Applied Bayesian Modeling module 11:
Model checking - Part 1 (in-sample)

Leontine Alkema, lalkema@umass.edu
Fall 2022

Model checking

- We can now fit a whole range of Bayesian models (using Stan).
- Important question: how well does a model fit the data?
- George Box: "All models are wrong some are useful".
- T. Tarpey (http://andrewgelman.com/wp-content/uploads/2012/03/tarpey.pdf)
  - "This quote is useful ... but wrong"; "All models are right but most are useless".
  - All models are approximations to the truth, some are useful approximations to the truth.

Model checking: how?

- What "goodness of fit" outcomes you need to check depends on your outcome of interest.
- Example
  - For radon data set, if a model is to be used for risk assessment, users may want to make sure that the model does not *underpredict* radon outcomes.
  - For educational testing, with interest in identifying students who perform below average, users want to make sure that the model does not *overpredict*.
- Checking a Bayesian model: checking appropriateness of the probability distribution for the data, imposed by prior and likelihood function
- To discuss:
  - General diagnostic plots, e.g. using residuals
  - Posterior predictive checks
  - Measures of predictive accuracy based on out-of-sample validation
- All of these are based on generating data from a fitted model, and comparing the "new" data to what's observed

Generating replicated data sets

- Approach to generate replicated data sets from a fitted model:
  - For each posterior sample $s = 1, 2, \ldots, S$, generate a new replicated data set $\tilde{\boldsymbol{y}}^{(s)} = (\tilde{y}_1^{(s)}, \tilde{y}_2^{(s)}, \ldots, \tilde{y}_n^{(s)}) \sim p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$.
  - This results in S replicated data sets.
- Can we do that?

  Yes, for any model, we can generate new data points $\tilde{\boldsymbol{y}} \sim p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$
- Minor note on notation:
  - BDA prefers y^{rep} over $\tilde{y}$ to indicate that it is a replicate for a specific y, as opposed to any model-based prediction.
  - Here (as in Gabry et al) we stick to $\tilde{y}$ to avoid too many subscripts and are explicit what it refers to

Generating replicated data sets

- Example for the radon data (module 8):
  Suppose

$$y_i|\alpha_{j[i]}, \sigma_y^2 \quad \sim \quad N(\alpha_{j[i]}, \sigma_y^2), \text{ (independent)}$$

  then we can sample each $\tilde{y}_i \sim p(\tilde{y}_i|\boldsymbol{y})$ in two steps:
  (1) Sample $(\alpha_{j[i]}^{(s)}, \sigma_y^{(s)}) \sim p(\alpha_{j[i]}, \sigma_y|, \boldsymbol{y})$, (we already have these)
  (2) Sample $\tilde{y}_i^{(s)} \sim p(\tilde{y}_i|\alpha_{j[i]}^{(s)}, \sigma_y^{(s)})$, here $\tilde{y}_i|\alpha_{j[i]}, \sigma_y^2 \sim N(\alpha_{j[i]}, \sigma_y^2)$.
- Details: this produces a sample $\tilde{y}_i \sim p(\tilde{y}_i|\boldsymbol{y})$ because
  $p(\tilde{y}_i|\boldsymbol{y}) = \int\int p(\tilde{y}_i|\alpha_{j[i]}, \sigma_y) p(\alpha_{j[i]}, \sigma_y|, \boldsymbol{y}) d\alpha_{j[i]} d\sigma_y$
- You can generate data sets in `R` or let brm/rstan do the work for you.

Creating replicated data sets $\tilde{y}^{(s)}$ in brm and with rstan

- brm function posterior_predict

```
ynew_si <- posterior_predict(fit) # adding si to indicate the dimension used
dim(ynew_si)
```

```
## [1] 2000  927
```

- When using rstan, add y_new to generated quantities block

```
33 generated quantities {
34   vector[N] mu;
35   for (i in 1:N)
36     mu[i] = mu_alpha + eta[county_id[i]] + x[i] * beta;
37
38   vector[N] y_new;  // replications from posterior predictive dist
39   for (i in 1:N)
40     y_new[i] = normal_rng(mu[i], sigma_y);
41
```

Back to big picture: In-sample checks using replicated data sets (ctd)

- Approach:
  - For each posterior sample $s = 1, 2, \ldots, S$, generate a new data set $\tilde{\boldsymbol{y}} = (\tilde{y}_1^{(s)}, \tilde{y}_2^{(s)}, \ldots, \tilde{y}_n^{(s)}) \sim p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$. This results in S replicated data sets.
  - Use these data sets in checks (compare to the observed data)
- To discuss: residuals, posterior predictive checks
- We illustrate functionality of the `bayesplot` package.

Residuals

- Example: radon model

$$y_i|\alpha_{j[i]}, \sigma_y^2 \quad \sim \quad N(\alpha_{j[i]}, \sigma_y^2), \text{ (independent)}$$

- For each sample s, we can calculate a residual for ith observation as $e_i^{(s)} = y_i - \tilde{y}_i^{(s)}$
- We can report a summary, e.g. $e_i = y_i - 1/S \sum_s \tilde{y}_i^{(s)}$ (MC approximation to y_i - $E(\tilde{y}_i|\boldsymbol{y})$), and use these in standard diagnostics plots
  - Check for deviations away from zero in the residuals wrt covariates

Residuals for subset model: old-school coding and plot

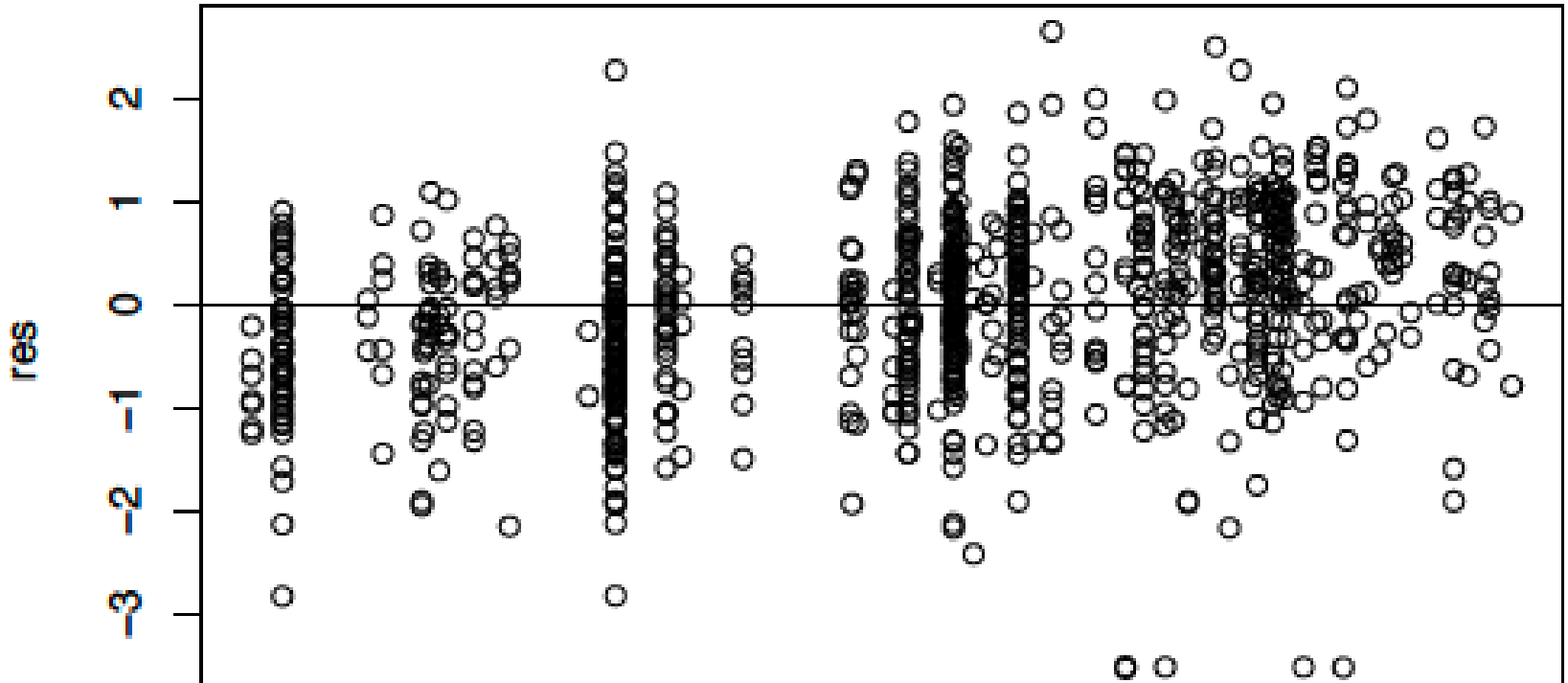
Residuals

Obtain point estimates from the replicated data

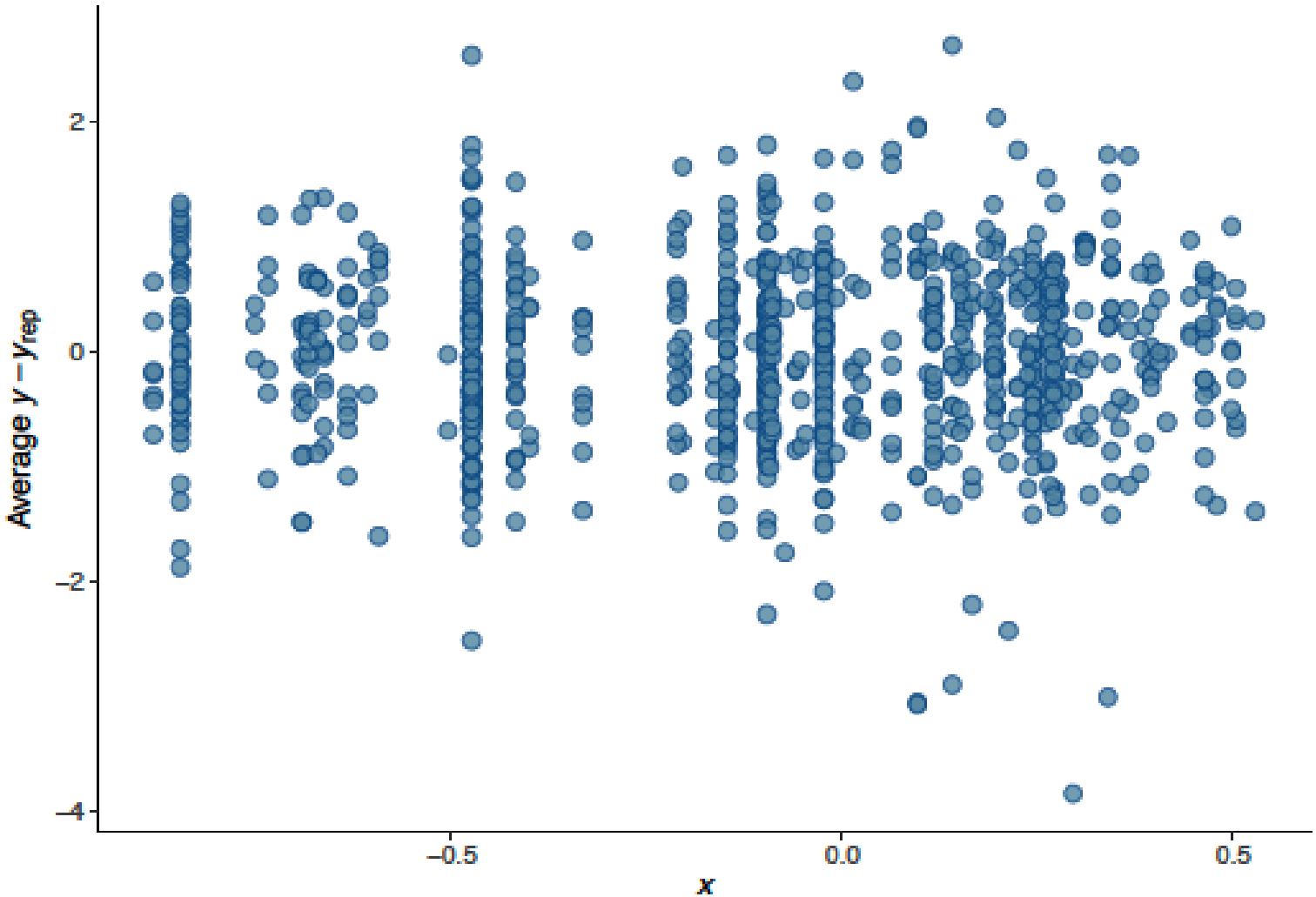
```
ytildehat_i <- apply(ynew_si, 2, mean)
res <- dat$y - ytildehat_i
```

and make some old-school plots (that I am sure you can improve upon :))

```
plot(res - dat$log_ur)
abline(h=0)
```



Residuals for full model: using `bayesplot`



Posterior predictive checks

- Main idea: a comparison of (summaries of) replicated data sets to an observed dataset can reveal problems with model fit/assumptions.
- Approach:
  - Simulate data from the fitted model.
  - Compare simulated data to the observed data.
  - Check whether those outcomes which you are most interested in are replicated well.

Comparing simulated data sets to an observed data set

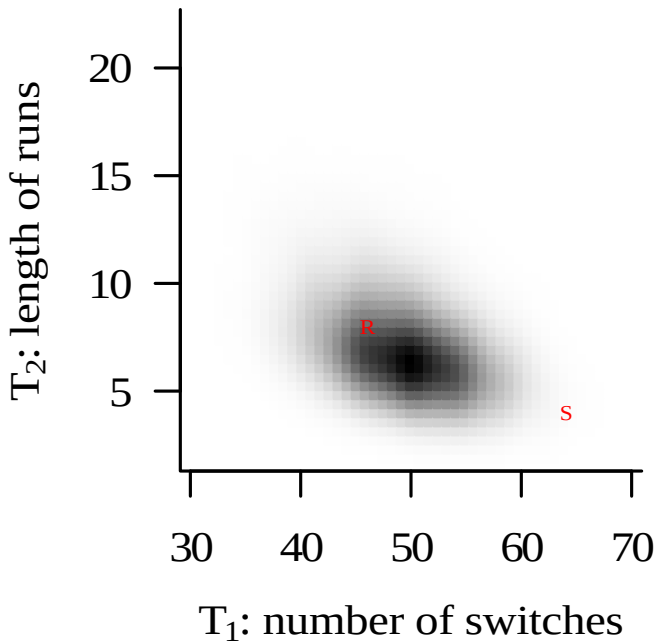
- Example: "Coin tosses and fake coin tosses" (from Alex Cook, NUS)
- Suppose two people obtain a sequence of heads and tails:
  - Leontine the statistician simulates the series in R using independent Bernouilli draws (or actually tosses a coin 200 times).
  - Her 6-yo daughter writes down a sequence.

How can we check which series was generated by Leontine?

Data series A	Data series B
TTHTHHTHHH	HHHTTTHTHH
THTTHTTHTH	THHHHHHHHT
TTHHTHHTHT	THTTTTHTTH
HTHTTTHHTH	THHTTHHHTT
THHHHTTHTH	TTHHTHHHHH
HHTHHTHHTT	TTTTTHTHTH
THHHTHTHTT	TTTHHHTTTT
HHTHTHHTTT	HTTTTTHTHH
HHTHHTHHTH	THTTHHTTTH
THHTHHTHTT	THTTHHHTTH

Summary statistics

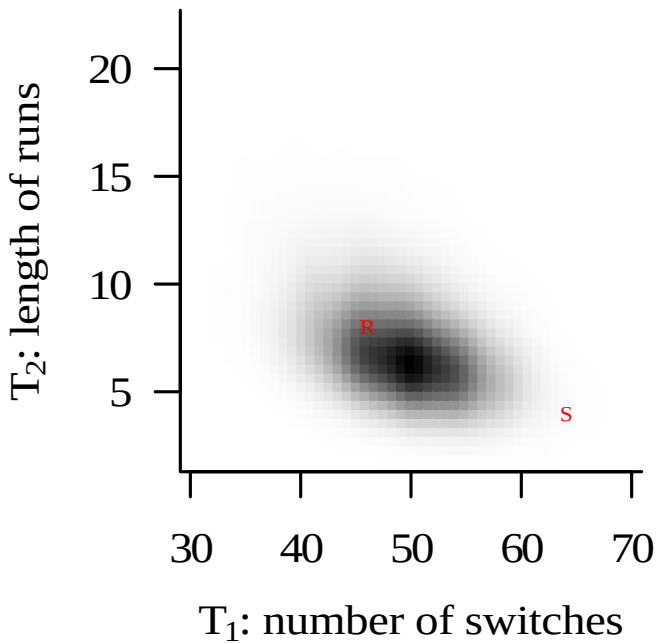
- We may be able to figure out which series was not simulated based on two data summary statistics:
  - T_1, the number of switches from H to T or T to H;
  - T_2, the maximum run of consecutive Hs or Ts.
- We can calculate these statistics for the two data series.
- We can also calculate these statistics for simulated data series, where each simulation is given by a series of 200 random draws from a Bernouilli distribution with $p = 0.5$.



Comparing the observed values to the distribution of expected values suggests which series is likely to be made up.

Introducing some notation

- Let $\boldsymbol{y}$ the real data; here we have data $\boldsymbol{y}^{(A)}$ and $\boldsymbol{y}^{(B)}$
- $\tilde{\boldsymbol{y}}^{(s)}$ denote simulated data from a model ($y_i \sim Bern(0.5)$, independent draws)
- We define summary statistics $T_1(\boldsymbol{y})$ = number of switches, $T_2(\boldsymbol{y})$ = maximum run



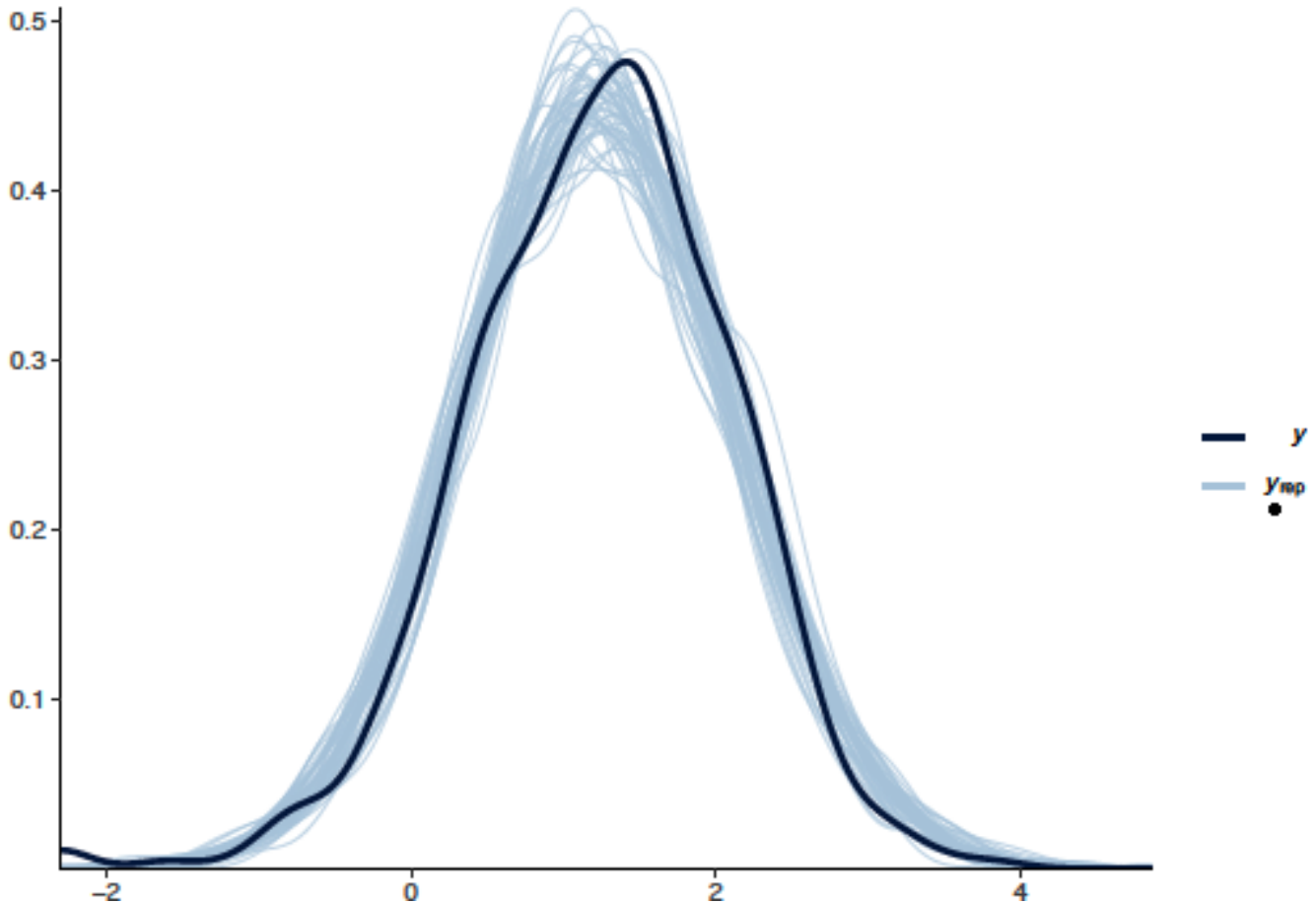
If the model generated the observed data, then we do NOT expect the observed summary statistic to be extreme, relative to the simulated summary statistics $T(\tilde{\boldsymbol{y}})$.

Back to using posterior predictive checks for model checking

- If model assumptions are reasonable then we should be able to use the fitted model to generate data that resemble the data we observed.
  - Basis: 'Simulate data from the fitted model';
    Simulate replicated data sets from the posterior predictive distribution.
  - With the replicated data sets:
    - graphical checks: compare (some summary or subset of) replicated data sets to real data
    - define summary statistic(s) and compare the observed summary statistic in your data to the posterior sample of replicated summary statistics.

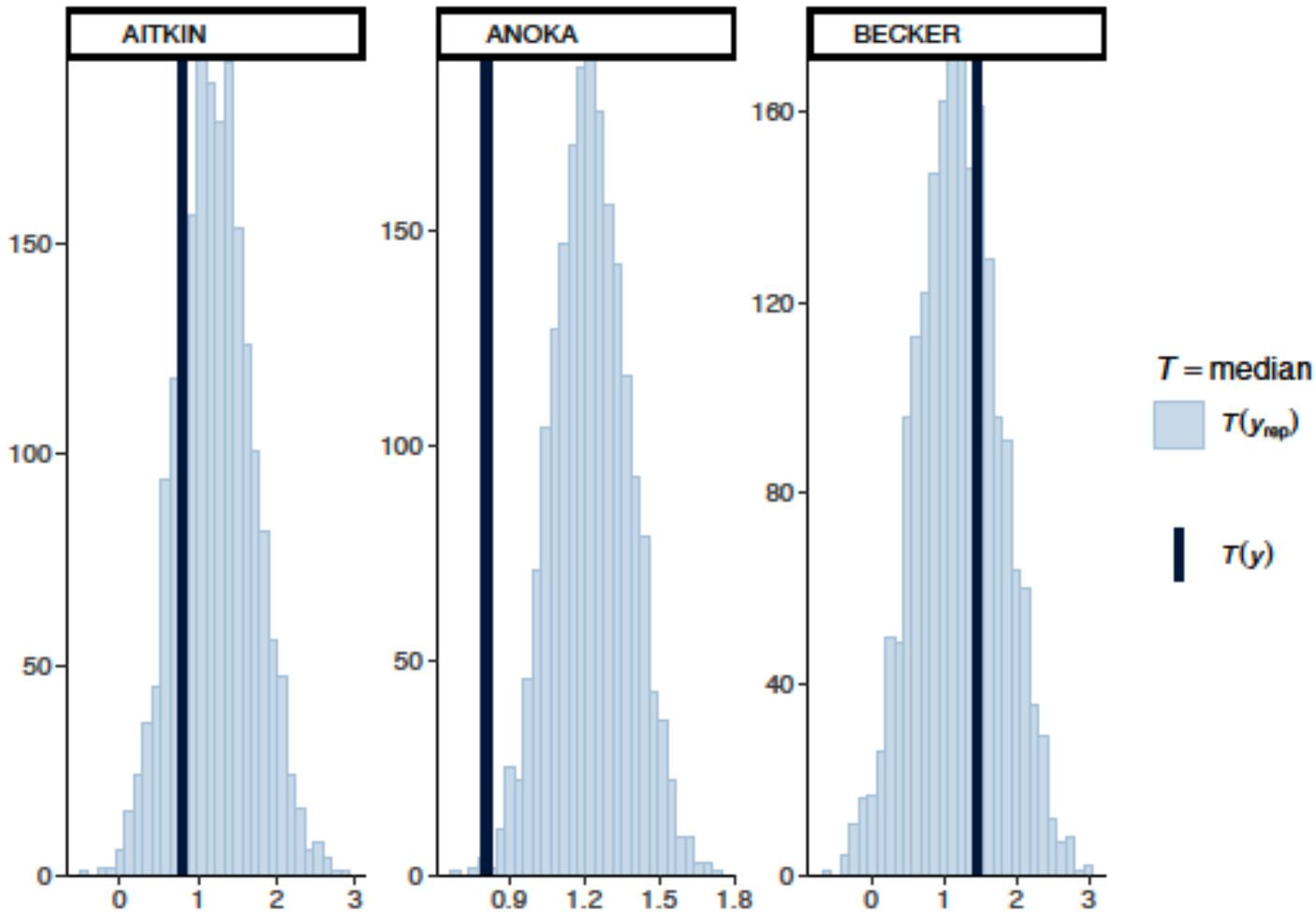
Posterior predictive checks for radon data I

Simple comparison of observed density and examples of replicated data sets



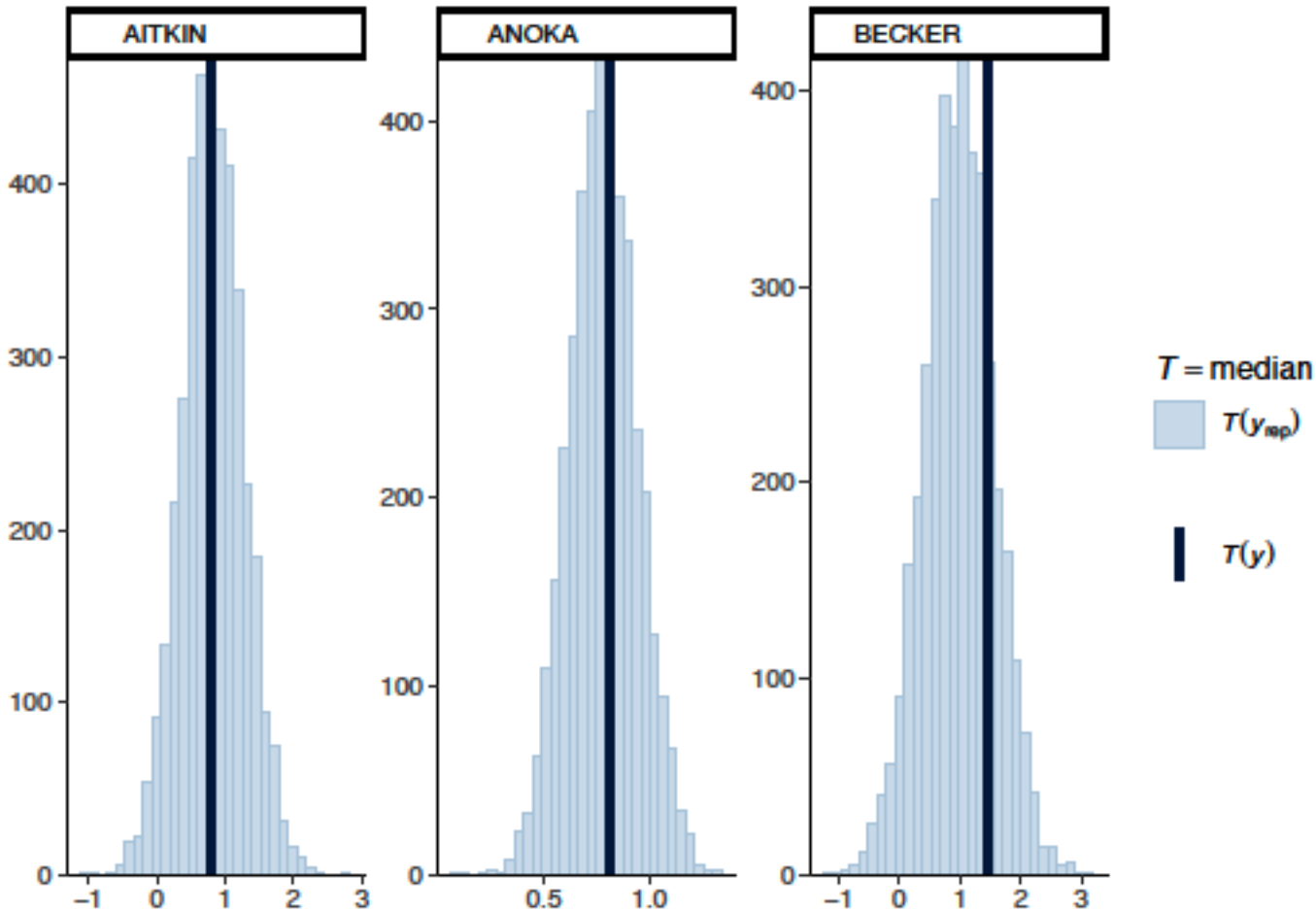
Posterior predictive checks for radon data II

- Summary outcome = county-specific median
- Does this look ok for model w/o county intercepts?



Posterior predictive checks for radon data II

- Summary outcome = county-specific median
- What about for the model with county intercepts?

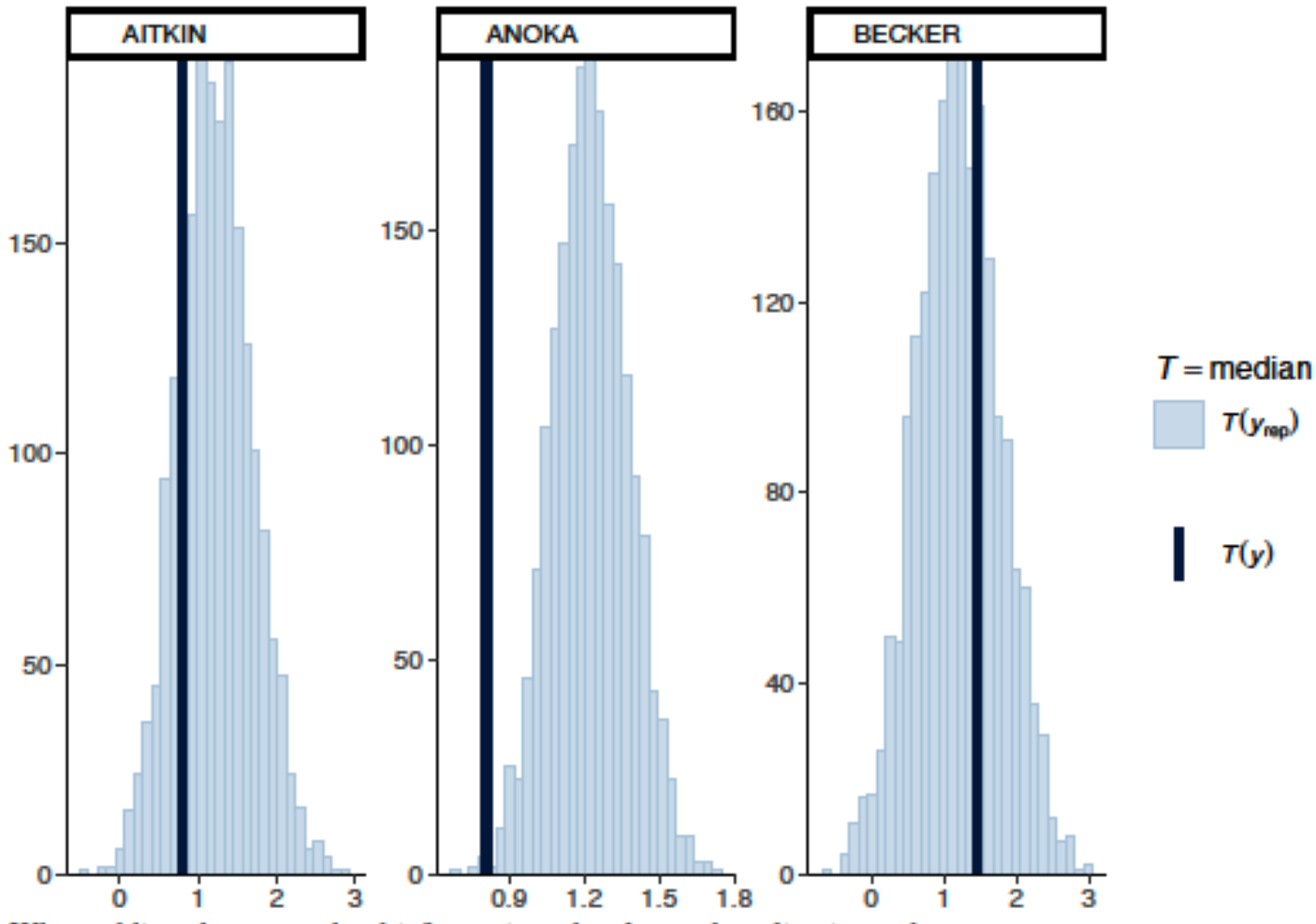


Test quantities/discrepancy measures

- Check outcomes that you're interested in using test quantity or discrepancy measure $T(\boldsymbol{y})$ or $T(\boldsymbol{y}, \boldsymbol{\theta})$, e.g. median in a group
  - Calculate these statistics for the real data, $T(\boldsymbol{y})$, and the replicated data, $T(\tilde{\boldsymbol{y}})$
  - Display and/or calculate posterior predictive p-value: probability that replicated data is more extreme than the observed data.

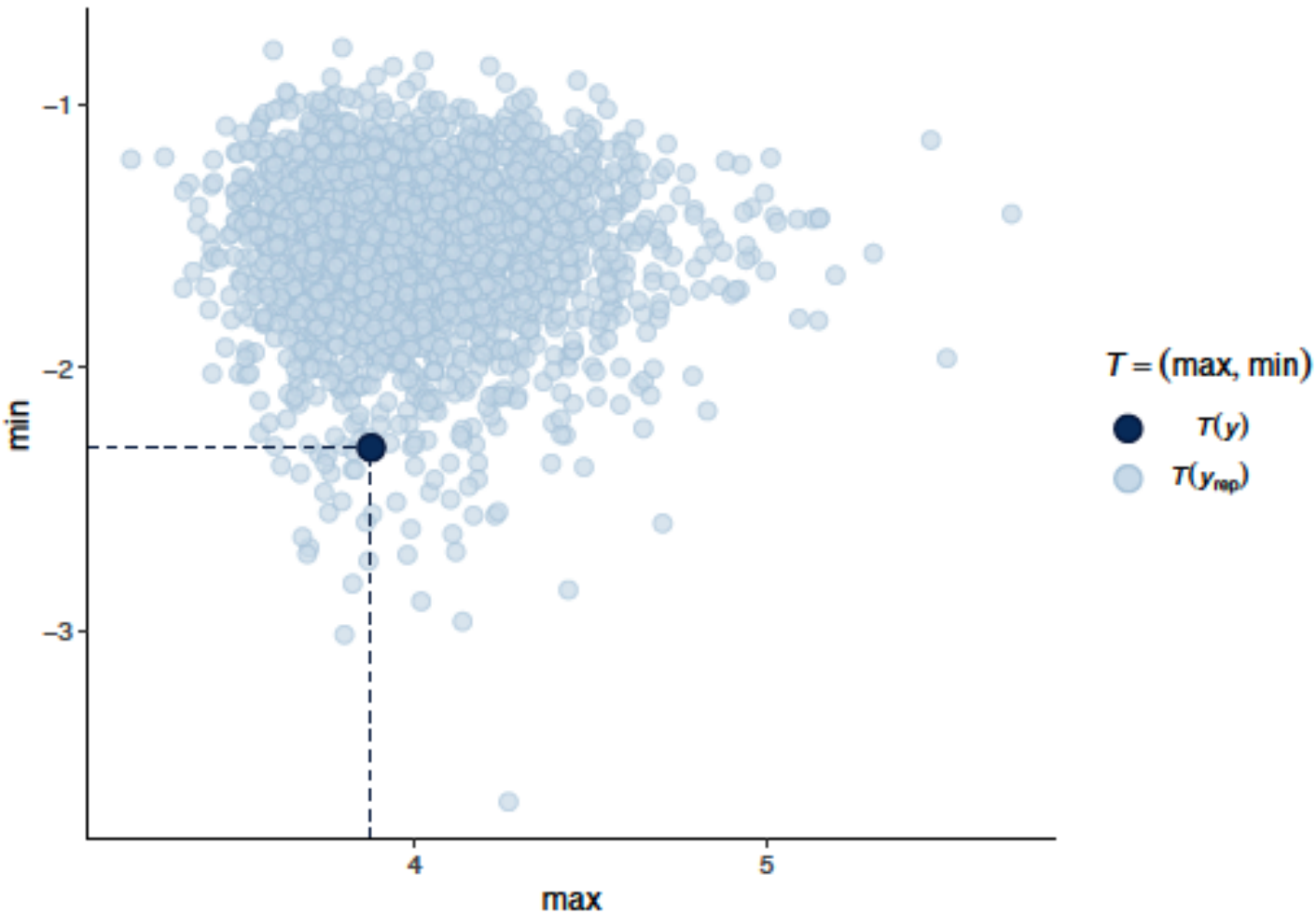
Posterior predictive checks for radon data

- Test quantity $= T(\boldsymbol{y}) =$ median outcome in a county
- Posterior predictive p-value close to 0 for Anoka



Posterior predictive checks for radon data: maximum

- Test quantity $= T(\boldsymbol{y}) = \max_i\{y_1, \ldots, y_n\}$ (minimum shown too)
- See code for calculation of probability



Summary posterior predictive checks

- All models are approximations to the truth, some are useful approximations to the truth.
- Posterior predictive checks can be very useful to check if model assumptions are reasonable and to inform model improvements.
  - Basis: 'Simulate data from the fitted model';
    Simulate replicated data sets from the posterior predictive distribution.
  - With the replicated data sets:
    - compare some replicated data sets to real data
    - define summary statistic(s) and compare the observed summary statistic in your data to the posterior sample of replicated summary statistics.
- Part 2: model checking based on (approximate) cross-validation