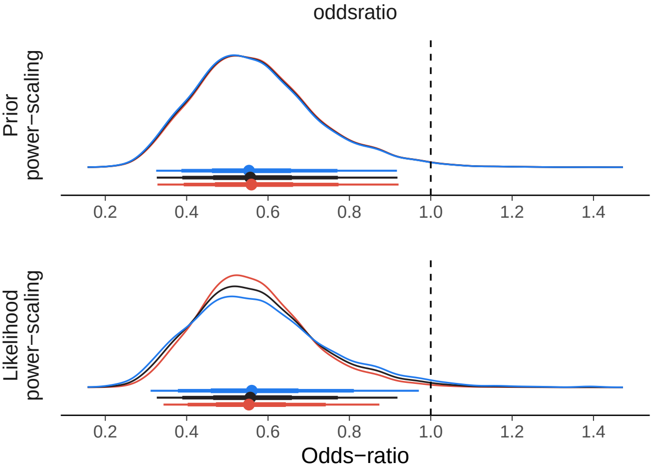
- How much different choices in model structure and priors affect the results
  - test different models and priors
    `priorsense` and `adjustr` packages use importance sampling for faster prior sensitivity analysis
    - Kallioinen, Paananen, Bürkner, and Vehtari (2024). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. https://doi.org/10.1007/s11222-023-10366-5
  - 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

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
    `priorsense` and `adjustr` packages use importance sampling for faster prior sensitivity analysis
    - Kallioinen, Paananen, Bürkner, and Vehtari (2024). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. https://doi.org/10.1007/s11222-023-10366-5
  - 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

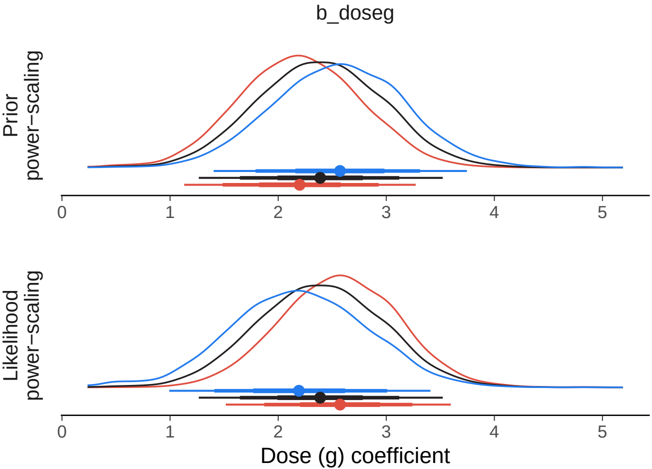
priorsense — prior and likelihood sensitivity analysis

- Power-scale prior and likelihood separately as $p(\theta)^\alpha$ and $p(y|\theta)^\alpha$
- Beta blockers — randomized control-treatment experiment
  - no prior sensitivity
  - likelihood is informative



priorsense — prior and likelihood sensitivity analysis

- Power-scale prior and likelihood separately as $p(\theta)^\alpha$ and $p(y|\theta)^\alpha$
- Sorafenib Toxicity — Binomial model meta analysis
  - prior-data conflict



priorsense — prior and likelihood sensitivity analysis

- Power-scale prior and likelihood separately as $p(\theta)^\alpha$ and $p(y|\theta)^\alpha$
- Sorafenib Toxicity — Binomial model meta analysis
  - prior-data conflict
  - due to accidentally too narrow prior