# Lecture 6

# Model Checking

Textbook Ch6

邓婉璐

wanludeng@tsinghua.edu.cn

清华大学统计学研究中心

# Outline



- Introduction – What & Why
- External Validation
- Internal Validation
  - Posterior Predictive Checking
  - Graphical Posterior Predictive Checking

# Objectives for Today



- 理解:
  - ✓ 为何需用模型检验(目的)
  - ✓ 选择$T(y,\theta)$的原则/导向
  - ✓ 与频率学派的对比
  - ✓ 如何理解Internal validation中的p-value

- 掌握下述方法的基本步骤，及简单情形中的应用:
  - ✓ Internal Validation中的Test quantity $T(y,\theta)$和p-value法
  - ✓ 基本的图像法(Graphical Posterior Predictive Checks 中基于$T(y,\theta)$的散点图、直方图，重复数据的直接呈现)

# What & Why?

# Model Checking: What & Why?



"*All models are wrong but some models are useful*"
– George E. P. Box

So far we have looked at a number of models and examined them with example data sets. Do the models used accurately describe the data used?

In standard analyses, we will often check model assumptions. For example, in standard regression we will check for

- Correct form of the regression function (e.g. linear vs quadratic)
- Constant variance of the residuals
- Independence of of the residuals
- Normality of the residuals

# Model Checking: What & Why?



- *Basic question: How sensitive are our posterior inferences to our modelling assumptions?*
- Rat Example: Will the following models give significantly different answers about tumor rates in each group?
  1. Original model
     - Data model: $y_i$ = number of tumors in group $i$

$$y_i|\theta_i \sim^{ind} Bin(n_i, \theta_i), \qquad i = 1, \ldots, 71$$

     - Parameter model: $\theta_i$ = tumor rate in group $i$

$$\theta_i|\alpha, \beta \sim^{i.i.d.} Beta(\alpha, \beta)$$

     - Hyper-parameter model:

$$p(\alpha, \beta) \propto \frac{1}{(\alpha + \beta)^{5/2}}$$

# Model Checking: What & Why?



- *Basic question: How sensitive are our posterior inferences to our modelling assumptions?*
- Rat Example: Will the following models give significantly different answers about tumor rates in each group?
  2. Alternative model 1
     - Data model: $y_i$ = number of tumors in group $i$

$$y_i|\theta_i \sim^{ind} Bin(n_i, \theta_i), \qquad i = 1, \ldots, 71$$

     - Parameter model: $\theta_i$ = tumor rate in group $i$

$$logit(\theta_i)|\mu, \sigma^2 \sim^{i.i.d.} N(\mu, \sigma^2)$$

     - Hyper-parameter model:

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

# Model Checking: What & Why?



- *Basic question: How sensitive are our posterior inferences to our modelling assumptions?*
- Rat Example: Will the following models give significantly different answers about tumor rates in each group?
  3. Alternative model 2
     - Data model: $y_i$ = number of tumors in group $i$

$$y_i|\alpha_i,\beta_i \sim^{ind} Beta-bin(n_i,\alpha_i,\beta_i), \qquad i=1,\dots,71$$

     - Parameter model: $(\alpha_i,\beta_i)$ = tumor rate in group $i$

$$\alpha_i,\beta_i|\gamma_\alpha,\delta_\alpha,\gamma_\beta,\delta_\beta \sim^{i.i.d.} \Gamma(\alpha_i|\gamma_\alpha,\delta_\alpha)\Gamma(\beta_i|\gamma_\beta,\delta_\beta)$$

       The tumor rate for group $i$ is $E\left(\frac{y_i}{n_i}\middle|\alpha_i,\beta_i\right)=\frac{\alpha_i}{\alpha_i+\beta_i}$.

     - Hyper-parameter model:

$$p(\gamma_\alpha,\delta_\alpha,\gamma_\beta,\delta_\beta) \propto 1$$

# Model Checking: What & Why?



- Note that we are not trying to answer the question of whether our model is correct or not. It's not (see Box). We are interested in whether the inaccuracies matter.
- Examples you may have seen in the past where deviations from assumptions don't hurt much (at least in large samples):

  Linear Regression: $Y = X\beta + \varepsilon$

  – $\hat{\beta} = (X'X)^{-1}X'Y$ is unbiased if $E[\varepsilon] = 0$
  – $\hat{\beta}$ is minimum variance unbiased estimator if $E[\varepsilon] = 0$ and constant variance. (Gauss-Markov theorem)
  Neither of these results require normality of ε.
- Examples where assumptions can matter:

  F-test for examining $H_0: \sigma_1^2 = \sigma_2^2$ $vs$ $H_a: \sigma_1^2 \neq \sigma_2^2$. The results of this test can be highly dependent on the *iid* normal assumptions for each group.

# Model Checking: What & Why?



- Steps of a Bayesian analysis:
  - constructing a probability model
  - computing the posterior distribution of all estimands
  - assessing the fit of the model to the data and to our substantive knowledge
  - model improvement

- **Remarks:**
  - Checking the model is crucial to statistical analysis.
  - Bayesian prior-to-posterior inferences can yield misleading inferences when the model is poor.

# Sensitivity Analysis from a ‘Super-Model’ Perspective

- It is possible that more than one reasonable model can provide an adequate fit to the data in a scientific problem
- These models may differ substantially in the prior specification, the sampling distribution, or in what information is included
- Key point of a sensitivity analysis: how much do posterior inferences change when different reasonable probability models are used
- Ideally, model checking is done by setting up a comprehensive joint distribution over all possible ‘true’ models (i.e., a super-model).
- In practice, however, setting up such a super-model to include all possibilities and all substantive knowledge is both conceptually impossible and computationally infeasible in all but the simplest problems.

# Model Checking



- Instead we will base these checks on the posterior predictive distribution. Does our data look like our fitted model says it should?

- This can either be done by
  - External validation: future data is compared with the posterior predictive distribution.
  - Internal validation: observed data is compared with the posterior predictive distribution.

# Approach I: External Validation

# External Validation



- Basic logic
  - using the model to make predictions about future data
  - collecting those data and comparing to their predictions
- Limitations
  - often we need to check the model before having new data
- Practical concerns
  - a single model can be used to make different predictions (e.g. existing group or new group)
  - selecting the focus of predictions could be tricky

# Approach II: Posterior Predictive Checking

## INTERNAL VALIDATION

# Posterior Predictive Checking



- Basic logic:
  - If the model fits, replicated data generated under the model should look similar to observed data
  - If we see some discrepancy, it is due to model misspecification or due to chance.
  - It is really a self-consistency check
- Procedure:
  - draw simulated data from the posterior predictive distribution
  - compare these samples to the observed data
  - Any systematic differences between the simulations and the data indicate potential failings of the model

# Example: Checking Independence in Binomial Trials



- Model: a binary sequence, $y_1, \ldots, y_n$, modeled as independent trials with a common probability of success $\theta$
- Prior of $\theta$: a uniform distribution (i.e. $\theta \sim Beta(1,1)$ )
- Posterior: $\theta \sim Beta(\sum y_i + 1, n - \sum y_i + 1)$
- Observed data in order: 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0

Simulation procedure to get posterior prediction?

# Step 1: Generate Replicated data

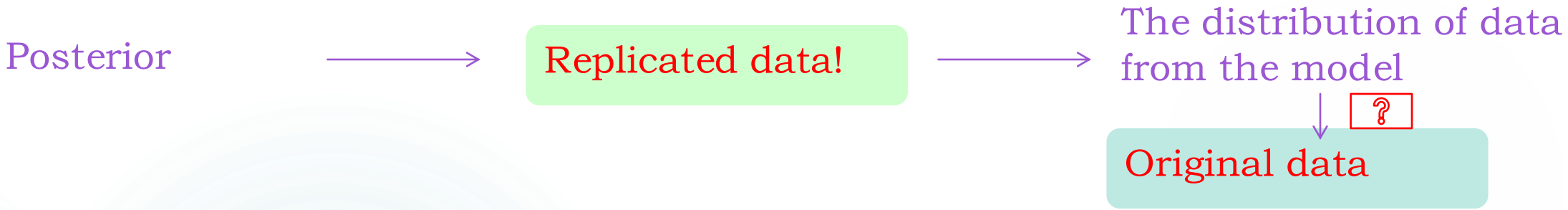


Posterior ⟶ Replicated data!

- Generate $L$ datasets, $y_1^{rep}, \ldots, y_L^{rep}$ from the posterior predictive distribution $p(y^{rep}|y)$. $y^{rep}$ corresponds to replicated data. Notice if there are any covariates that are conditioned on in the original data.
- For example, in the rat tumor example, we need to use the same group sample sizes as in the original data set.
- $\tilde{y}$ represents any future outcome whereas $y^{rep}$ indicates a replication exactly like the observed $y$. $\tilde{y}$ does not need to have the same covariate structure as the original data.

# Step 2: Compare Obs. vs Rep.






- The approach has a similar feel to hypothesis testing, where a test statistic $T(y, \theta)$ needs to be defined to measure the discrepancy between the data and the predictive simulations.
- The lack of fit of the data as compared to the posterior predictive distribution can be compared by a tail-area probability (e.g. $p$-value) of the test statistic $T(y, \theta)$. To calculate this probability we will use the replicates sampled from $p(y^{rep}|y)$.

# Detailed Calculation - Formula



- Posterior predictive distribution

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

- Classical p-values (frequentist test)

$$p_C = \Pr(T(y^{\text{rep}}) \geq T(y)|\theta)$$

Which is random?

- Posterior predictive p-values (Bayesian test)

$$p_B = \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y)$$

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

Similar to choosing a powerful test statistic when conducting a hypothesis test

Test quantity to be specified

What difference do you find here?

# Detailed Calculation - Simulation



❖ Calculation of Posterior predictive p-values (Bayesian test)

$$p_B = \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y) = \iint I_{T(y^{\text{rep}},\theta)\geq T(y,\theta)} p(y^{\text{rep}}|\theta) p(\theta|y) dy^{\text{rep}}\, d\theta$$

- Usually we can't calculate the Bayesian p-value exactly, but can do it by simulation.
- Suppose that we have $L$ simulations of $\theta$ $(\theta^1, \ldots, \theta^L)$ from the posterior distribution $p(\theta|y)$. Then for each of these samples, say $\theta^l$, generate one sample $y^{rep,l}$ from $p(y^{rep}|\theta^l)$.
- We want to compare each of the $T(y^{rep,l}, \theta^l)$ with $T(y, \theta^l)$.
- Then $\hat{p}_B = \frac{1}{L}\sum_{l=1}^{L} I(T(y^{rep,l}, \theta^l) \geq T(y, \theta^l))$ is an estimate of $p_B$.

(i.e. the proportion of samples where $T(y^{rep,l}, \theta^l) \geq T(y, \theta^l)$ )

# Example: Checking Independence in Binomial Trials



- Model: a binary sequence, $y_1, \ldots, y_n$, modeled as independent trials with a common probability of success $\theta$
- Prior of $\theta$: a uniform distribution (i.e. $\theta \sim Beta(1,1)$ )
- Posterior: $\theta \sim Beta(\sum y_i + 1, n - \sum y_i + 1)$
- Observed data in order: 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
- The observed autocorrelation suggests the model is flawed

# Example: Checking Independence in Binomial Trials

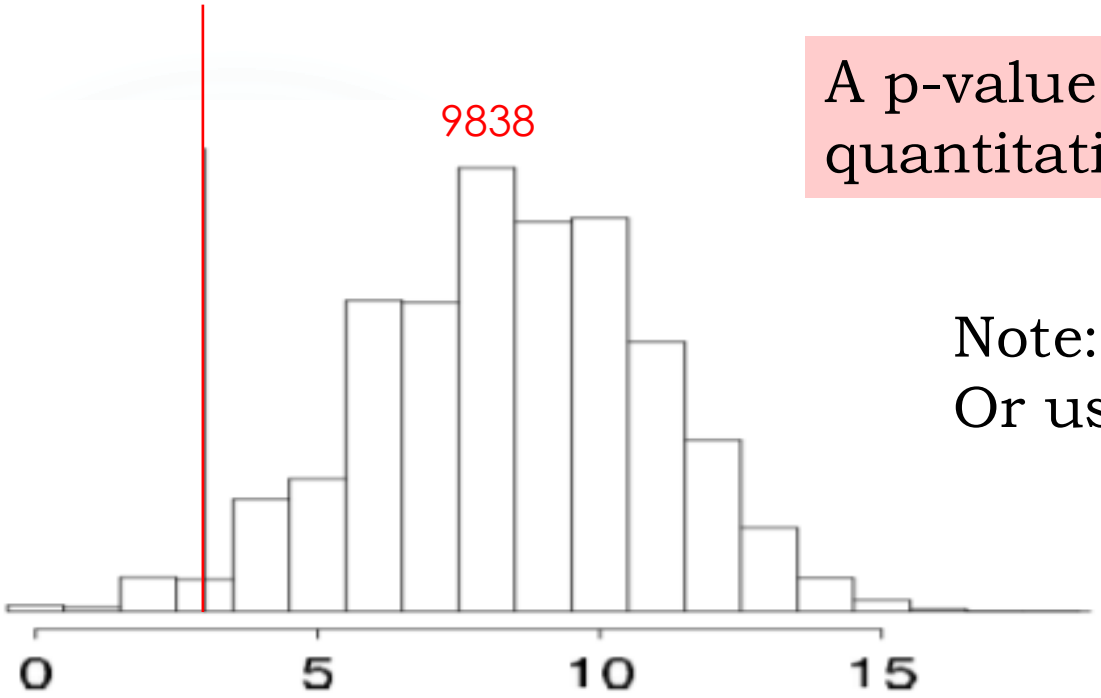


- Observed data in order: 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
- Summary statistics for observed data: $n = 20, \Sigma y_i = 7$.
- Posterior: $Beta(7 + 1, 20 - 7 + 1) = Beta(8, 14)$

- Test quantity $T$ = number of switches between 0 and 1 in the sequence
- $T(y) = 3$
- Get posterior predictive distribution of $T(y^{rep})$ by simulation
  - Draw $\theta^s$ from its $Beta(8, 14)$ , then draw $y^{rep's} = (y_1^{rep's}, \ldots, y_{20}^{rep's})$ as independent Bernoulli variables with probability $\theta^s$
  - Repeat the procedure above for $s = 1, \ldots, 10000$.
- Compare $T(y)$ to the posterior predictive distribution of $T(y^{rep})$

# Example: Checking Independence in Binomial Trials






A p-value of 9838/10000=98.38% gives a quantitative evidence of model failure.

Note: here $P(T(y^{rep},\theta) \leq T(y,\theta)|y)$.
Or use $-T$ with $P(T(y^{rep},\theta) \geq T(y,\theta)|y)$.

Figure 6.5 Observed number of switches compared to 10,000 simulations from the posterior predictive distribution of the number of switches.

# Principles for Choosing Test Quantities



- Because a probability model can fail to reflect the process that generated the data in any number of ways, posterior predictive p-values can be computed for a variety of test quantities in order to evaluate more than one possible model failure.
- 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, ranks of the sample, or correlation of residuals with some possible explanatory variable

# Example: Newcomb's speed of light



In this example, a worry was the effect of outliers.

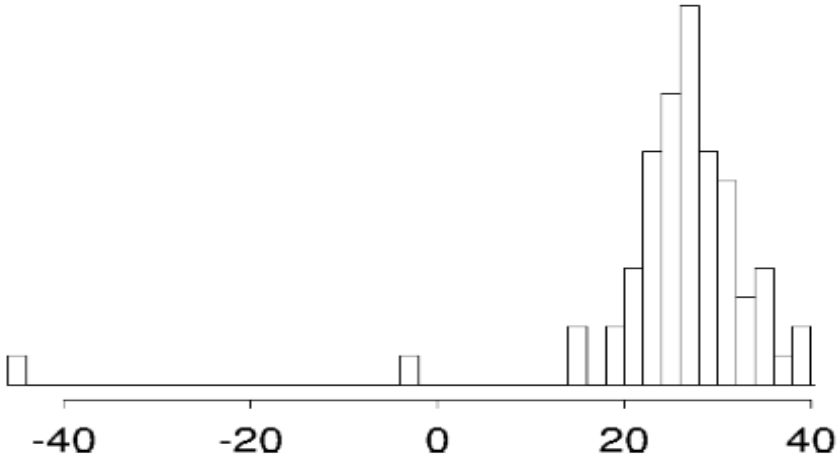



Figure 3.1 Histogram of Simon Newcomb's measurements for estimating the speed of light (n = 66)

Likelihood: $N(\mu, \sigma^2)$

Prior: $p(\mu, \log(\sigma)) \propto 1$

# Example: Newcomb's speed of light



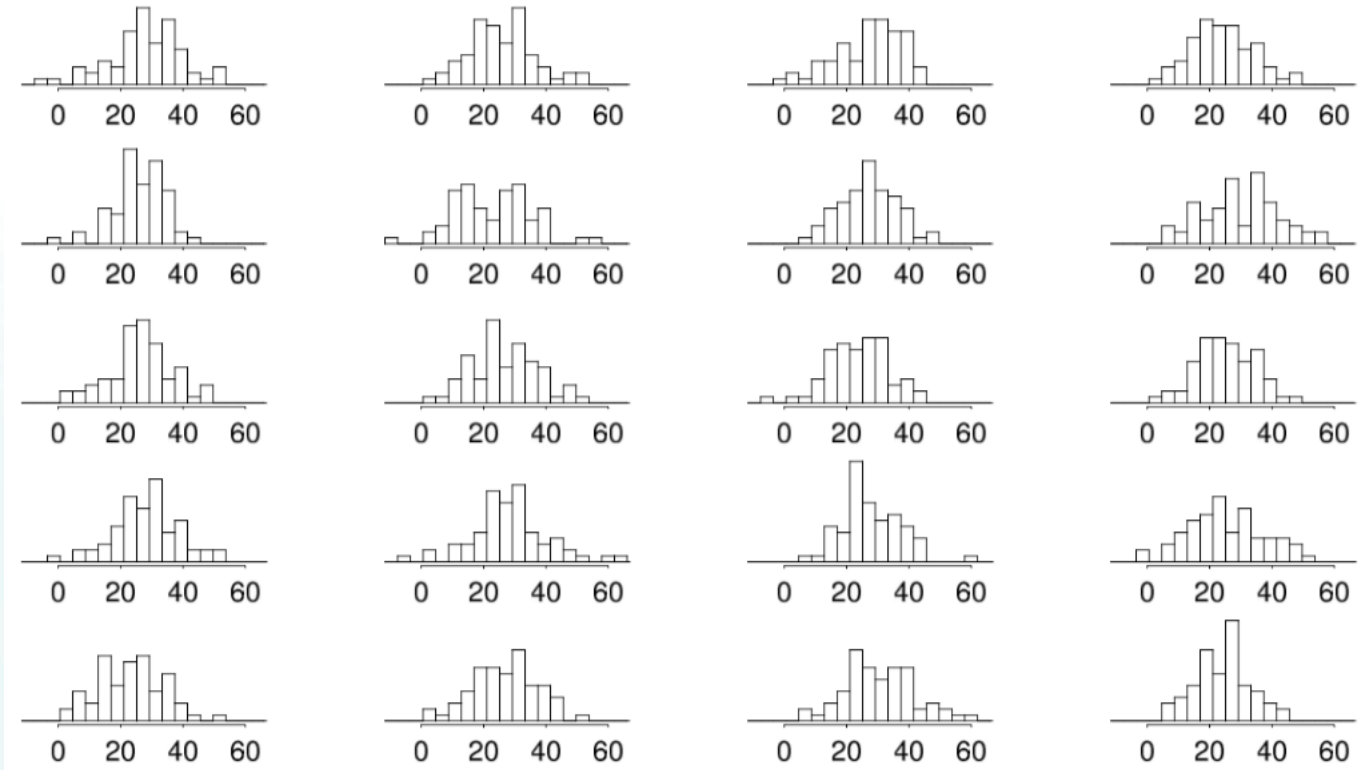



Figure 6.2 Twenty replications, $y^{rep}$, of the speed of light data from the posterior predictive distribution, $p(y^{rep}|y)$; compare to observed data, y, in Figure 3.1. Each histogram displays the result of drawing 66 independent values $\tilde{y}_i$ from a common normal distribution with mean and variance $(\mu, \sigma^2)$ drawn from the posterior distribution, $p(\mu, \sigma^2|y)$, under the normal model.

# Example: Newcomb's speed of light



In this example, a worry was the effect of outliers.
Thus $T(y, \theta)$ needs to be chosen to focus on this issue.

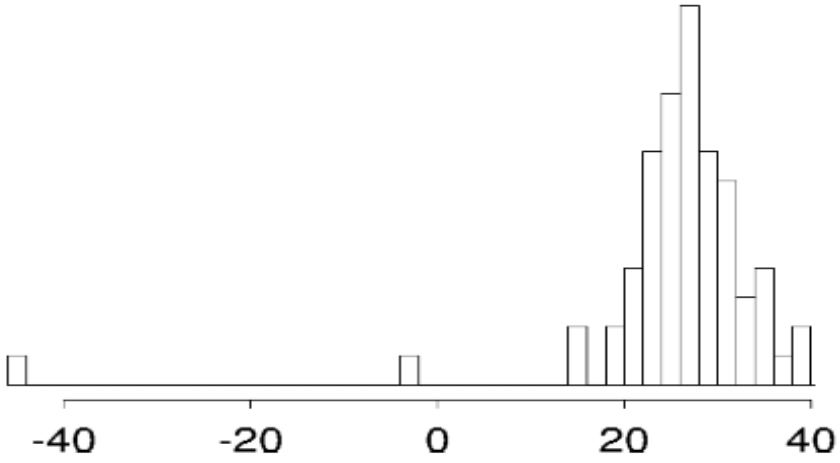



Figure 3.1 Histogram of Simon Newcomb's measurements for estimating the speed of light (n = 66)

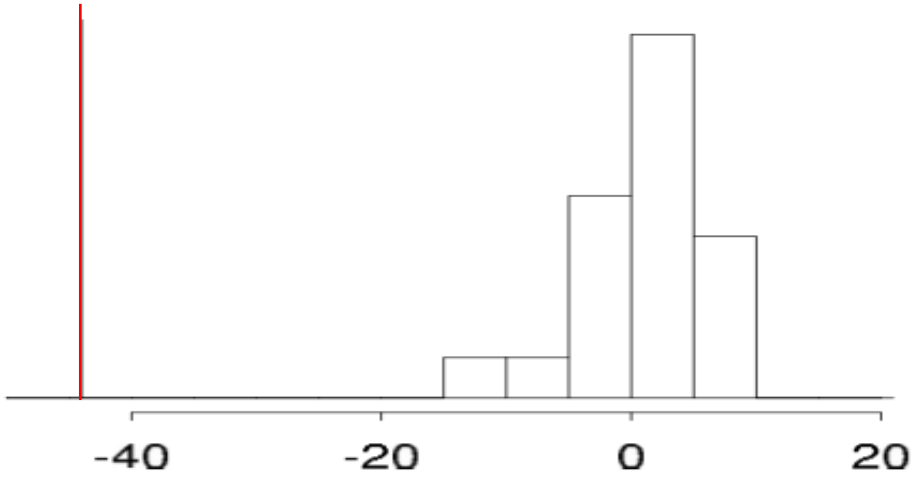



Figure 6.3 Smallest observation of Newcomb's speed of light data (the vertical line at the left of the graph), compared to the smallest observations from each of the 20 posterior predictive simulated datasets

# Example: Newcomb's speed of light



- Concern: The effect of outlier.

- Choice of $T(y, \theta)$:
  - $\min y_i$ : worry about low outliers
  - $\max |y_i - \mu|$: This would be appropriate if the worry was either big positive or big negative residuals.

- A revised model might be

$$y_i|\mu, \sigma^2 \sim t_\nu(\mu, \sigma^2)$$

instead of

$$y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$$

for the sampling distribution in the analysis.

# Advantages of Bayesian over Frequentist



- Difference: The test statistic $T(y, \theta)$ can depend on the data y and the parameters and hyperparameters $\theta$, which is different from standard hypothesis testing where the test statistic only depends on the data, but not the parameters.

- Reality: For many problems, a function of data and parameters can directly address a particular aspect of a model in a way that would be difficult or awkward using a function of data alone.

- Implementation: Test quantity $T(y, \theta)$ as well as its replication $T(y^{rep}, \theta)$ are unknowns and are represented by $S$ simulations.

# Interpreting Posterior Predictive p-values



- An extreme $p$-value for a test statistic $T(y,\theta)$ (e.g. near 0 or 1) indicates that the observed pattern in the data would be unlikely of the data if the model were true.
- While it is a probability, it is not $P[model\ is\ true \mid data]$. As we have seen before, it is $P[T(y^{rep},\theta) \geq T(y,\theta)|y]$, a statement about probabilities of data sets, not models.
- If a $p$-value is extreme, it usually doesn't matter how extreme. For example a $p$-value of 0.00001 is effectively no stronger that a $p$-value of 0.001.
- As with normal $p$-values, these measure "statistical significance" not "practical significance". Small changes to the model can make large changes in the p-value.

# Other Related Issues



- Limitations of posterior tests
  - Finding an extreme p-value is never the end of an analysis
  - Model improvements are often needed in the next step
- Multiple comparisons
  - We often evaluate a model with several test quantities from multiple angles
  - We can do a ‘multiple comparisons’ adjustment by calculating the probability of the most extreme p-value (e.g. Bonferroni correction)
  - But, we do not suggest this adjustment in model checking
  - We are not concerned with ‘Type I error’ (accept or reject a model)
  - We try to understand the limits of its applicability in realistic replications
- Likelihood principle

# Approach III: Graphical Posterior Predictive Checking

## INTERNAL VALIDATION

# Graphical Posterior Predictive Checks



- Basic idea:
  - Display the data alongside simulated data from the fitted model
  - Look for systematic discrepancies between real and simulated data.

- Three kinds of graphical display:
  1. Direct display of all the data ——— Psychology Example
  2. Display of data summaries or parameter inferences.
  3. Graphs of residuals or other measures of discrepancy between model and data. ——— Rat: hierarchical Binomial; Beer: hierarchical Normal

# Example



- Data from an experiment in psychology.
- 6 persons in total.
- For each person, he / she answered ‘yes’ or ‘no’ to each of 15 possible reactions to 23 situations.
- Illustrative examples for possible reactions: 生气、大笑、大哭
- Illustrative examples for situations: 得知有期中考试、得知期中考试取消了

# Direct Data Display

15 possible reactions (rows)
to 23 situations (columns)
of 6 individuals

7 independently simulated replications $y^{rep}$ from a fitted logistic regression model

look fuzzy and 'random'

have strong rectilinear structures that are clearly not captured in the model.

Suggesting a model failure

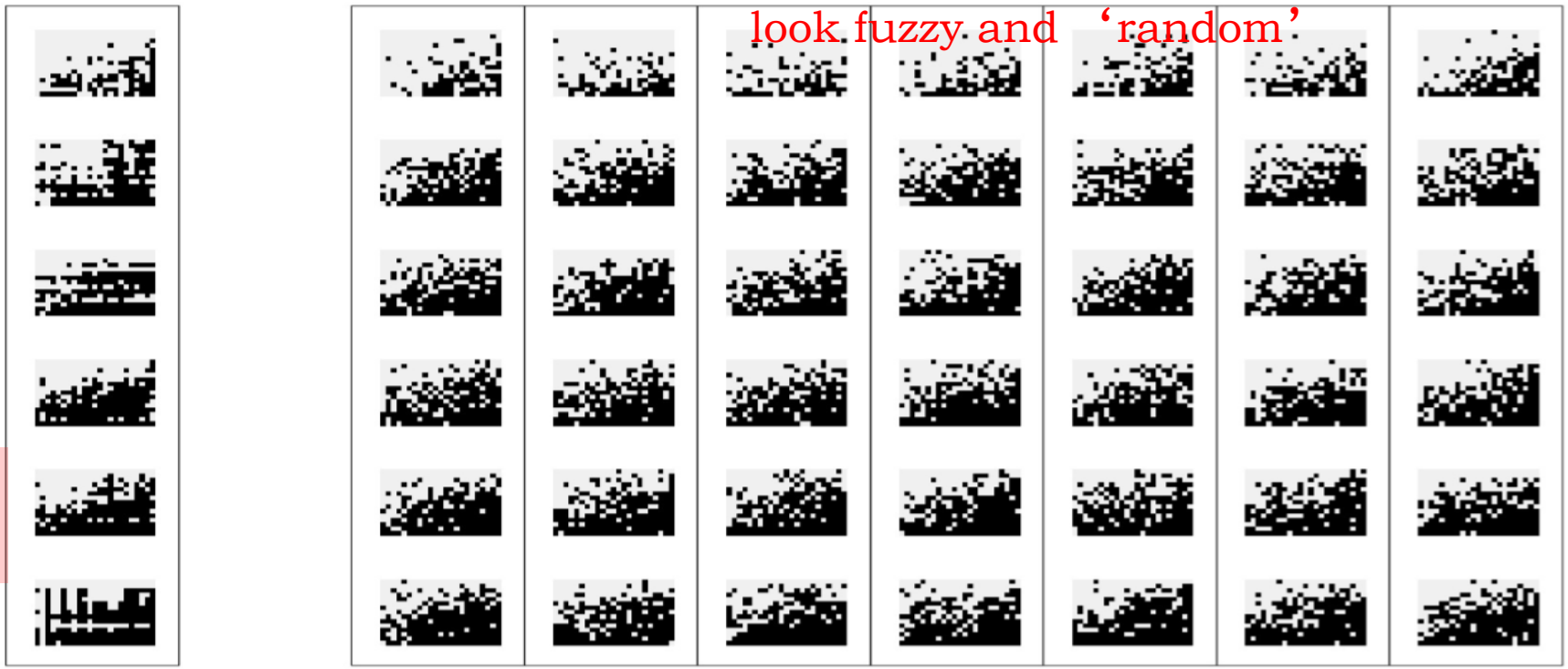


Figure 6.7 *Left column displays observed data y (a 15 × 23 array of binary responses from each of 6 persons); right columns display seven replicated datasets $y^{rep}$ from a fitted logistic regression model. A misfit of model to data is apparent: the data show strong row and column patterns for individual persons (for example, the nearly white row near the middle of the last person's data) that do not appear in the replicates. (To make such patterns clearer, the indexes of the observed and each replicated dataset have been arranged in increasing order of average response.)*

# Direct Data Display

15 possible reactions (rows)
to 23 situations (columns)
of 6 individuals

7 independently simulated replications $y^{rep}$ from a fitted logistic regression model

Look fuzzy and ‘random’ as wel

No model failures can Be detected

**Remark:**
find a proper way to display the data is a real **art**!

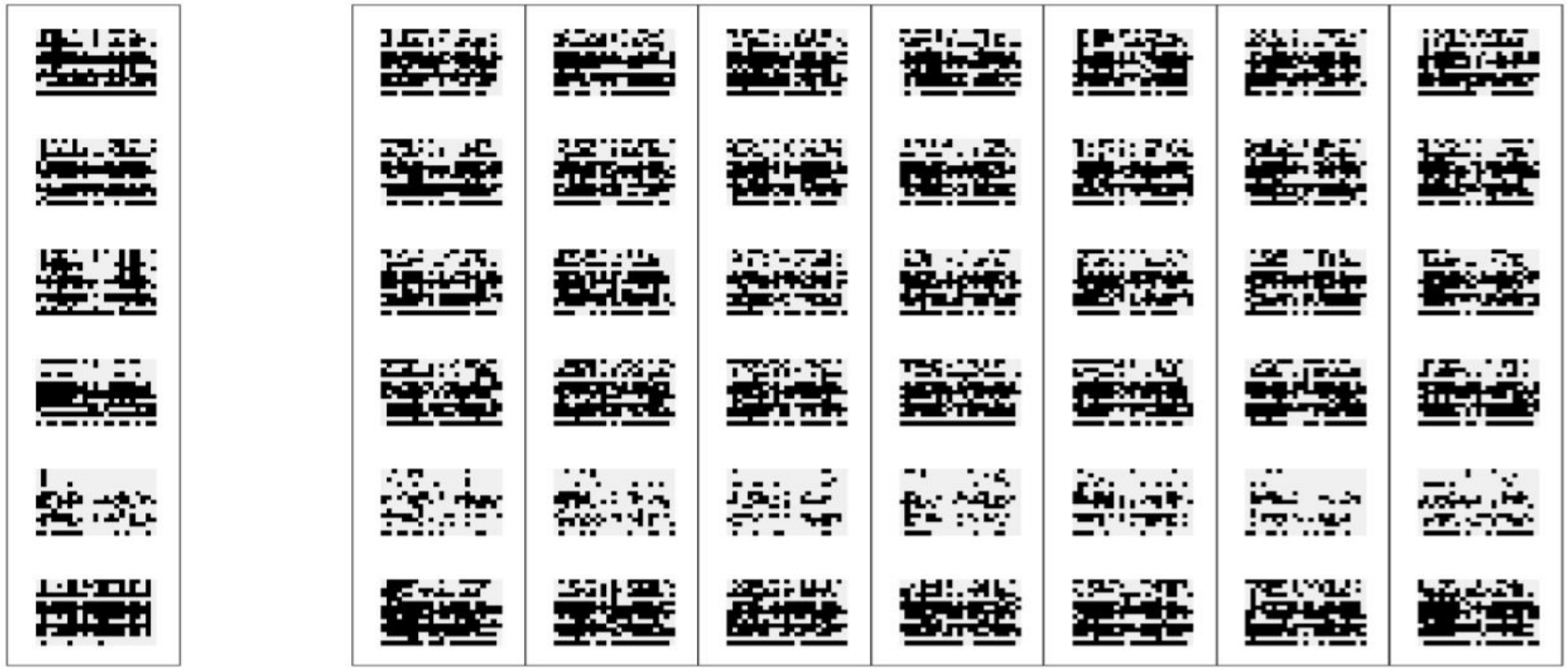


Figure 6.8 *Redisplay of Figure 6.7* without *ordering the rows, columns, and persons in order of increasing response. Once again, the left column shows the observed data and the right columns show replicated datasets from the model. Without the ordering, it is difficult to notice the discrepancies between data and model, which are easily apparent in Figure 6.7.*

# Graphs based on $T(y, \theta)$



- Note that the Bayesian p-values only tell part of the story. It is also useful to look at the relationship between $T(y, \theta^s)$ and $T(y^{rep's}, \theta^s)$.

- The comparison can be displayed either as a scatterplot of the values $T(y, \theta^s)\ vs\ T(y^{rep's}, \theta^s)$ or a histogram of the differences $T(y, \theta^s) - T(y^{rep's}, \theta^s)$, or a histogram of their ratio, etc.

- Under the model, the scatterplot should be symmetric about the 45° line and the histogram should include 0, or 1, etc.

# Example: Rat Tumors



- Consider two models:

1. Variable tumor rates

$$y_i|\theta_i \sim^{ind} Bin(n_i, \theta_i)$$

$$\theta_i|\alpha,\beta \sim^{i.i.d.} Beta(\alpha,\beta)$$

$$p(\alpha,\beta) \propto \frac{1}{(\alpha+\beta)^{5/2}}$$

2. Common tumor rates

$$y_i|\theta \sim^{ind} Bin(n_i, \theta)$$

$$\theta|\alpha,\beta \sim Beta(\alpha,\beta)$$

$$p(\alpha,\beta) \propto \frac{1}{(\alpha+\beta)^{5/2}}$$

# Omnibus tests



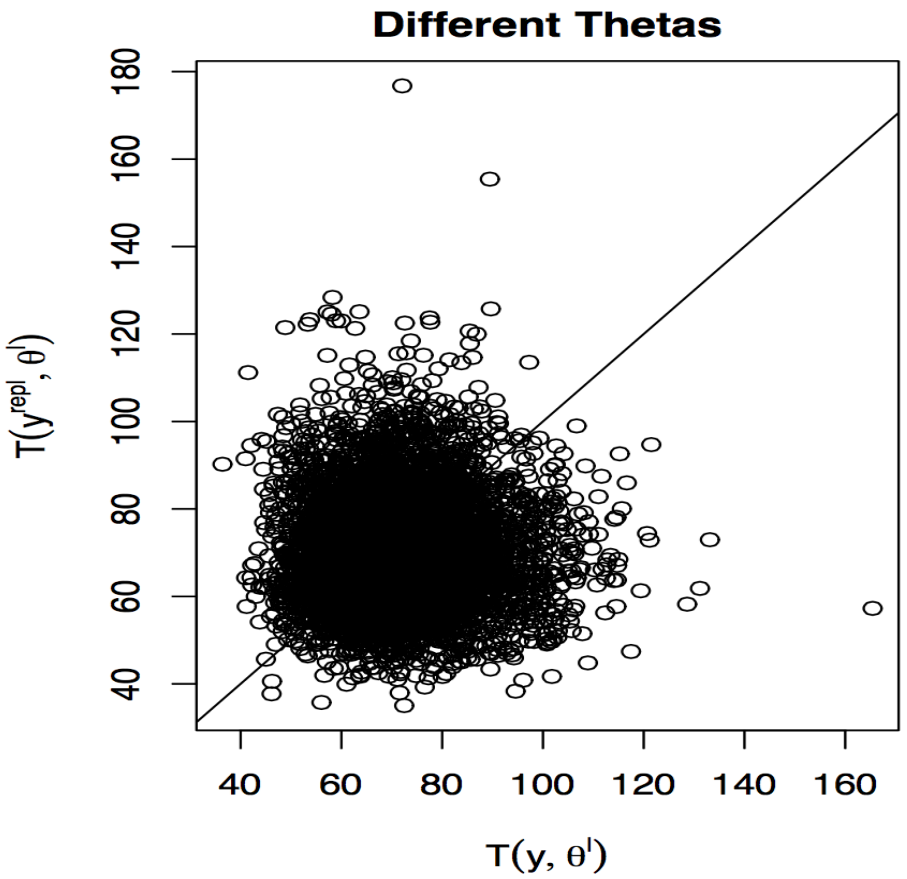
- In addition to focused test statistics, there are more general measures of fit. The most common one is the $\chi^2$ discrepancy

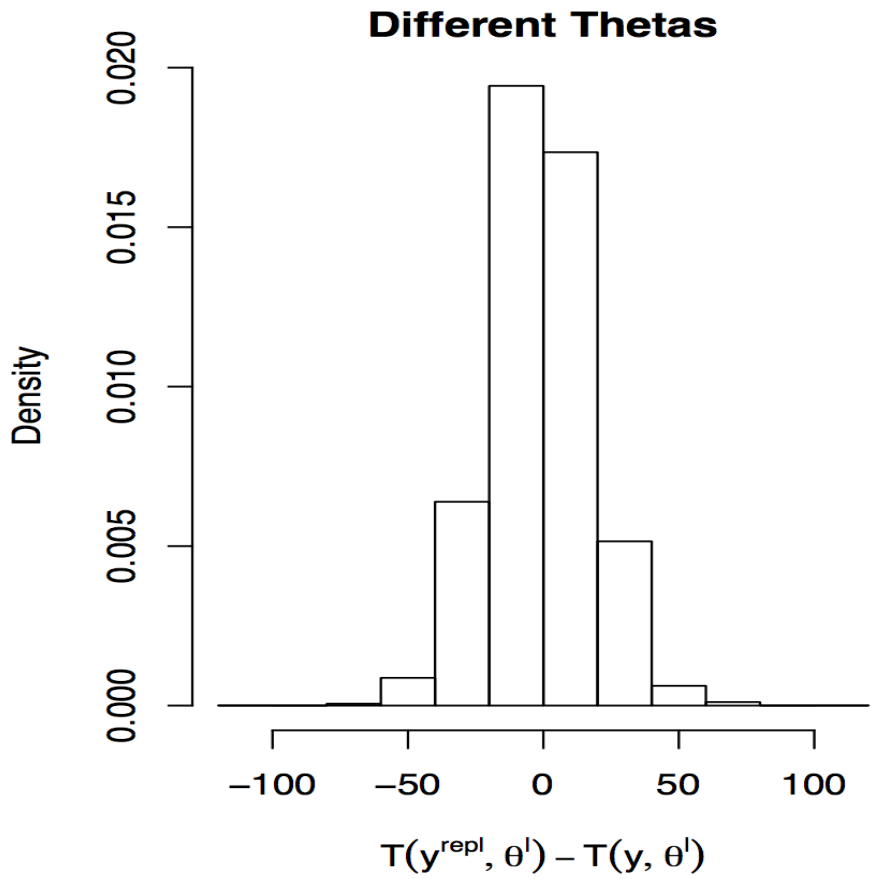
$$T(y,\theta)=\sum_i \frac{(y_i-E[y_i|\theta_i])^2}{\mathrm{Var}(y_i|\theta_i)}$$

- If $\theta$ is known, this is similar to the classical $\chi^2$ goodness of fit statistic.

# Example: Rat Tumors





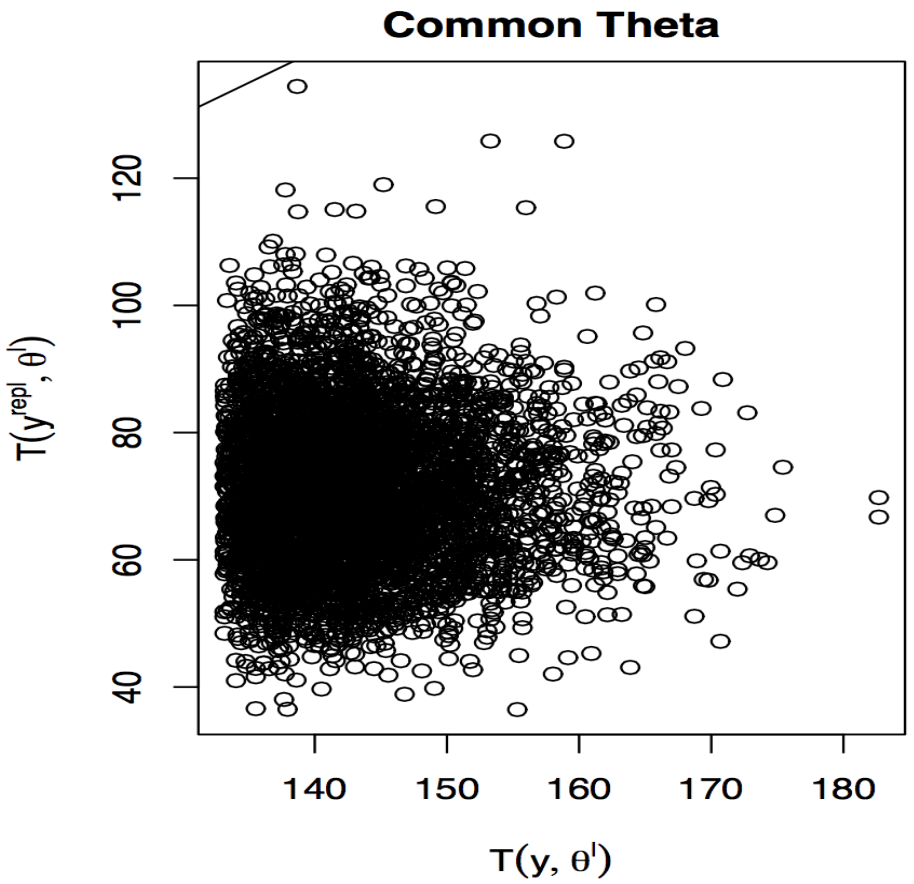



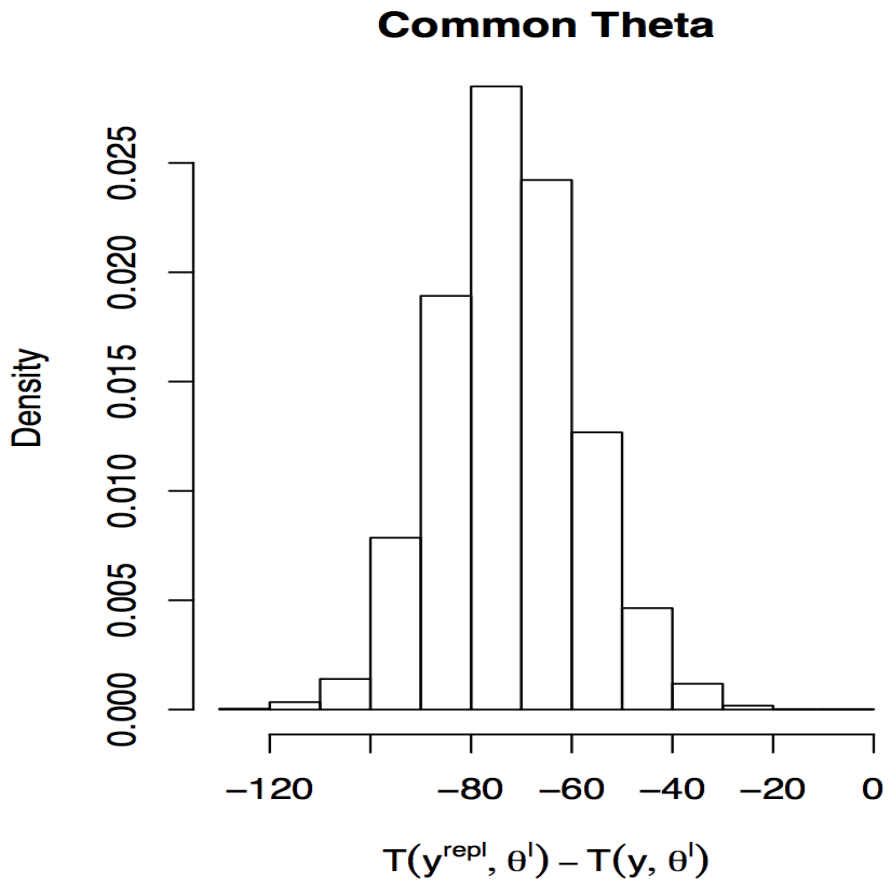

$$\hat{p}_B = 0.48$$

$$\hat{E}[T(y^{rep,l}, \theta^l)] = 71.28 \qquad \hat{E}[T(y, \theta^l)] = 72.09$$

# Example: Rat Tumors





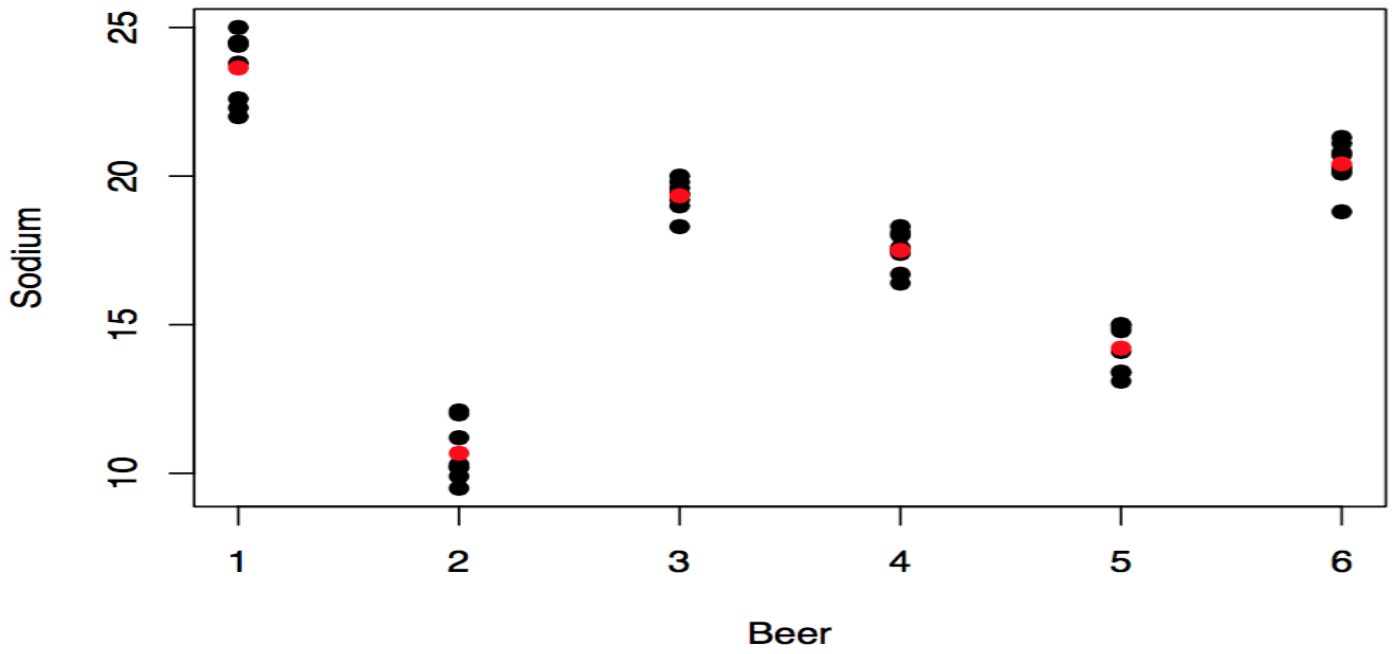

$$\hat{p}_B = 0$$

$$\hat{E}[T(y^{rep,l}, \theta^l)] = 70.89 \qquad \hat{E}[T(y, \theta^l)] = 143.09$$

# Example: Sodium Content in Beer

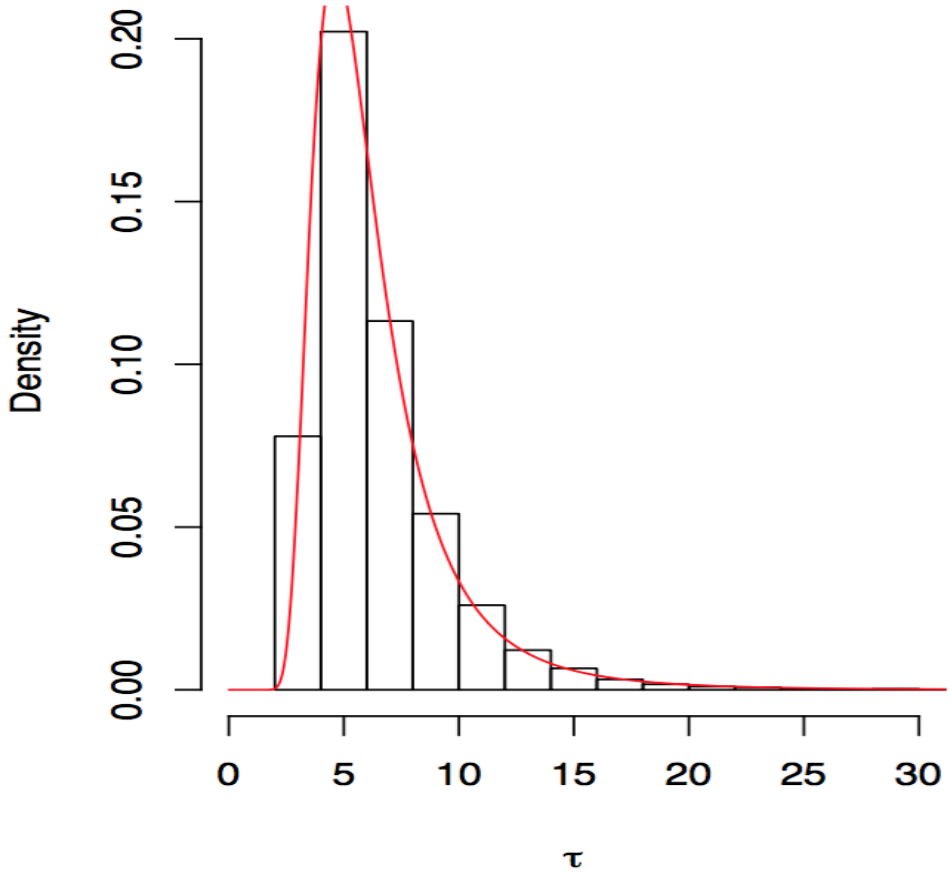


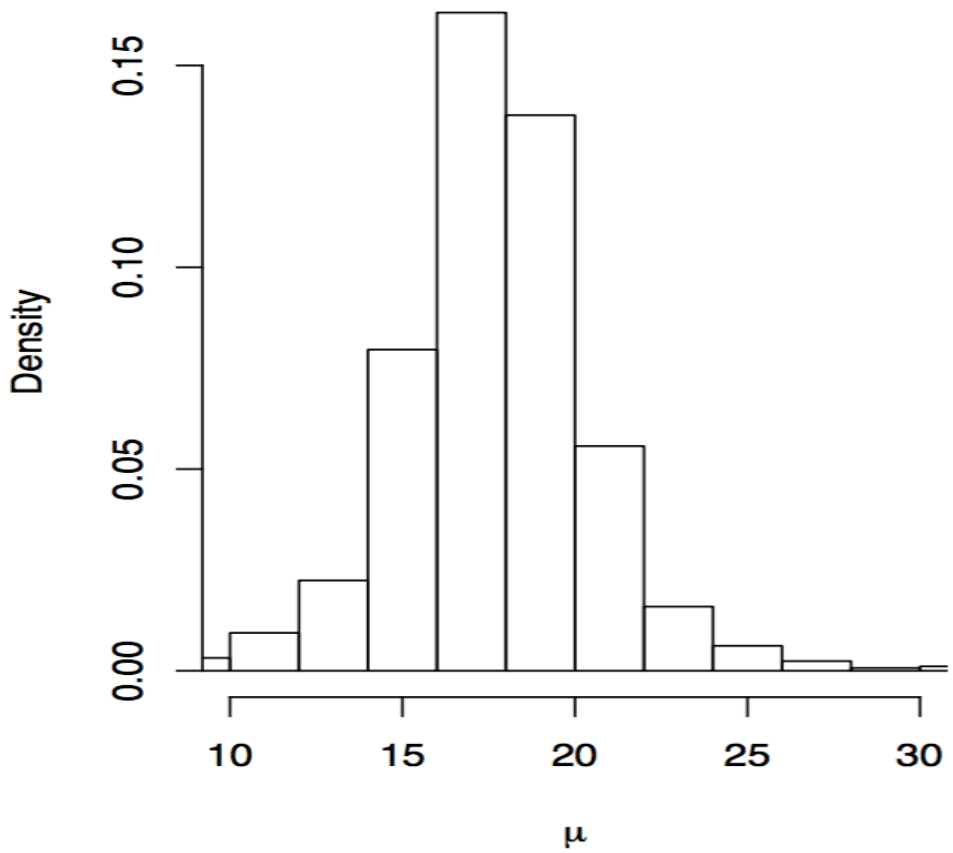
- A study was done to investigate the sodium content of 6 randomly chosen brands of beer. For each brand, 8 randomly chosen bottles or cans were analyzed to measure the sodium content (in mg) of each bottle or can.
- For this analysis, $\sigma_j^2 = 0.0895$, which is based on the MSE from the 1-way ANOVA.



# Example: Sodium Content in Beer






$E(\tau|y) = 6.448$
$Mode(\tau|y) = 4.61$
$E(\mu|y) = 17.67$

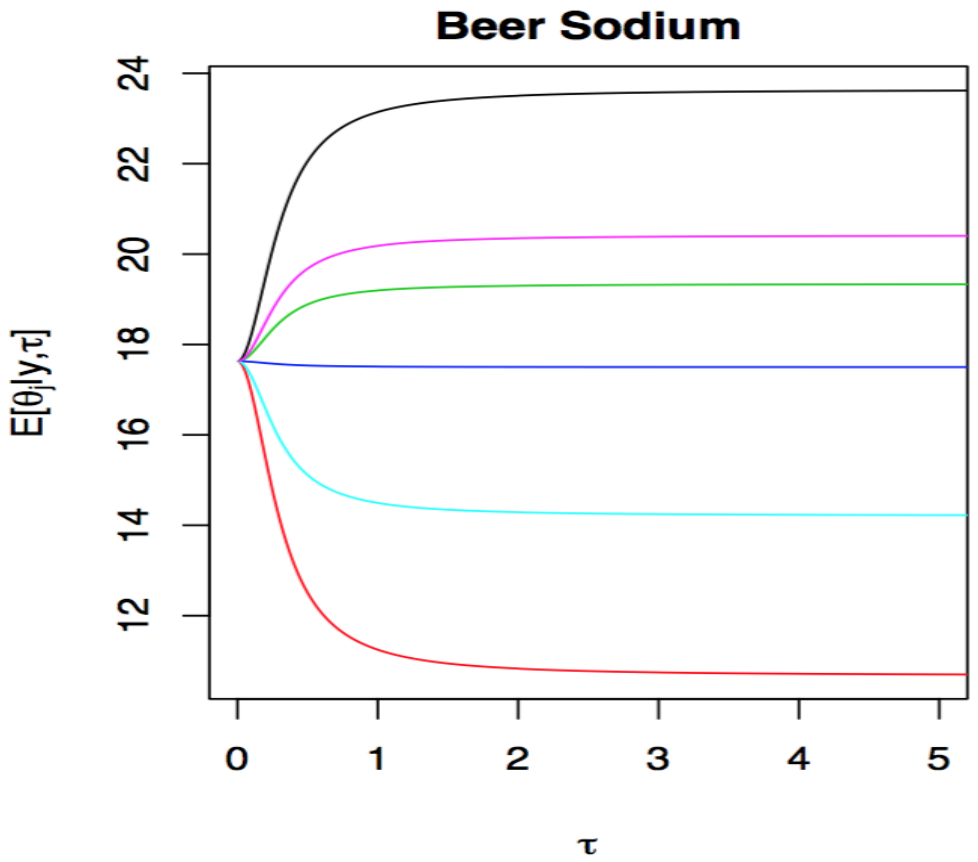
$E(\tau^2|y) = 50.847$
$Mode(\tau^2|y) = 21.25$
$SD(\mu|y) = 2.928$

# Example: Sodium Content in Beer



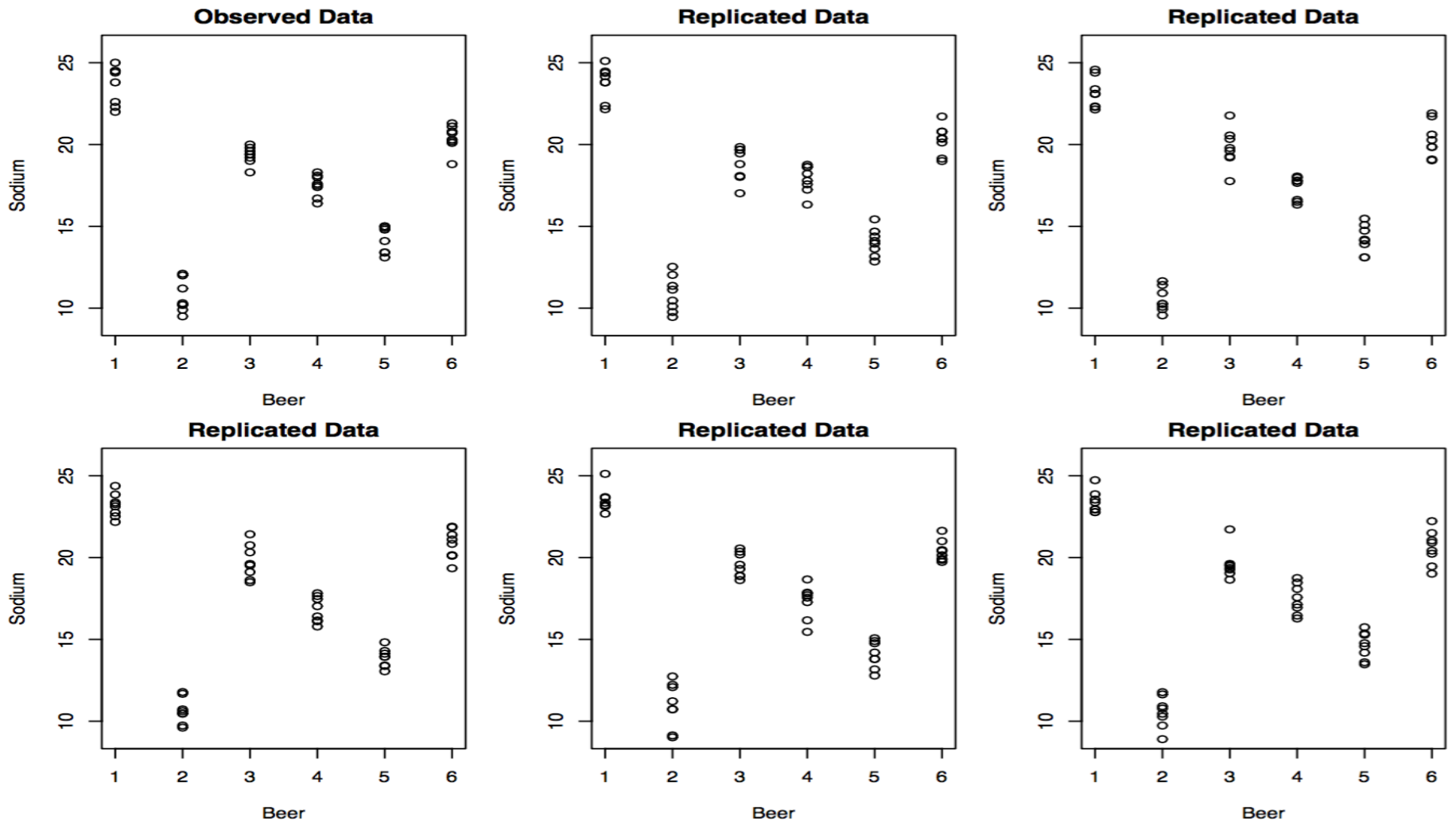
The relationship between the amount of shrinkage and $\sigma_j^2$ and $\tau^2$ can be seen by

$$E(\theta_i|\mu,\tau,y) = \frac{\tau^2}{\sigma_j^2+\tau^2}\bar{y}_{.j} + \frac{\sigma_j^2}{\sigma_j^2+\tau^2}\mu$$



# Example: Sodium Content in Beer





# Example: Sodium Content in Beer



- For this random effects model example, two concerns might be

1. Conditional normality of the observations

2. Constant variance of observations between groups

- Note that these are probably of limited concern in this example, as the total sample size is fairly large and there are equal numbers of observations in each group.

# Example: Sodium Content in Beer



- Possible test statistics to evaluate these are

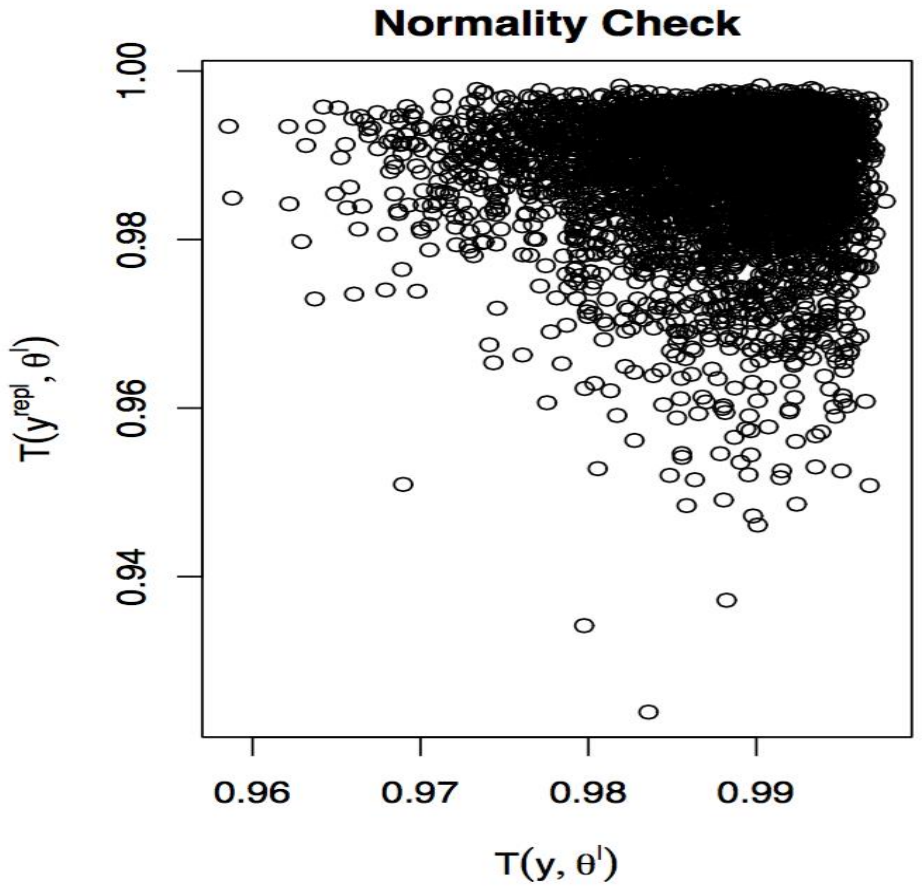
1. Normality:
   - Let $e_{ij} = y_{ij} - \theta_j$ and $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$ be the ordered residuals. Let

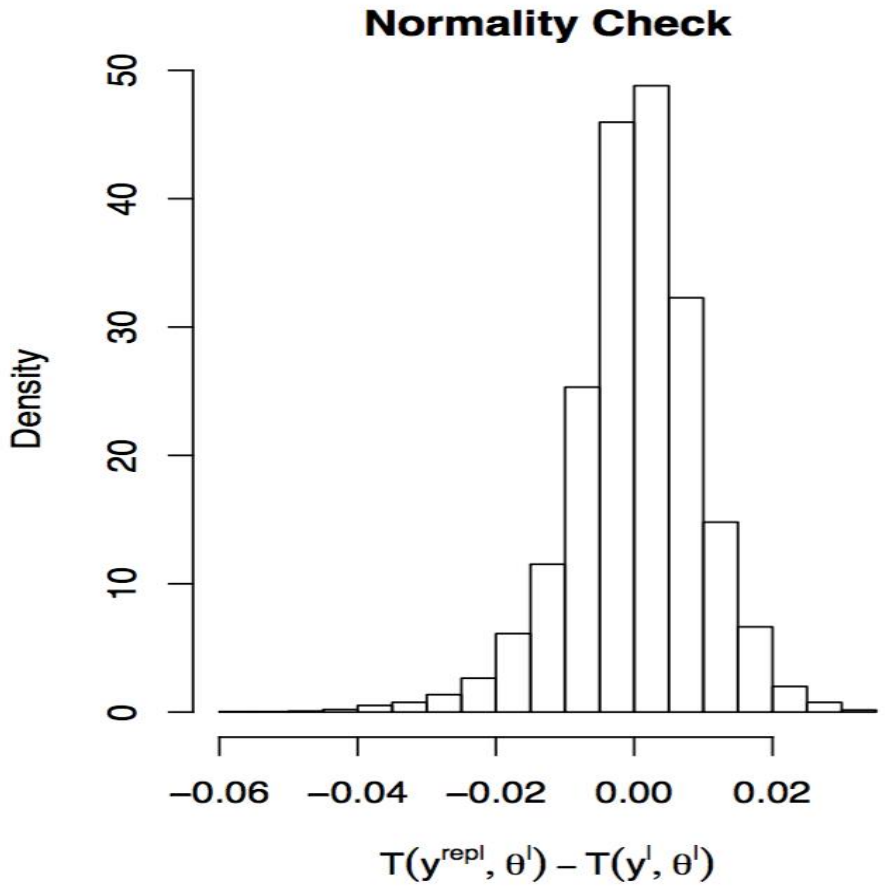
$$T(y, \theta) = corr(e_{(i)}, \Phi^{-1}\left(\frac{i - \frac{1}{2}}{n}\right))$$

   - (e.g. Correlation of points in a normal scores plot / QQ plot).
   - If the data is conditionally normal, this correlation should be close to one. Otherwise the normal scores plot will have some non-linearity, which will pull this correlation down from one.

# Example: Sodium Content in Beer






The histogram being centered at approximately 0 suggests that the fit of the observed data is roughly in the middle of what would be expected based on the posterior predictive distribution.



Normality test: $\hat{p}_B = 0.53$

# Example: Sodium Content in Beer



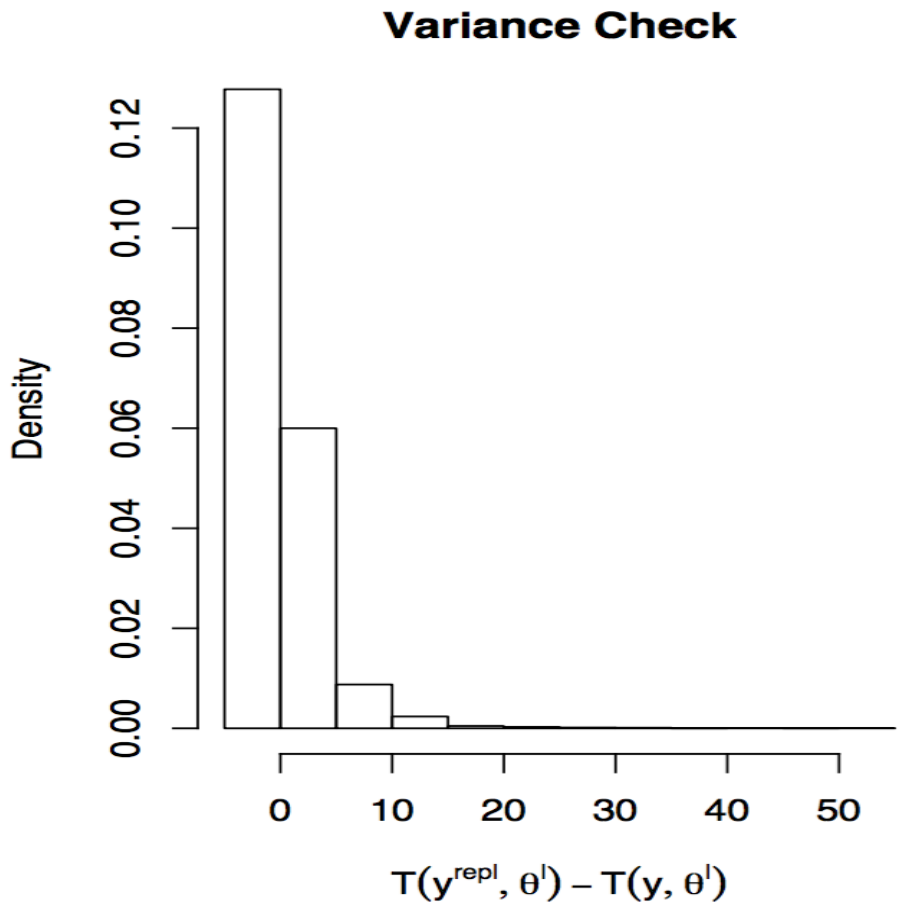
2. Equal variance:
   - Let $s_i^2$ be the sample variance of the observation in group $i$. If the constant variance assumption is reasonable,

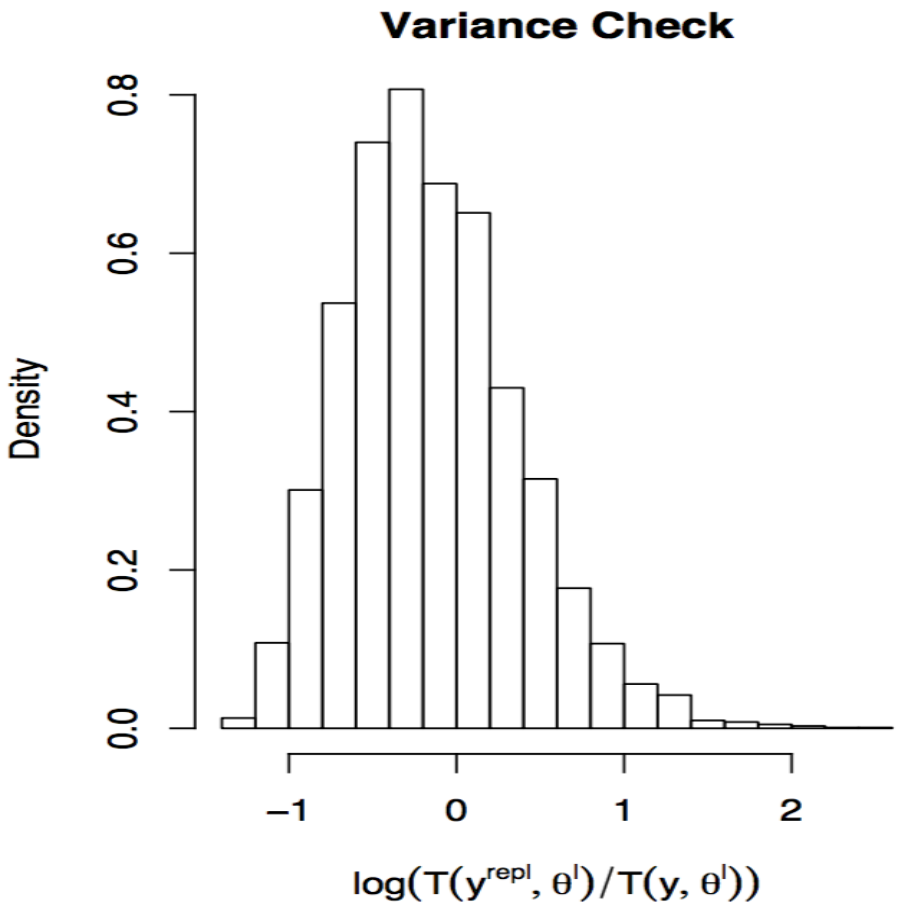
$$\mathrm{T}(\mathrm{y}, \theta) = \frac{\max s_i^2}{\min s_i^2}$$

   - should not be much bigger than one.

# Example: Sodium Content in Beer






Equal variance test: $\hat{p}_B = 0.35$

# Summary

# Key Points for Today



- Model checking helps us to decide if we need model improvement and find the direction for improvement.
- Posterior Predictive Checking
  - ✓ Similar to the idea of hypothesis testing, but not the same.
  - ✓ Test statistics $T(y, \theta)$: able to include parameter, unnecessary to know its exact distribution, multiple choices.
  - ✓ $p$-value: interpretation.
  - ✓ Implementation by simulation.
- Graphical Posterior Predictive Checking