Statistics 568 Bayesian Analysis

Posterior Predictive Model Checking

Ruobin Gong

Department of Statistics
Rutgers University

Week 5 - 02/18/21

Model checking in (Bayesian) statistical modeling

All statistical model building, Bayesian or otherwise, requires an examination on the **goodness of fit**.

In Bayesian model building, model fit encompasses all the following aspects:

- Sampling distribution, including
  - Model family, over- or under-parametrization;
  - Dispersion, tail-area;
  - Inclusion of covariates;
- Prior distribution, including
  - Choice of family;
  - Hierarchical structure.

Sensitivity analysis

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

- Test different models and priors;
- Combine different models into 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.

Do inferences from my model make sense?

We're not interested in whether the model is "right" or "wrong". (All models are wrong.) We want to know whether the model is **adequate** in:

- Capturing the probabilistic information in the data;
- Delivering sensible statistical inference with **external validity**.

Some peculiar examples that warrant re-examination:

- Prior-data conflict;
- *Discrepant posterior phenomenon.

Replication

Let y^{rep} denote **replicated data**: data that could have been obtained by replicating the whole experiment under the same conditions, covariates, and parameters.

This is a similar, but generally different, notion from $\tilde{y}$: the future observation yet to be observed.

In other words, y^{rep} is a replica of y, whereas $\tilde{y}$ may not be.

Posterior predictive distribution

If the model fits, replicated data y^{rep} generated under the model should look similar to observed data y.

In other words, the observed data y should look plausible under the **posterior predictive distribution**

$$p(y^{\text{rep}} \mid y) = \int p(y^{\text{rep}} \mid \theta) p(\theta \mid y) d\theta.$$

Test quantities and tail-area probabilities

y^{rep} is typically high-dimensional. Other than visual examination, we need to rely on test quantities and tail-area probabilities of the test quantities.
In frequentist/likelihood inference, a *test statistic* $T(y)$ is a function of the data. The tail area

$$p = Pr(T(y^{\text{rep}}) \geq T(y) \mid \theta)$$

is the p-value.

Posterior predictive p-value

We generalize test statistics to allow dependence on the model parameters under their posterior distribution:

$$T(y, \theta).$$

The (Bayesian) **posterior predictive p-value** is the probability that the replicated test quantity could be more extreme than the observed test quantity:

$$\begin{aligned} p_B &= Pr\left(T(y^{\text{rep}}, \theta) \geq T(y, \theta) \mid y\right) \\ &= \int \int \mathbf{1}\left\{T(y^{\text{rep}}, \theta) \geq T(y, \theta)\right\} p\left(y^{\text{rep}} \mid \theta\right) p\left(\theta \mid y\right) dy^{\text{rep}} d\theta, \end{aligned}$$

with the understanding that $p\left(y^{\text{rep}} \mid \theta, y\right) = p\left(y^{\text{rep}} \mid \theta\right)$.

Posterior predictive p-value: simulation strategy

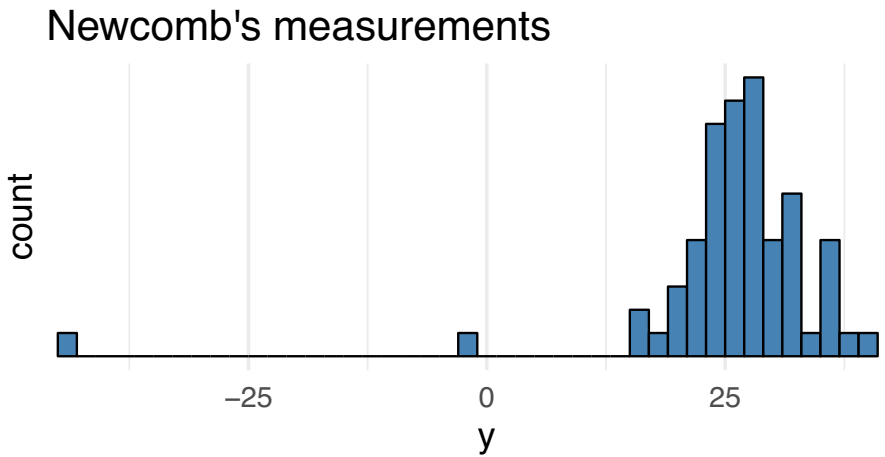
1. Draw $\theta^{(s)}$ from the posterior $p(\theta \mid y)$
2. Draw $y^{\text{rep},(s)}$ from $p(y^{\text{rep}} \mid \theta^{(s)})$
3. Repeat to get S replicates of y^{rep}
4. Estimated p_B as the proportion of these simulations for which the test quantity equals or exceeds its realized value:

$$\frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left\{ T(y^{\text{rep},(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}) \right\}$$

Simon Newcomb's light speed experiment in 1882

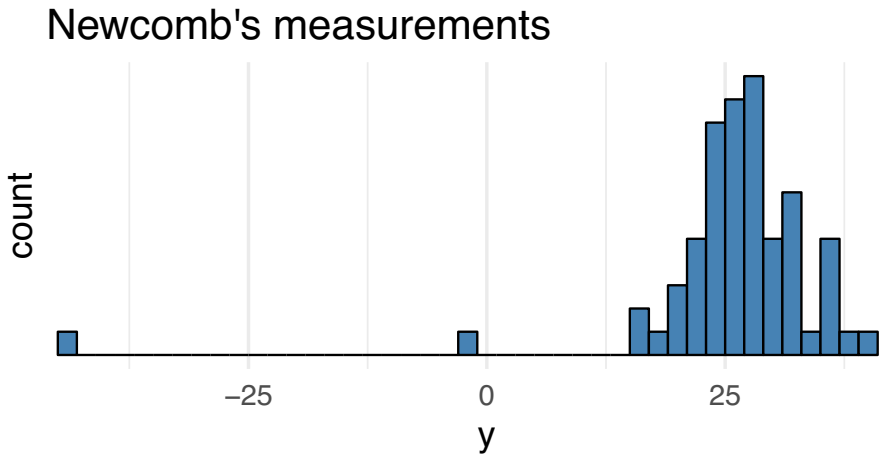
Newcomb measured ($n = 66$) the time required for light to travel from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of 7422m. Apply (inappropriately) the location-scale Normal model with a noninformative prior:

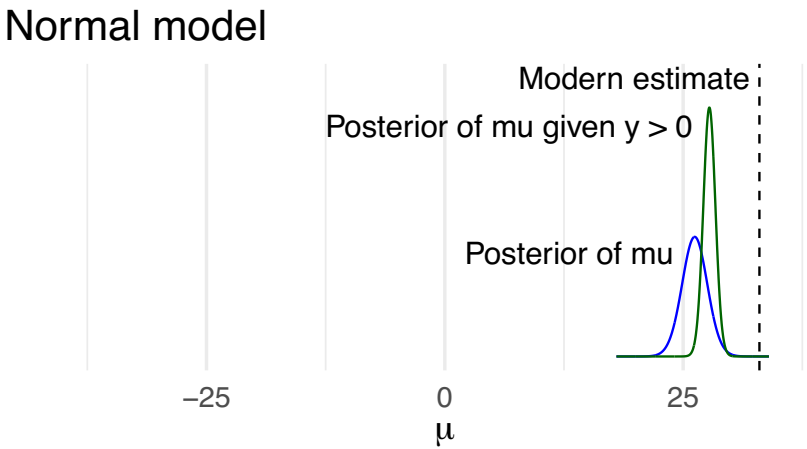
$$p(y \mid \mu, \sigma^2) \sim N(y \mid \mu, \sigma^2)$$
$$p(\mu, \sigma^2) \propto \sigma^{-2}$$



Simon Newcomb's light speed experiment in 1882

See R.





Choice of test quantity

In hypothesis testing, a good test statistic is ancillary (or almost ancillary).

- **Ancillarity**: if a statistic depends only on observed data and if its distribution is independent of the parameters of the model

On the other hand, bad test statistic is highly dependent of the parameters.

Choice of test quantity

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

- Test quantities should measure a feature of the data *not* directly addressed by the probability model.

A probability model can fail to reflect the process that generated the data in any number of ways.

- Choose a variety of test quantities in order to evaluate **more than one** possible model failure.

Choice of test quantity

That is,

If you don't look, you won't find the defect.
If you do, you may or may not find the defect!

Example. Variance for the normal model (Newcomb data)

Interpreting posterior predictive p-values

The posterior predictive p-value, p_B, has a **direct interpretation** as a posterior probability:

> If we believe the model, we think there is a $100p_B\%$ chance that tomorrow's test value will exceed today's test value.

Note: p_B is NOT $Pr(\text{model is true} \mid y)$!

Interpreting posterior predictive p-values

If the model is true:

- If the parameter θ is known (or estimated to a very high precision), or if the test statistic $T(y)$ is ancillary and continuous, then

$$Pr\left(T(y^{\text{rep}}) \geq T(y) \mid y\right)$$

  has a uniform distribution.

- On the other hand, if there is posterior uncertainty about θ, which propagates to the distribution of $T(y \mid \theta)$, the distribution of the p-value is more concentrated near the middle range, i.e. *stochastically less variable* than the uniform. **(Why?)**

Posterior calibration

This is the basis of criticism that posterior predictive checks are *conservative* or *uncalibrated*.

Calibration. A model is calibrated (with respect to T), if we observe many replications $T(y^{\text{rep}})$ and the event $T(y^{\text{rep}}) > T(y)$ indeed happens $100p_B\%$ times, where p_B is the posterior predictive probability for the event.

Marginal predictive checking

Consider each replicated observation separately in its marginal predictive distribution $p(\tilde{y}_i \mid y)$:

- marginal posterior p-values

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

- If $T(y_i) = y_i$,

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

Cross-validation predictive checking

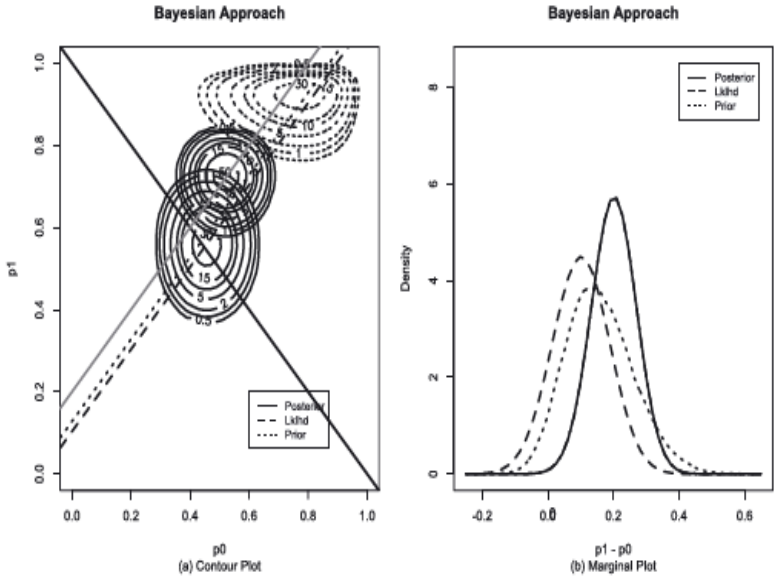
A related approach is to use *cross-validation* predictive distributions, giving CV predictive p-values

$$p_i = Pr(T(y_i^{\text{rep}}) \leq y_i \mid y_{-i}).$$

For continuous data, CV predictive p-values have uniform distribution if the model is calibrated. On the downside, they generally requires additional computation.

Example. Newcomb data (again)

*Discrepant posterior phenomenon



- Xie, M., Liu, R. Y., Damaraju, C. V., & Olson, W. H. (2013). Incorporating external information in analyses of clinical trials with binary outcomes. *The Annals of Applied Statistics*, 342-368.
- Chen, Y., Gong, R., & Xie, M. G. (2020). Geometric Conditions for the Discrepant Posterior Phenomenon and Connections to Simpson's Paradox. *arXiv preprint arXiv:2001.08336*.

*Discrepant posterior phenomenon