L5. Model Checking, Evaluating, Comparing and Expanding models

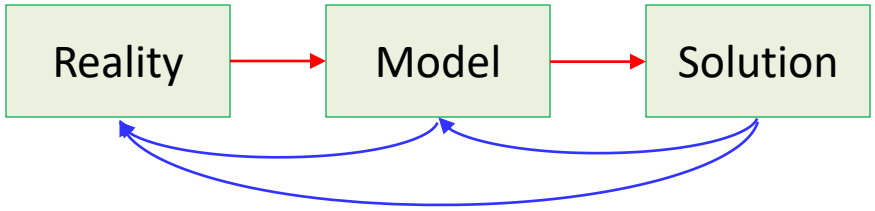
Model checking is a crucial step

1. Constructing a probability model

2. Computing the posterior distribution of all estimands

3. Assessing the fit of the model to the **data** and to our substantive **knowledge**

It is impossible to include in a probability distribution all of one's knowledge about a problem, so it is wise to investigate what aspects of reality are not captured by the model

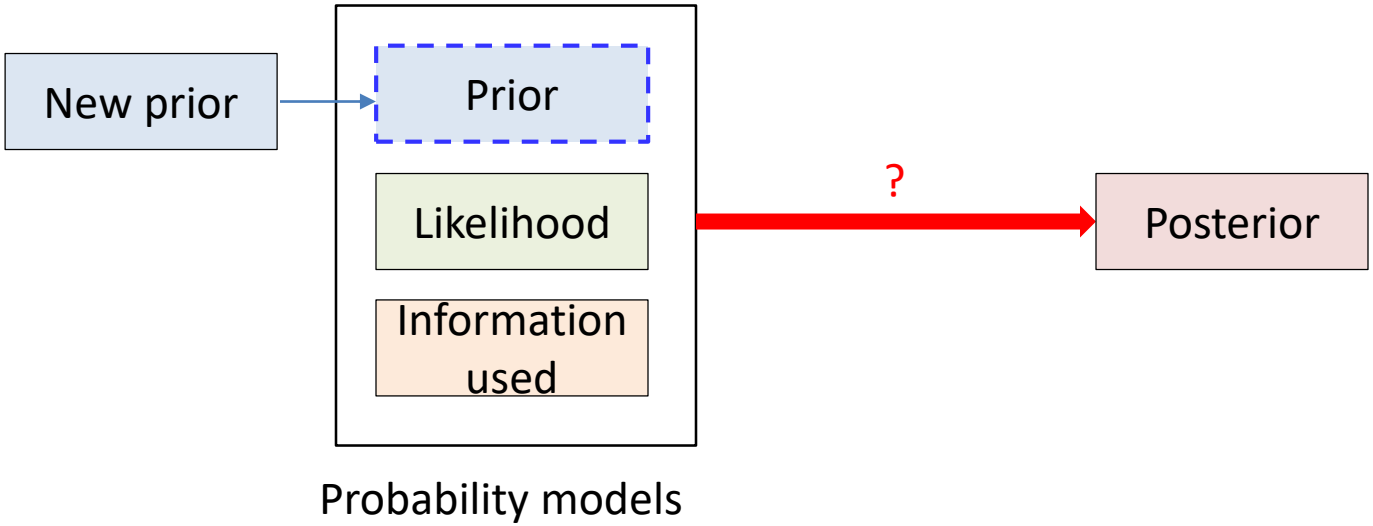


We need to check….

- prior
- sampling distribution
- hierarchical structure
- explanatory variables…

Sensitivity analysis :

how much do posterior inferences change when other reasonable probability models are used in place of the present model?



We need to examine:

- How our model fails to fit reality
- How sensitive the resulting posterior distributions are to arbitrary specifications

Sensitivity analysis

Is our model true or false?

v.s.

Do the model's deficiencies have a noticeable effect on the substantive inferences?

- Rat tumor example : Beta population distribution for tumor rates
- Eight schools example : Normal distribution for the eight school effects

→ Little bit arbitrary and convenience chaise, but they have little impact on the inferences of most interests

How to judge when *assumptions of convenience* can be made safely is a central task of Bayesian sensitivity analysis

Do the inferences from the model make sense?

For reasons of convenience or objectivity,
there will be knowledge that is not included formally in either the prior or likelihood

If the additional information suggests that posterior inferences of interest are false

Potential for creating a more accurate probability model for the parameters and data collection process

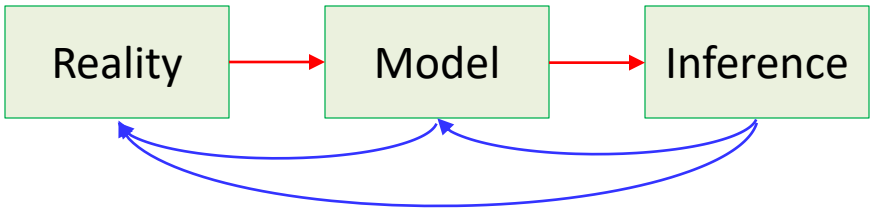
External validation

- Prediction about future data v.s. collected external data
- Often we need to check the model before obtaining new data or waiting for the future to happen

→ Require a method of approximating external validation using the data we already have

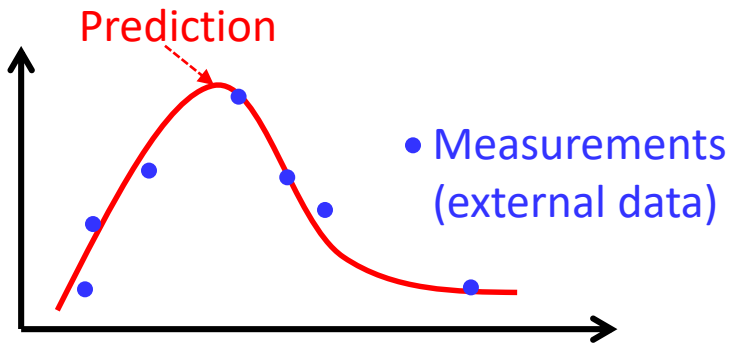
Posterior predictive checking
- Use global summaries to check the **joint predictive distribution** $p(\tilde{y}|y)$

Do the inferences from the model make sense?



When inference does not capture reality,
we need more accurate probability model for the parameters and data collection process

External validation



- Prediction about future data v.s. collected external data
- When there is no external data received yet,

→ Require a method of approximating external validation using the data we already have

Two approaches of modal checking

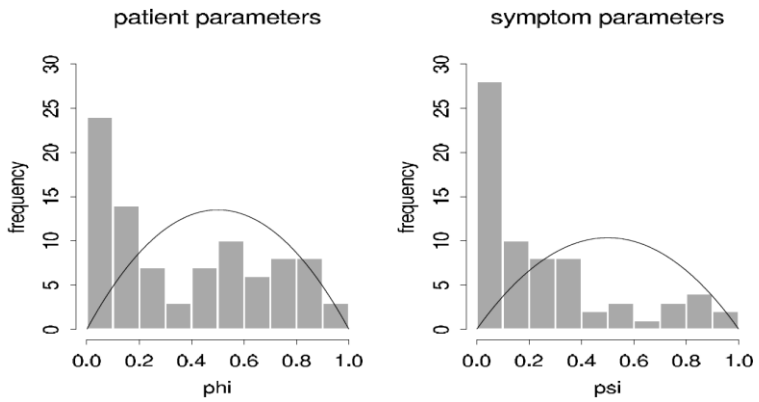
- **Posterior predictive checking**

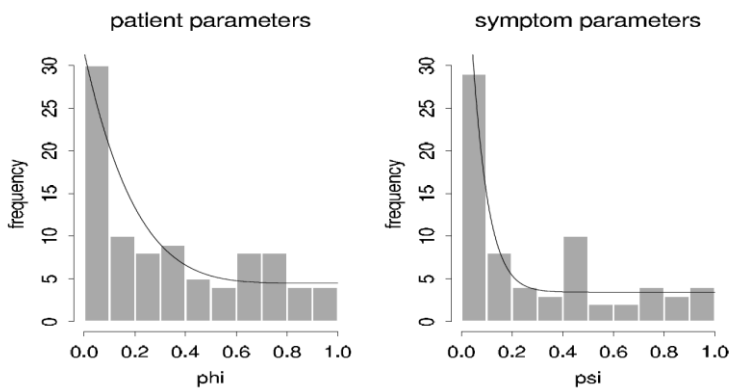
$$p(\tilde{y}|y)$$

$\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_m)$: joint prediction on future data

$y = (y_1, \ldots, y_n)$: observed data

- **Graphical Checking**

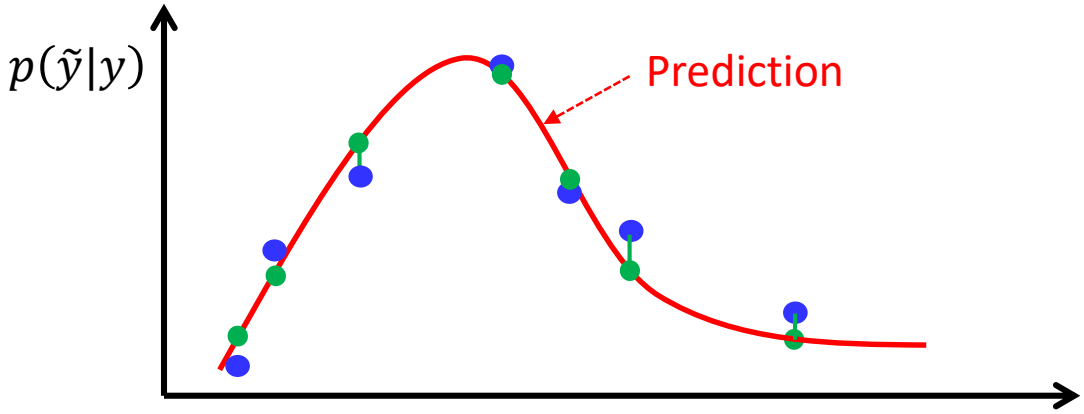




Posterior predictive checking

Self-consistency check:

Replicated data generated under the model should look similar to observed data. That is, the observed data should look plausible under the posterior predictive distribution



y : observed data

y^{rep} : Replicated (training) data

$\tilde{y}$: prediction on future data

Posterior predictive checking

y^{rep} is replication (simulation) of the observed data y using the trained model

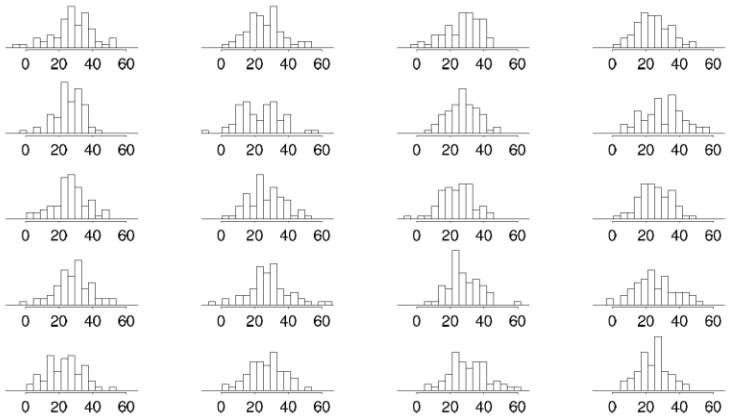
Observed : x → y, y^{rep} Unobserved : $\tilde{x}$ → $\tilde{y}$

Posterior predictive distribution on the replication is

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$$

prior	$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$
Likelihood	$y|\mu, \sigma^2 \sim N(\mu, \sigma^2)$

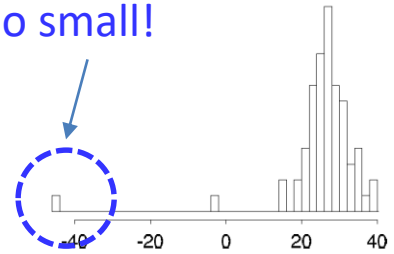
$p(\mu, \sigma^2|y)$ → Sample (μ, σ^2) → sample 66 y_i^{rep}



Replicated $p(y^{rep}|y)$

Is it similar?

Too small!



Actual measurement $p(y)$

Posterior predictive checking

y^{rep} is replication (simulation) of the observed data y using the trained model

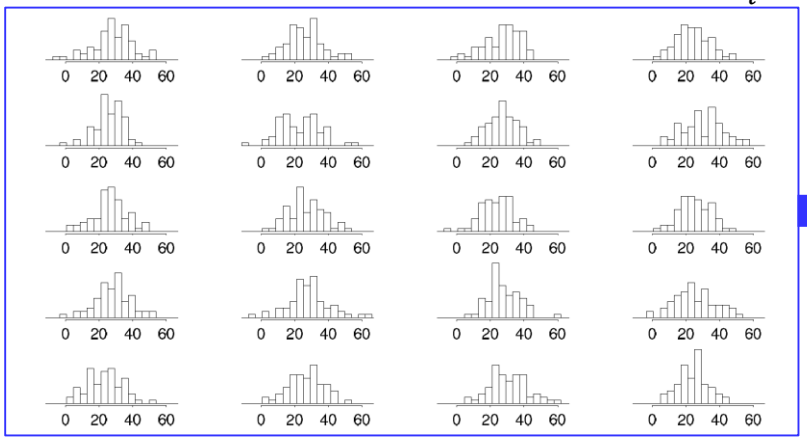
Observed : x → y, y^{rep} Unobserved : $\tilde{x}$ → $\tilde{y}$

Posterior predictive distribution on the replication is

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$$

prior	$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$
Likelihood	$y|\mu, \sigma^2 \sim N(\mu, \sigma^2)$

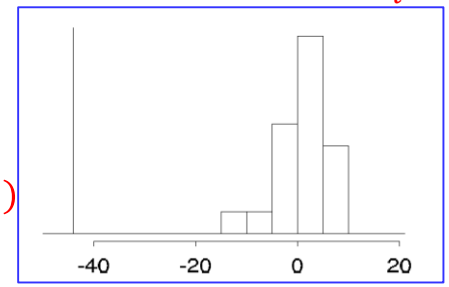
$p(\mu, \sigma^2|y)$ → Sample (μ, σ^2) → sample 66 y_i^{rep}

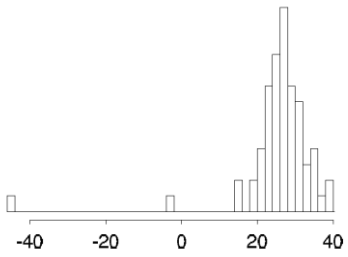


Replicated $p(y^{rep}|y)$

Test statistics $T(y^{rep}) = \min_i y_i^{rep}$

$T(y^{rep})$





The model does not capture the real data well in terms of $T(y^{rep}) = \min_i y_i^{rep}$

(Posterior predictive distribution is represented as histogram)

Posterior predictive checking

Test quantities

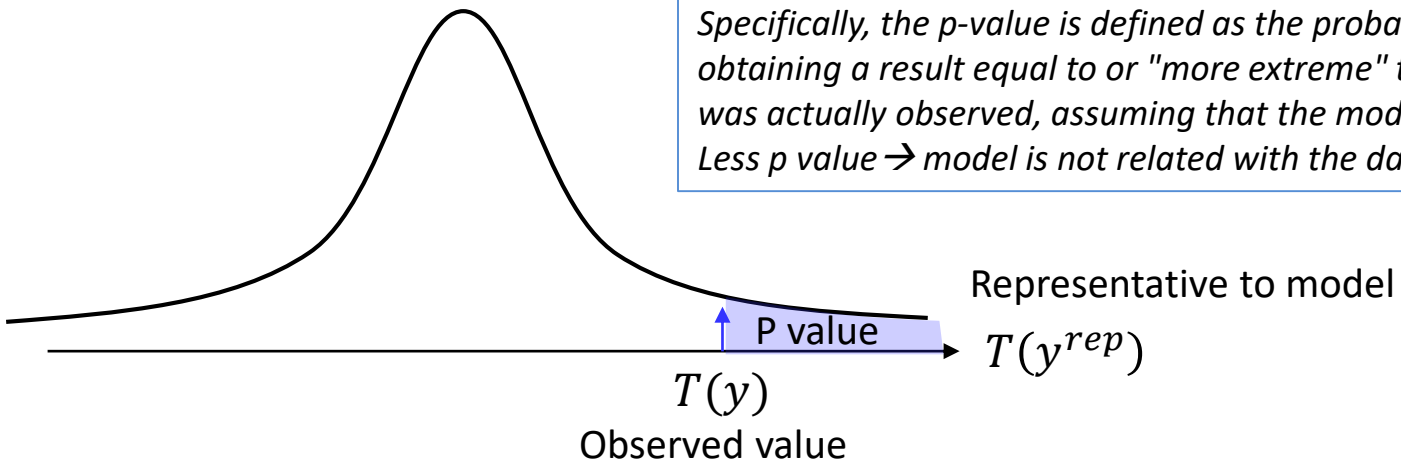
- In cases with less obvious discrepancies than the outliers in the speed of light data, it is often also useful to measure the "statistical significance" of the lack of fit
- Measure the discrepancy between model and data by defining *test quantities*
  - ✓ Classical setting : $T(y)$, called test statistic
  - ✓ Bayesian setting : $T(y, \theta)$, which take into account the dependency on the model parameter θ

The procedure for carrying out a posterior predictive model check

1. Specify a test quantity, $T(y)$ or $T(y, \theta)$
2. Specify an appropriate predictive distribution for the replications y^{rep}
3. Measure discrepancy between $T(y)$ and $T(y^{rep})$, for example using $p - value$

Posterior predictive checking

Tail area probability



Specifically, the p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming that the model is true
Less p value → model is not related with the data

- Classical $p-$values : $p_C = Pr(T(y^{rep}) \geq T(y)|\theta)$: Parameter is fixed

  θ is fixed and should be substituted

- Bayesian $p-$values : $p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y)$: Data is fixed

$$= \int \int I_{T(y^{rep},\theta) \geq T(y,\theta)} p(y^{rep}, \theta|y) dy^{rep} d\theta$$

✓ Where the probability is taken over the posterior distribution of θ and the posterior predictive distribution of y^{rep}, i.e., $p(\theta, y^{rep}|y)$

✓ Usually, the posterior predictive distribution can be computed using simulation :
count the number of $T(y^{rep_s}, \theta^s) \geq T(y, \theta^s), s = 1, \ldots, S$

Posterior predictive p-values

- An extreme p-values implies that the model cannot be expected to capture this aspect of the data.

→ Extreme tail-area probabilities (less than 0.01 or higher than 0.99) can be addressed by expanding the model appropriately.

- Typically, we evaluate a model with respect to several test quantities

- Wrong question : Do the data come from the assumed model?
- Correct question : Quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model's own assumptions.

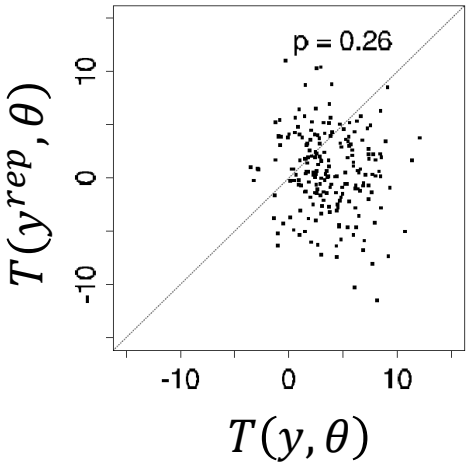
- Bayesian predictive checking generalizes classical hypothesis testing by averaging over the posterior distribution of the unknown parameter vector θ rather than fixing it at some estimate $\hat{\theta}$.

Speed of light example with test quantity

- Use other test quantities to illustrate how the fit of a model depends on the aspects of the data and parameters being monitored.
- *Assess whether the model is adequate except for the extreme tails by considering a model check based on a test quantity sensitive to asymmetry in the center of the distribution*

$$T(y, \theta) = |y_{61} - \theta| - |y_6 - \theta|$$

- The 61st and 6th order statistics are chosen to represent approximately the 90% and 10% points of the distribution.
- The test quantity should be scattered about zero for a symmetric distribution.



Choosing test quantities

Example: Check the independence assumption

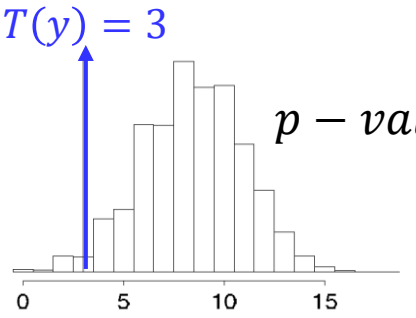
- A sequence of binary outcomes, $y_1, \dots, y_n$, modeled as specified number of independent trial with a common probability of success, θ, that is given a uniform prior distribution.
- The posterior distribution is $p(\theta|y) \propto \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}$

Perform a posterior predictive test using with

$T(y) =$ number of switches between 0 and 1

Observed data: $y = (y_1, \dots, y_n) = (1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0) \rightarrow T(y) = 3$

1. draw θ^s for $s = 1, \dots, S = 1{,}0000$
2. for each s draw $y^{rep_s} = (y_1^{rep_s}, \dots, y_{20}^{rep_s})$
3. compute $T(y^{rep_s})$



$p - value$ = 0.9838:

The discrepancy cannot be explained by chance, but is likely to be explained by modeling failure

Choosing test quantities

- Posterior predictive checking is a useful direct way of assessing the fit of the model to these various aspects of the data.

- Ideally, the test quantities T will be chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied.

- Test quantities are commonly chosen to measure a feature of the data not directly addressed by the probability model; for example, ranking and correlation

Interpreting posterior (Bayesian) p-values

- The goal of examining p-value is not to answer the question 'Do the data come from the assumed model?, but to quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model's own assumption

- An extreme p-value implies that the model cannot be expected to capture this aspect of the data

- Bayesian predictive checking generalized classical hypothesis testing by averaging over the posterior distribution of the unknown parameter vector θ rather than fixing it at some estimate $\widehat{\theta}$

Marginal predictive checks

- The focus has been on replicated data from the joint posterior predictive distribution

  → An alternative approach is to compute the probability distribution for each marginal prediction $p(y_i^{rep}|y)$ separately and then compare these separate distributions to data in order to find outliers or check overall calibration.

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

If y_i is scalar,

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

A related approach is to replace predictive distributions with cross-validation predictive distributions, for each data point comparing to the inference given all the other data.

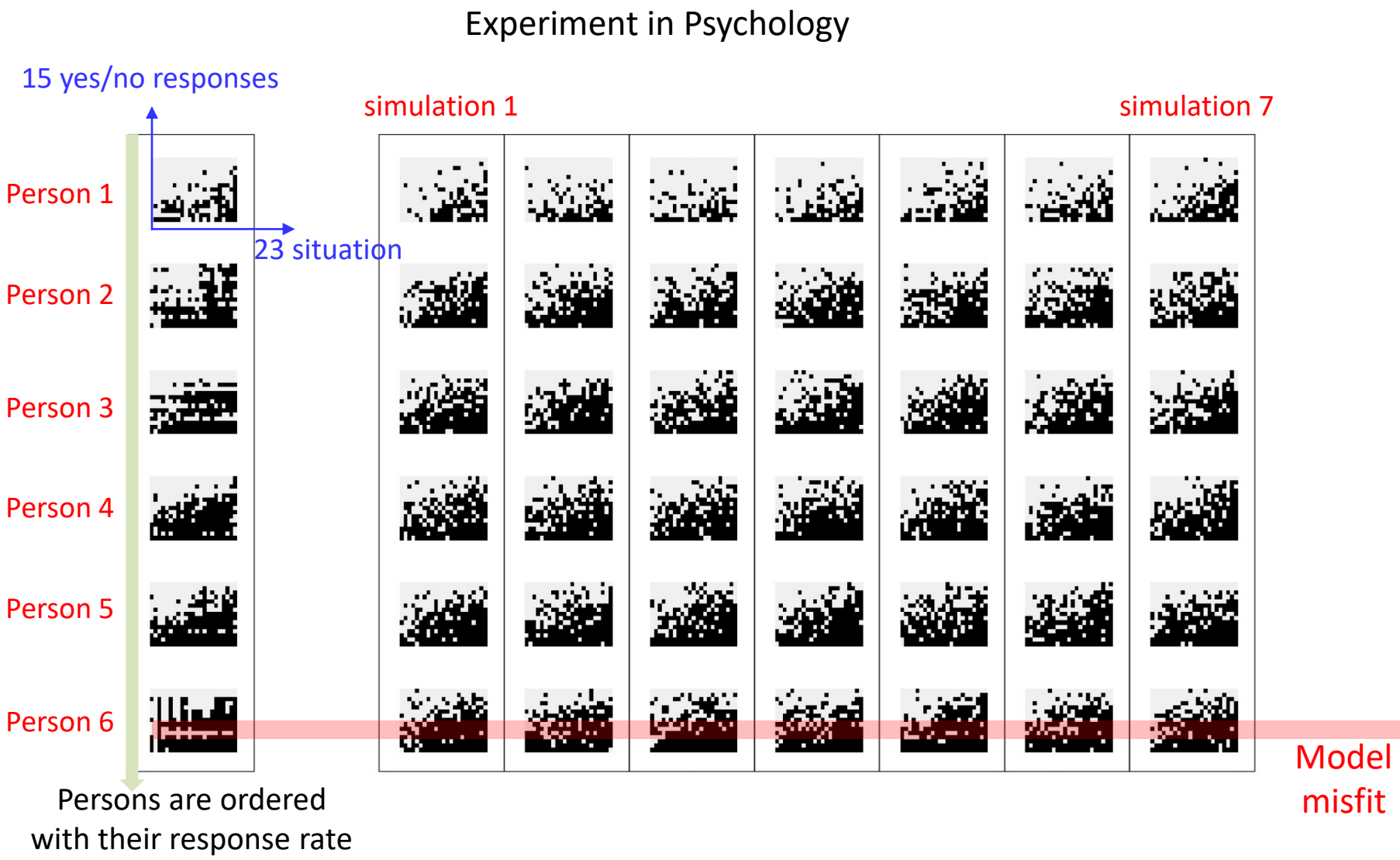
$$p_i = \Pr\left(y_i^{rep} \leq y_i|y_{-i}\right)$$

Graphical posterior predictive checks

Graphical model checking is to display the data alongside simulated data from the fitted model, and to look for systematic discrepancies between real and simulated data from the fitted model, and look for systematic discrepancies between real and simulated data.

- Direct display of all the data
- Display of data summaries or parameter inferences
- Graphs of residuals or other measures of discrepancy between model and data

Direct display of all the data



Graphical representations with special ordering itself is a test statistics

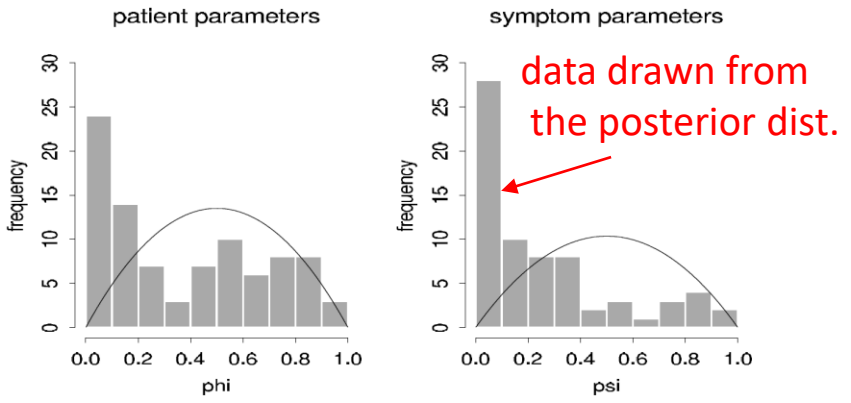
Displaying summary statistics or inferences

Inference from a hierarchical model from psychology (Model specification is now shown here)

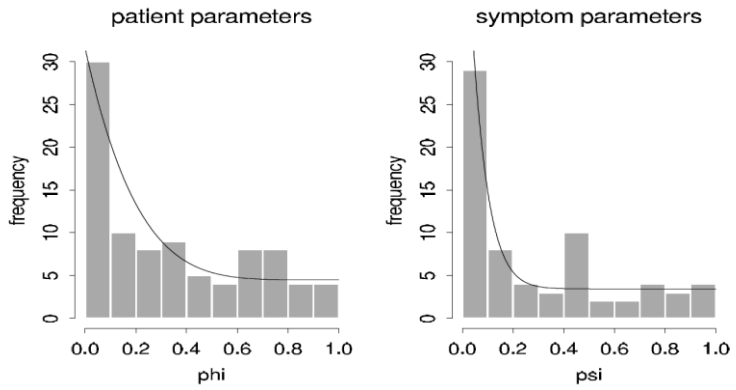
Two vector of parameters:

- Patient parameters: $\phi_1, \ldots, \phi_{90}$
- Psychological symptom parameters: $\psi_1, \ldots, \psi_{69}$

$$p(\phi_i) \sim Beta(\phi_i|2,2)$$
$$p(\psi_i) \sim Beta(\psi_i|2,2)$$



$$p(\phi_i) = 0.5\text{Beta}(\phi_i|1,6) + 0.5\text{Beta}(\phi_i|1,1)$$
$$p(\psi_i) = 0.5\text{Beta}(\psi_i|1,16) + 0.5\text{Beta}(\psi_i|1,1)$$



Prior distribution can be viewed as a posterior predictive check in which a new set of patients and symptom parameters.

Graphical model checking for the 8 schools example

8 schools example assumptions:

(1) Normality of the estimates y_j given θ_j and σ_j (known)
→randomization, large sample size,…

(2) Exchangeability of the prior distribution of the θ_j
→ we will let the data tell us about the relative ordering and similarity of effects in the schools.
→ In the absence of any information, the exchangeability assumption implies that the prior distribution of the θ_j's can be considered as independent samples from a population whose distribution is indexed by some hyper-parameters

(3) Normality of the prior distribution of each θ_j given μ and τ

$$\theta_j \sim N(\mu, \tau)$$

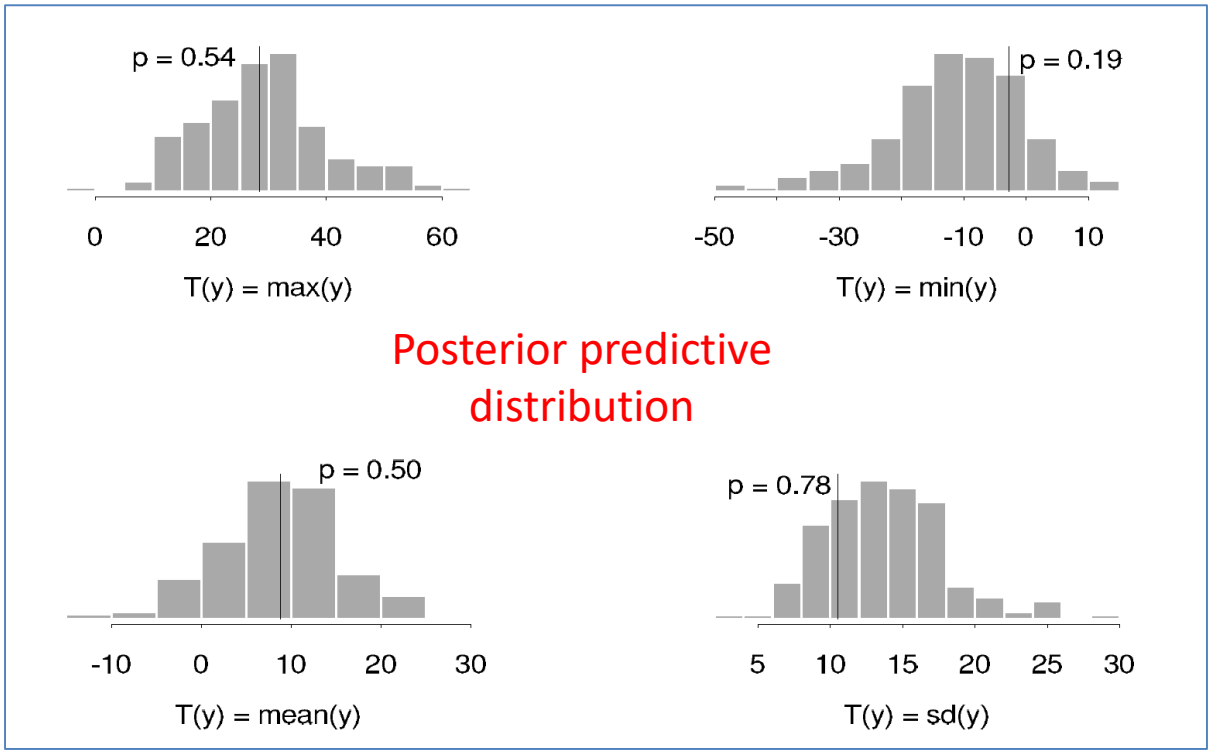
(4) Uniformity of the hyper prior distribution of (μ, τ)

$$(\mu, \tau) \sim Uniform$$

(3) And (4) are hard to justify (before seeing data)
Mathematical tractability is one reason for the choice of models, but if the family of probability models is inappropriate, Bayesian answers can be misleading

Posterior predictive checking

School Coaching Example: Suppose we perform 200 posterior predictive simulations of the coaching experiments



- The summaries suggest that the model generates predicted results similar to the observed data in the study: that is, the actual observations are typical of the predicted observations generated by the model
- Other reasonable models might provide just as good a fit but lead to different conclusions. **Sensitivity analysis** can then be used to assess the effect of alternative analyses on the posterior inferences.