Model building and model criticism

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Outline

Likelihood construction

- The specification of the likelihood is often the most important part of model building.
- Sometimes default choices of likelihood are used for convenience.
- While defaults can be useful, it is preferable to use a model that reflects the "physics" of the data generating process.
- Various aspects to consider:
  - Special properties of distributions
  - Parametrization (e.g., multiplicative versus additive effects)
  - Conditional independence properties (e.g., Markov, etc.)

Likelihood construction: Special properties

- Many distributions have special properties that make them well-justified in particular applications.

- Examples
  - Gaussian: Central limit theorem
  - Poisson: Law of small numbers
  - Exponential: Memorylessness
  - Pareto: Power law
  - Exponential families: Maximum entropy
  - Poisson process: Limit of Bernoulli processes

- Combining or transforming distributions according to the physics of the data generating process is also useful.

Likelihood construction: Poisson process example

- Consider dividing $\mathbb{R}^d$ into tiny boxes and putting an independent $\text{Bernoulli}(p)$ random variable in each box, where $p = \int_{\text{box}} f(x)dx$ for some nonnegative function f.

- This converges to a Poisson process as the resolution of the grid goes to infinity.

- The "limit of Bernoulli processes" perspective gives intuition into when a Poisson process model might be reasonable.

- Examples
  - times of neuron spikes
  - locations of mutations in a genome
  - times of speciation events in phylogenetic history
  - emission times of radioactively decaying particles
  - locations of organisms in a habitat at a given time

Outline

Objective Bayesian inference

- If there is universally-accepted prior information, then almost no one would argue with using it.
- But what if you really have no idea at all?
- Or, more likely, what if it is critical that your results not depend on any personal biases? e.g.,
  - clinical trials for a new drug,
  - testing of a medical device,
  - evidence to be presented in a court of law.
- The original motivation of *objective Bayes* was to find priors that contain little or no information.
- That has evolved into a more attainable goal of finding "default" priors that provide reliable and interpretable results, to be used as conventions when more specific prior information can't or shouldn't be used.

Non-informative priors

- Such priors are called *non-informative*, and they are often improper, in the sense that they do not integrate to a finite value.

- If we are using a gamma prior $\theta \sim \mathrm{Gamma}(a, b)$, then since the prior variance $\mathrm{Var}(\theta) = a/b^2$ is finite, in some sense the prior contains some information.

- This is apparent in the shrinkage that occurs, with the posterior mean being a convex combination of the sample mean and prior mean.

- If we take $a = b = \varepsilon$ for ε small, then the prior mean stays at $a/b = 1$ and the variance becomes large. As $\varepsilon \to 0$, the shape of the prior becomes $\propto 1/\theta$.

- In this limit, the gamma prior becomes non-informative.

Jeffreys priors

- The Jeffreys prior is defined (in the univariate case) as

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

  where $I(\theta) = \int \left(\frac{\partial}{\partial\theta} \log p(y|\theta)\right)^2 p(y|\theta) dy$ is the Fisher information for θ.

- The Jeffreys prior is a classical non-informative prior.

- Examples
  - Gaussian mean: $\pi(\mu) \propto 1$.
  - Gaussian standard deviation: $\pi(\sigma) \propto 1/\sigma$.
  - Poisson rate: $\pi(\lambda) \propto 1/\sqrt{\lambda}$.
  - Bernoulli success probability: $\pi(p) \propto 1/\sqrt{p(1-p)}$.

Jeffreys priors: Invariance property

- Suppose $\pi(\theta)$ is the Jeffreys prior for θ in some model $p(y|\theta)$.
- Suppose we have an alternate parametrization of the model, say $q(y|\phi) = p(y|\theta)$ where $\theta = h(\phi)$ and h is a smooth 1-to-1 function.
- By change of variables, the induced prior on ϕ is

$$p(\phi) \propto \pi(h(\phi))|h'(\phi)|.$$

- If $\tilde{\pi}(\phi)$ is the Jeffreys prior for ϕ then it turns out that

$$\tilde{\pi}(\phi) \propto \pi(h(\phi))|h'(\phi)|.$$

- Thus, the Jeffreys prior for ϕ coincides with the prior on ϕ induced by the Jeffreys prior for θ.

Comments on improper priors

- Bayesian inferences under improper priors are sometimes similar to frequentist inferences.
- If using an improper prior, it is important to make sure the resulting posterior is proper.
- Improper posteriors can arise, for example, when there is non-identifiability and the prior is improper.
- If the posterior is improper, inferences are typically meaningless — the posterior mean, credible intervals, etc., are undefined.
- Even if the posterior is proper, serious issues sometimes arise when using improper priors: contradictory probabilities, prior can dominate for large n, inadmissible estimators, marginalization paradoxes.

Weakly informative priors

- Often, weakly informative priors are preferable to non-informative priors.
- A *weakly informative prior* is proper, but is more diffuse than a typical subjective prior.
- In many situations, a weakly informative prior will outperform a non-informative one.
- In small sample sizes and data sparse situations, weakly informative priors stabilize inferences through mild shrinkage towards the prior mean.
- This is the bias-variance tradeoff — the prior introduces a bit of bias to greatly reduce variance.

Outline

Model criticism

- Once you have built and implemented a model, it is important to see how well it agrees with the data.
- In addition to checking how well the model fits the data distribution, it is also important to check whether the parameter inferences are appropriate.
- This is referred to as model criticism or model checking.
- Unfortunately, there does not seem to be a fully Bayesian way to do model criticism, since Bayes assumes the model is correct.
- Nonetheless, there are a number of commonly used techniques for model criticism in the Bayesian setting.

Model criticism

- Lack of fit can be due to the prior and/or the likelihood. The likelihood usually has a much bigger impact, especially asymptotically.
- The term "model" sometimes refers to just the likelihood, but here we say "model" to mean both the likelihood and the prior, together.
- *"All models are wrong, but some are useful."* – George Box
- In practice, we can never hope to model the data generating process perfectly, but we can often find a model that is adequate for a given application.

Goals of model criticism

- A good model criticism procedure not only tells you if the model is wrong, but also in what ways the model is wrong, and ideally, whether the models' inadequacies significantly affect the results.

- Comparison with sensitivity analysis
  - Sensitivity analysis = Seeing whether other plausible models give similar results.
  - This is important, but it is different than model criticism. All of the other models you consider might give similar results, and all of them might be very wrong.
  - Conversely, even if the model seems to fit well, it is still a good idea to do sensitivity analysis.

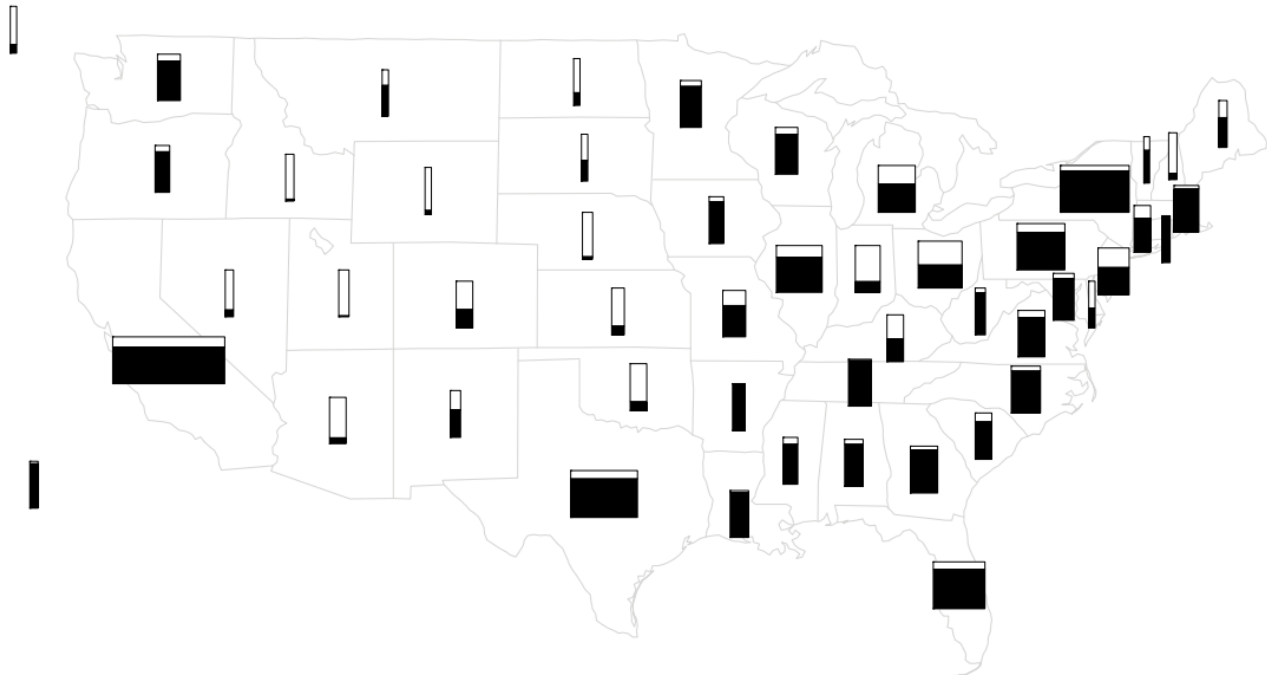
Approaches to model criticism

- Two general approaches to model criticism

    1. Checking whether the results are consistent with other information not used when constructing the model or computing the posterior. (For example, other domain knowledge, other data, etc.) This is the approach taken by cross-validation.

    2. Checking whether the results are consistent with the information used to construct the model and compute the posterior. This is the approach taken by "posterior predictive checks". This has some issues due to the fact that it is "using the data twice", but it serves as a useful check of internal consistency.

Example: 1992 election forecast



(Figure 6.1 from BDA3, Gelman et al., 2013)

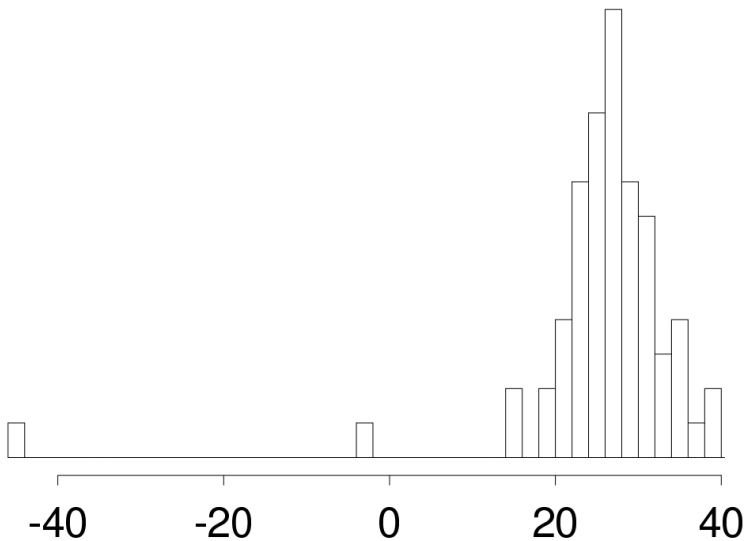
Forecast of 1992 US presidential election. For each state, the height of the black bar indicates the probability of Clinton winning the election, and the width is the number of electoral votes. Polls from Texas and Florida not used in the forecast indicated much less support for Clinton than predicted, indicating a possible inadequacy of the model.

Outline

Posterior predictive checks

- The basic idea of posterior predictive model checks is that the observed data set should look "typical", relative to replicate data sets sampled from the posterior predictive distribution.

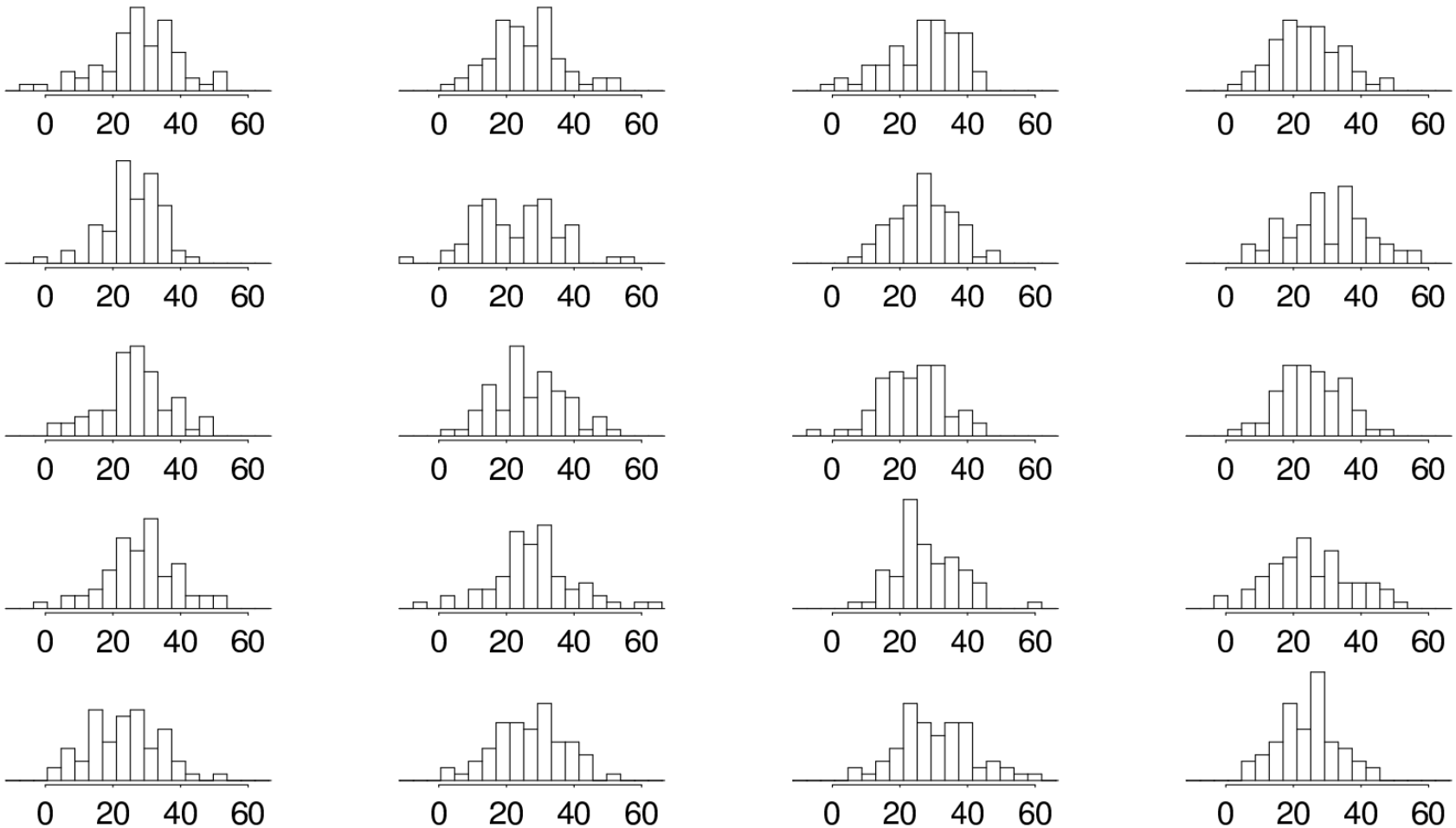
- Below are 66 measurements of the speed of light made by Simon Newcomb in 1882. This is a classic example of a data set with outliers. Note that the lowest two measurements are significantly lower than the rest. (Incidentally, Newcomb threw out the lowest one, but kept the second lowest one.)



Example: Newcomb's speed of light measurements

- A naive approach would be to model the data as $\mathcal{N}(\mu, \sigma^2)$.
- In this case, we can visually see that there may be issues with this model, but in more complicated, high-dimensional situations, it is difficult to visually assess whether a model will fit. Let's see what happens if we use the normal model.
- Below are 20 replicate data sets sampled from the posterior predictive (see BDA for prior details).
- Each replicate dataset was generated by first sampling (μ, σ^2) from the posterior, and then sampling 66 points x_i^{rep} from a normal with this mean and variance (i.e., using the same mean and variance for all 66 points).

Example: Newcomb's speed of light measurements

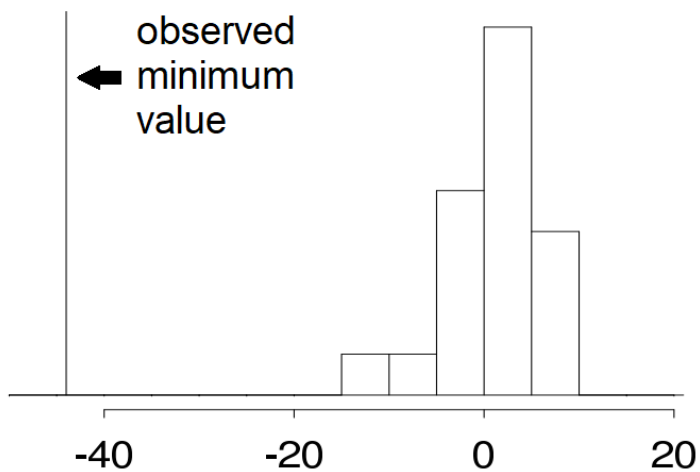


Example: Newcomb's speed of light measurements

- Visually, none of these posterior predictive datasets look much like the original observed data set.

- To quantify this, below is a histogram of the 20 minimum values, $\min\{x_1^{\text{rep}}, \ldots, x_{66}^{\text{rep}}\}$ from the 20 replicate data sets, compared to the observed minimum, $\min\{x_1, \ldots, x_{66}\}$.

- The minimum is just one possible choice of statistic that could be used to check the model fit.



The idea of posterior predictive checks

- If by some miracle the likelihood is exactly correct, and we have a very large amount of data, then the posterior should be highly concentrated at the true parameter value.

- In this case, the distribution of the observed data should be basically identical to the distribution of the replicate data sets. However, this is more than we can hope for.

- On the other hand, if the likelihood or prior are very wrong, then basically none of the replicate data sets will look like the observed data set, indicating mismatch between the model and the data.

The idea of posterior predictive checks

- If the likelihood and prior are close enough to being correct, relative to the amount of data, then some of the replicate data sets should look roughly like the observed data set. This is what we aim for.

- A nice thing about posterior predictive checks is that they are computationally cheap, since they only involve generation of replicate data sets, as opposed to computing the posterior for multiple data sets.

Definition of posterior predictive for replicate data sets

- Recall that the posterior predictive distribution of a single future sample is

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\theta)p(\theta|x_{1:n})d\theta$$

(assuming the data is conditionally independent given θ).

- Here, we will be considering the posterior predictive distribution of an entire replicate data set, $x_{1:n}^{\text{rep}}$,

$$p(x_{1:n}^{\text{rep}}|x_{1:n}) = \int p(x_{1:n}^{\text{rep}}|\theta)p(\theta|x_{1:n})d\theta.$$

- We can sample each of these posterior predictive datasets by first sampling $\theta|x_{1:n}$, and then sampling $x_{1:n}^{\text{rep}}|\theta$.

Test statistics and test quantities

- To perform a posterior predictive check, we compare the observed value of some test statistic/quantity to its distribution under the posterior predictive.
- Test statistic = function of the data only, $T(x_{1:n})$.
- Test quantity = function of data and parameter, $T(x_{1:n}, \theta)$. This is a generalization that is sometimes useful.
- The speed of light example used the test statistic $T(x_{1:n}) = \min\{x_1, \ldots, x_n\}$.
- Using different test statistics/quantities will allow you to probe different aspects of model fit.

Posterior predictive p-values

- To simplify the notation, let's abbreviate $x = x_{1:n}$.

- Recall that the classical frequentist p-value for a test statistic $T(x)$ is defined as

$$p_C = \mathbb{P}\big(T(X) \geq T(x) \mid \theta_0\big),$$

  where x is the observed data, X is distributed according to P_{θ_0}, and θ_0 corresponds to the null hypothesis.

- The idea is that if the p-value is very small, then the observed data is very atypical under the null hypothesis, providing evidence for rejecting the null.

Posterior predictive p-values

- The posterior predictive p-value for $T(x)$ is defined as

$$p_B = \mathbb{P}\big(T(X^{\text{rep}}) \geq T(x) \mid x\big),$$

  or more generally, for a test quantity $T(x, \theta)$,

$$p_B = \mathbb{P}\big(T(X^{\text{rep}}, \theta) \geq T(x, \theta) \mid x\big).$$

- In this expression, x is fixed and equal to the observed data, and $(X^{\text{rep}}, \theta)|x$ is distributed according to

$$p(x^{\text{rep}}, \theta|x) = p(x^{\text{rep}}|\theta)p(\theta|x),$$

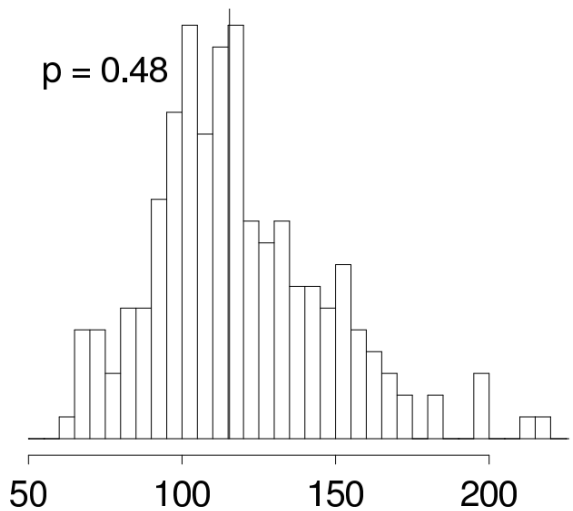
  which we sample from as described earlier.

- We use these samples in a Monte Carlo approximation,

$$p_B \approx \frac{1}{S}\sum_{s=1}^{S} \mathrm{I}\big(T(x^{\text{rep},s}, \theta^s) \geq T(x, \theta^s)\big).$$

Example: Speed of light (continued)

- How well is the model capturing the variance? What happens if we choose $T(x) = \hat{\sigma}^2$?
- The p-value is around 0.48, which seems to indicate that the model is doing okay. However, the sample variance is a poor choice of test statistic for this model.
- It is a sufficient statistic, so the model is already paying very close attention to matching it. Consequently, it isn't really helping us recognize poor model fit!



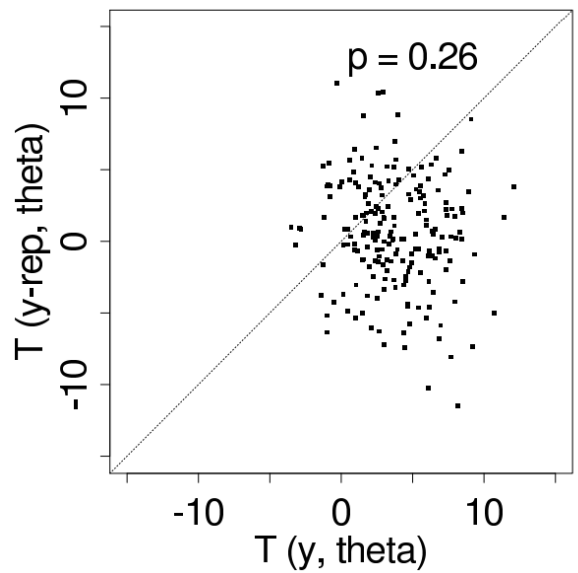
(BDA figure 6.4) Histogram of the sample variance over the 200 replicate data sets, compared with the observed sample variance.

Example: Speed of light (continued)

- How well is the model capturing any asymmetry in the distribution? We can compare distance from the 90% and 10% quantiles to the center, θ:

$$T(x, \theta) = |x_{(61)} - \theta| - |x_{(6)} - \theta|.$$

- The p-value is around 0.26, so the model actually seems to be doing okay in this respect.



(BDA figure 6.4) Scatterplot of $T(x, \theta)$ versus $T(x^{\text{rep}}, \theta)$ for the asymmetry statistic; the p-value is the fraction above the diagonal.

Interpretation of posterior predictive checks

- The purpose of posterior predictive checks is not to formally test goodness of fit, but rather, to explore and visualize how well the model is capturing various aspects of the data distribution.

- As usual, if you are computing multiple p-values, it is important to be aware of the fact that some of them may be small or large simply by chance—i.e., remember the multiple testing issue.

- We are not constructing a formal hypothesis test, but it is important to keep this in mind when interpreting the results.

Interpretation of posterior predictive checks

- Ideally, a frequentist p-value is uniformly distributed over the interval from 0 to 1.
- Posterior predictive p-values do not always have this property, which can complicate their interpretation somewhat.
- This is another reason to view them as tools for exploratory purposes, rather than precisely calibrated tests of model fit.

Interpretation of posterior predictive checks

- Also remember that each posterior predictive check is only assessing model fit with respect to one particular statistic.
- What should you do if your posterior predictive check indicates a problem?
- In some cases, the inferences you draw from the model might not be negatively affected by the lack of model fit.
- Otherwise, the posterior predictive check can provide valuable insight into how to modify the model in order to improve it.
- If you do modify the model, you need to be careful not to overfit—after all, you could match every statistic perfectly by using a model which simply reproduces the observed data set every time!

References and supplements

- Kass & Wasserman. *The selection of prior distributions by formal rules*. JASA, Vol. 91, No. 435, 1996.
- James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985. (Sec 3.3)
- Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin. *Bayesian Data Analysis*. CRC press, 2013.