Model Checking and Improvement II

Statistics 220

Spring 2005

Graphical Checks

As in standard frequentist analyses, graphical summaries are also useful examine the fit of a model. There are three common types of plots

1. Displaying data

2. Displaying data summaries or parameter inferences

3. Graphs of residuals or other discrepancy measures

Choosing $T(y, \theta)$

As mentioned last time, $T(y, \theta)$ should be chosen examine possible deviations of interest and examining more than one at a time is reasonable.

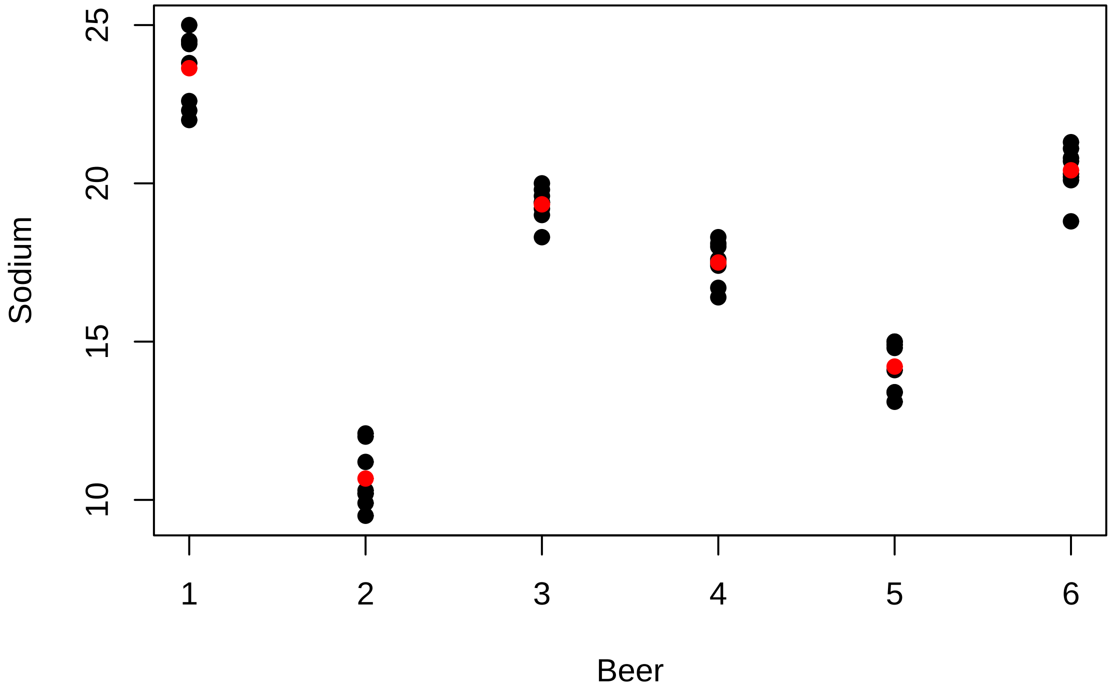
For example, in the examples last time, two statistics were studied, one investigating the normality of deviations from the means in a one-way ANOVA and the other investigating homoscedacity of the deviations.

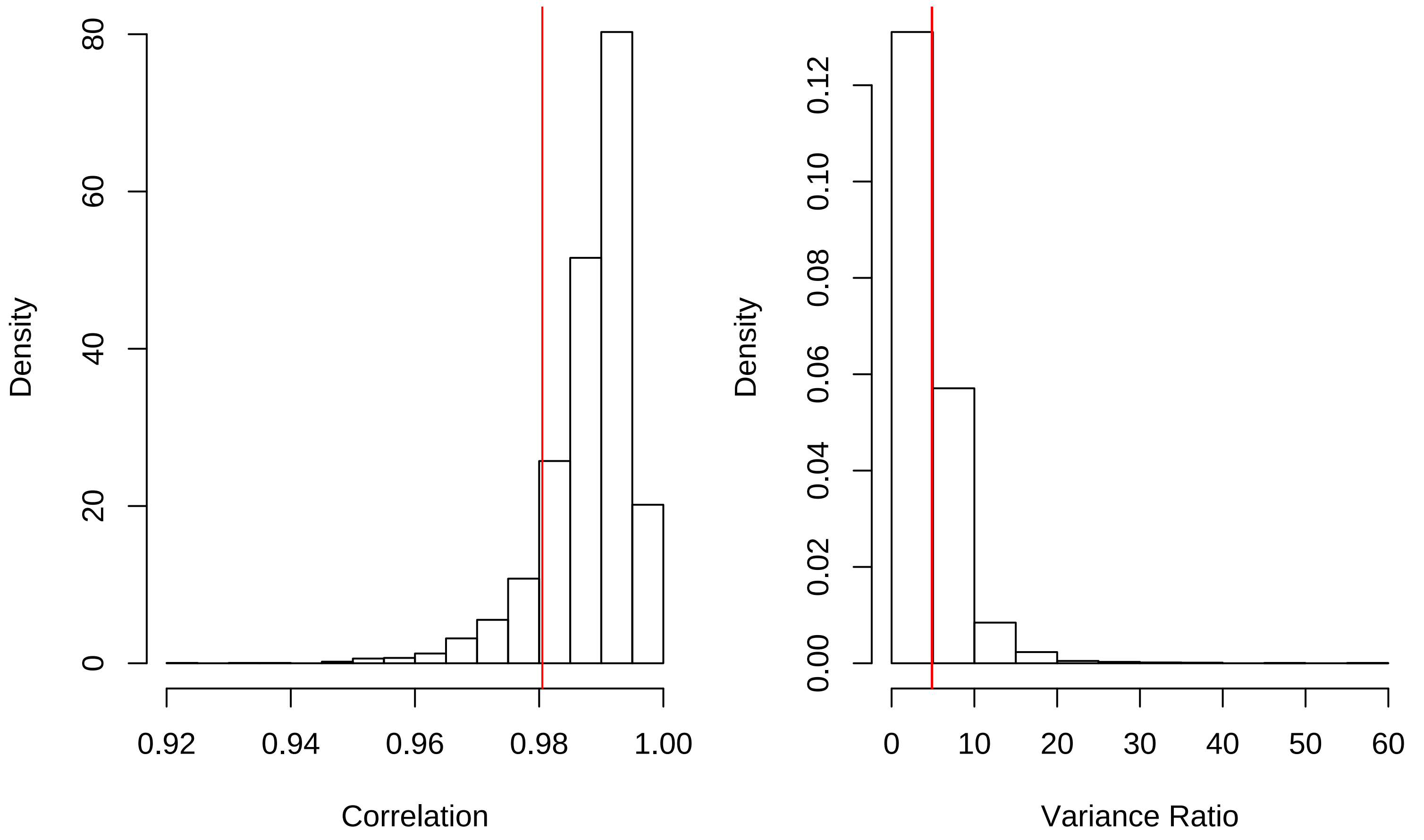
For each $T(y, \theta)$, an estimate of the Bayesian p-value

$$\hat{p}_B = \frac{1}{L} \sum_{l=1}^{L} I(T(y^{repl}, \theta^l) \geq T(y, \theta^l))$$

was calculated.

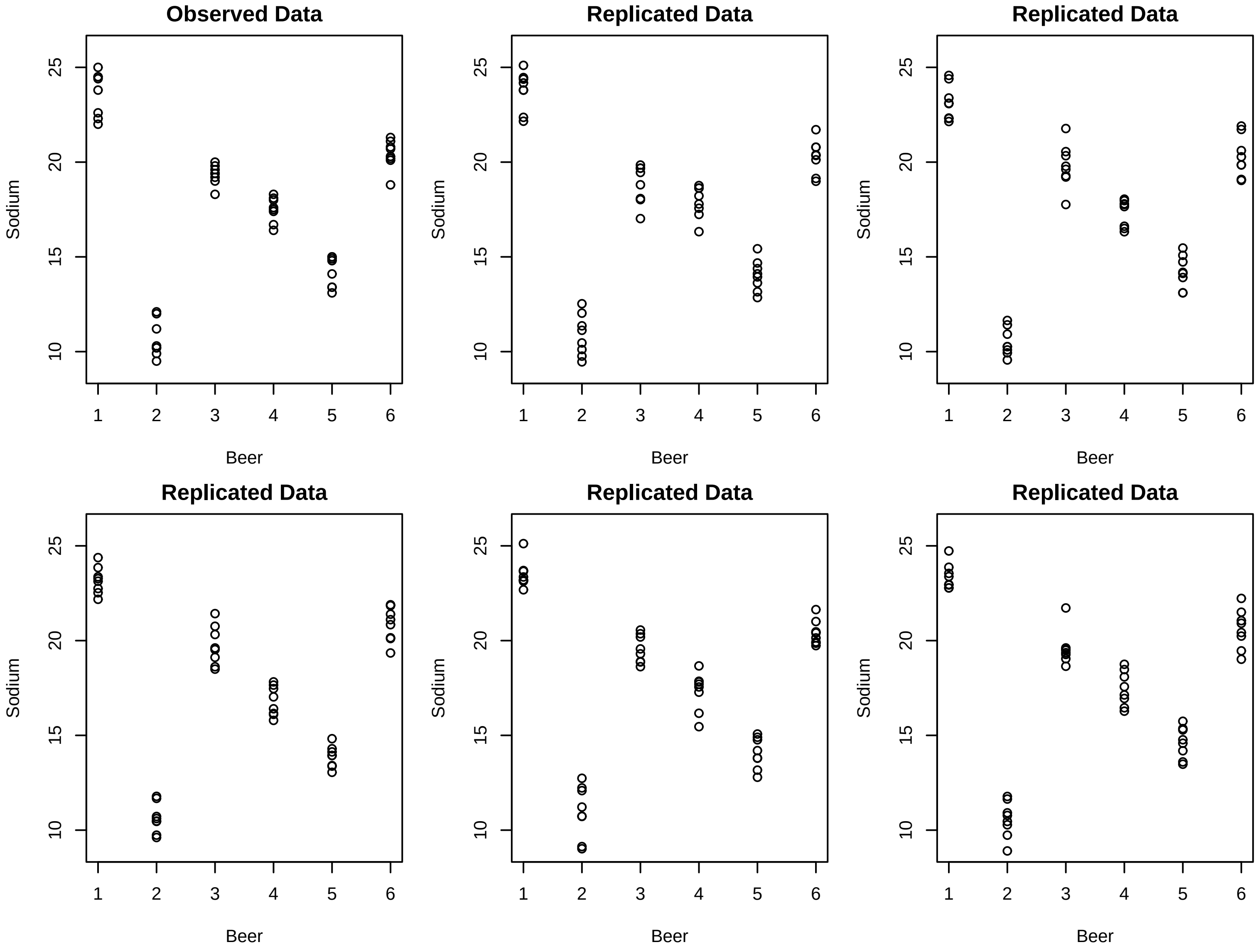
Beer example:





Normality test: $\hat{p}_B = 0.5420$

Equal variance test: $\hat{p}_B = 0.3408$


Observed Data
Replicated Data
Replicated Data
Replicated Data
Replicated Data
Replicated Data
Sodium
10
15
20
25
Beer
1
2
3
4
5
6

Note that the Bayesian p-values only tell part of the story.

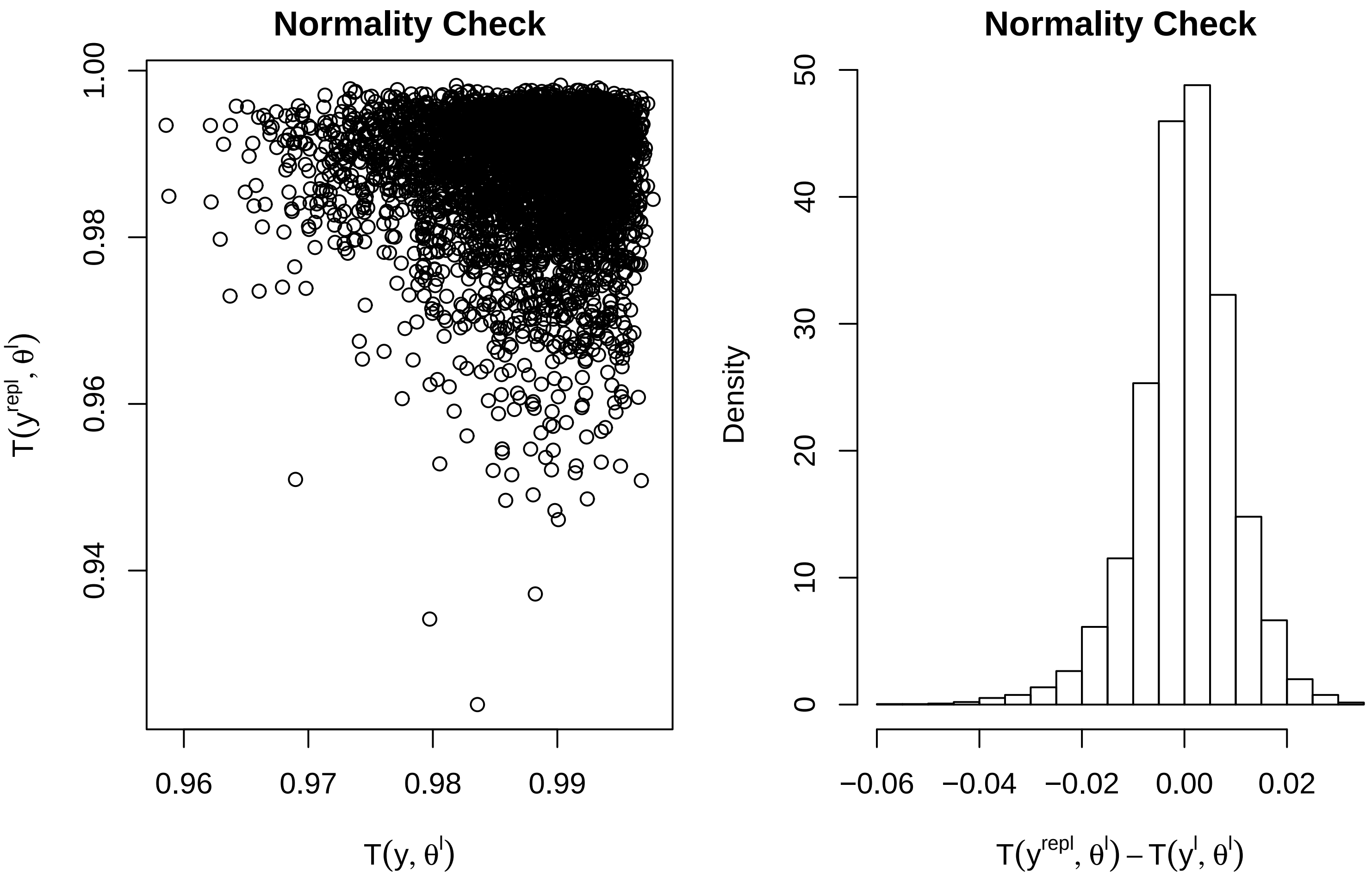
It is also useful to look at the relationship between $T(y^{repl}, \theta^l)$ and $T(y, \theta^l)$.

This could be done as in the previous graphs or by

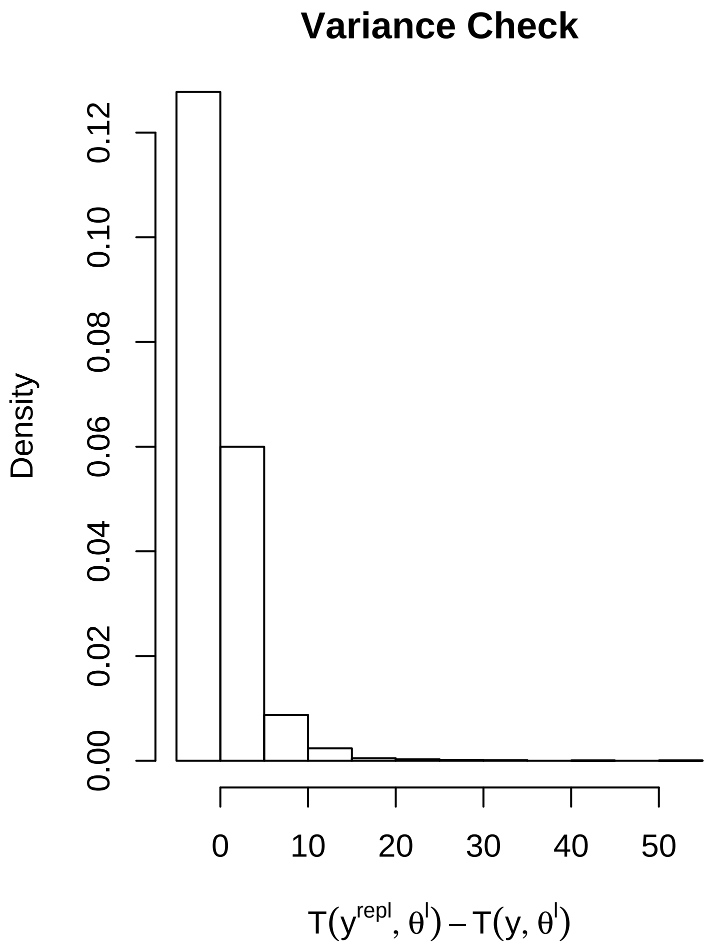
- plotting $T(y^{repl}, \theta^l)$ versus $T(y, \theta^l)$
- a histogram of $T(y^{repl}, \theta^l) - T(y, \theta^l)$
- a histogram of $\frac{T(y^{repl}, \theta^l)}{T(y, \theta^l)}$ or of $\log T(y^{repl}, \theta^l) - \log T(y, \theta^l)$

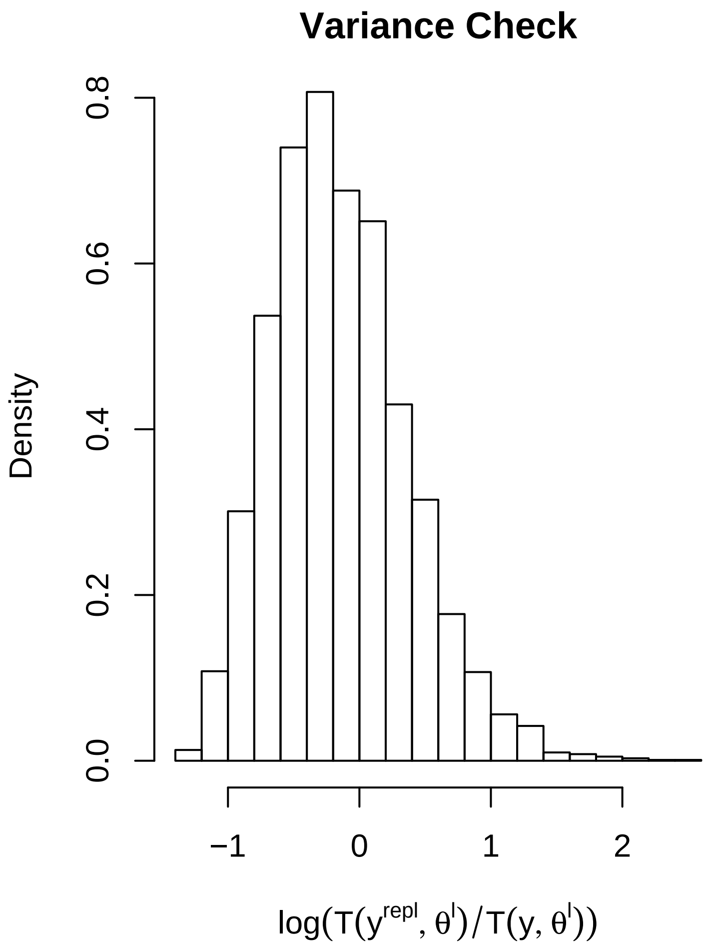
These are preferable to what I had last time for the normality check plot as the summary of the data fit on the plot (the red line) ignored the uncertainty in θ.

(The red line was the correlation in the normal scores plot using the residuals from a standard one-way ANOVA.)



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.


Variance Check
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30
40
50
$T(y^{repl}, \theta^l) - T(y, \theta^l)$


Variance Check
Density
0.0
0.2
0.4
0.6
0.8
−1
0
1
2
$\log(T(y^{repl}, \theta^l)/T(y, \theta^l))$

Multiple Comparisons:

In many situations with multiple test statistics you want to adjust for multiple looks at the data. For example, a Bonferroni correction (involving k p-values)suggests that the p-values should be compared with $\frac{\alpha}{k}$ instead of α.

This is not recommending in this setting. There is no worry about “Type I error” rates here. We are not using the p-values to accept or reject a model but as summaries to investigate limits of the model in realistic replications.

Aside: While strictly not doing hypothesis tests, if the values of $T(y^{repl}|\theta^l)$ indicate how the model can be improved, the model should be abandoned in favour of a better one. So in one sense, you are sort of acting like rejecting one model in favour of another one.

Omnibus tests:

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

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

If θ is known, this is similar to the classical χ^2 goodness of fit statistic.

An alternative to this is $T(y, \theta) = -2 \log p(y|\theta)$, the deviance.

In the the classical setting, θ must be specified. This might be by

- $\theta = \theta_{null}$
- $\theta = \theta_{mle}$
- $\theta = \arg\min_\theta T(y, \theta)$

In the Bayesian approach, we average over θ and the sampling distribution is automatically calculated by the posterior predictive simulations.

Example: Rat Tumors

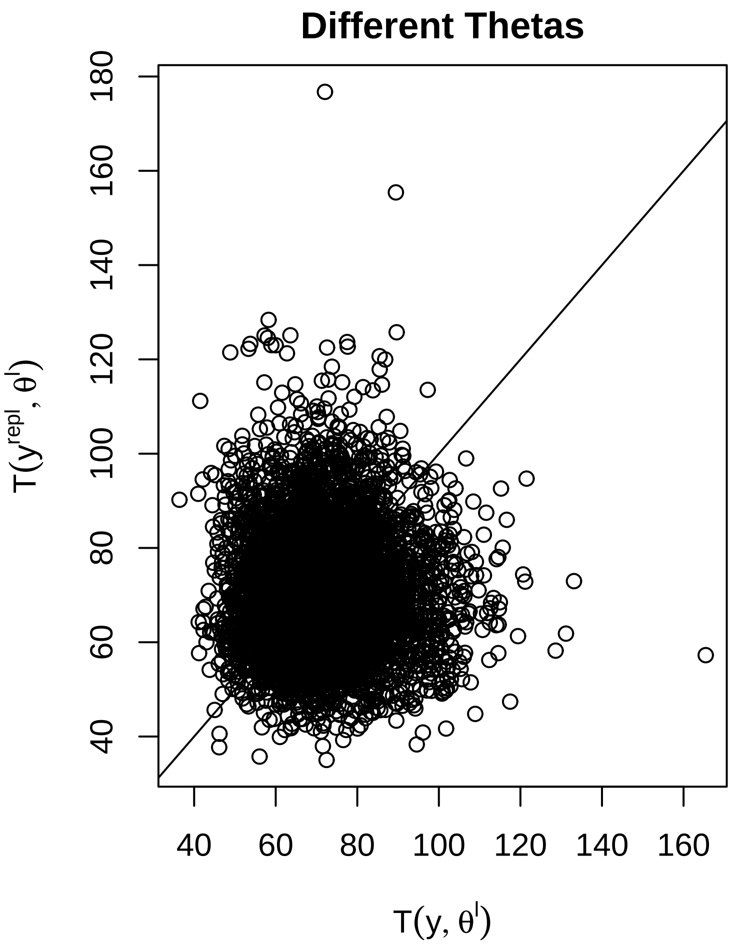
Compare two models:

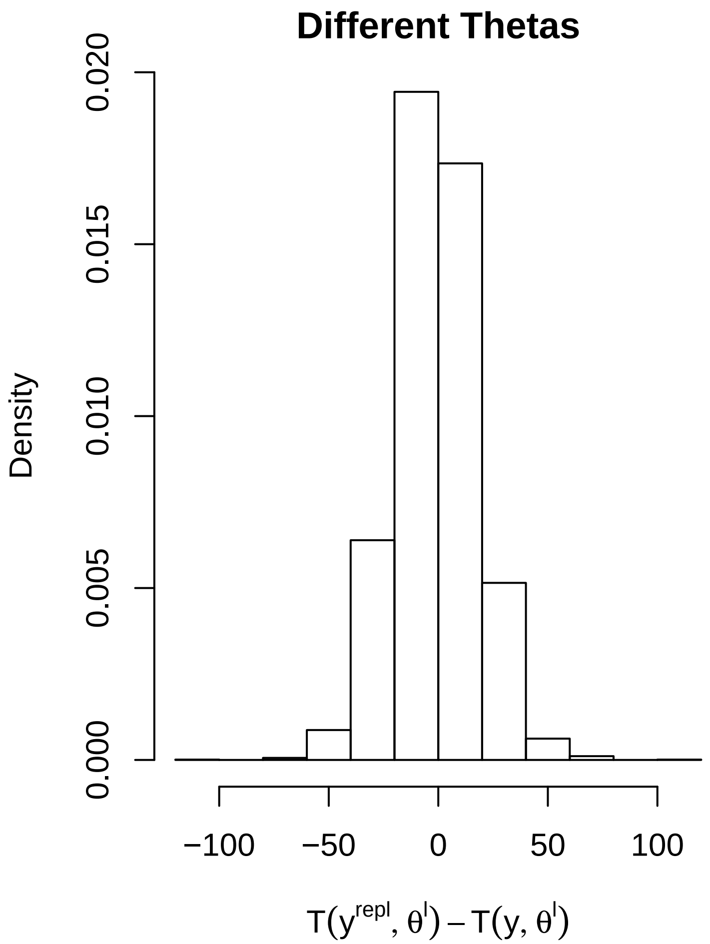
1. Variable tumor rates

$$\begin{aligned} y_i|\theta_i &\overset{ind}{\sim} Bin(n_i, \theta_i) \\ \theta_i &\overset{iid}{\sim} Beta(\alpha, \beta) \\ p(\alpha, \beta) &\propto \frac{1}{(\alpha+\beta)^{5/2}} \end{aligned}$$

2. Common tumor rates

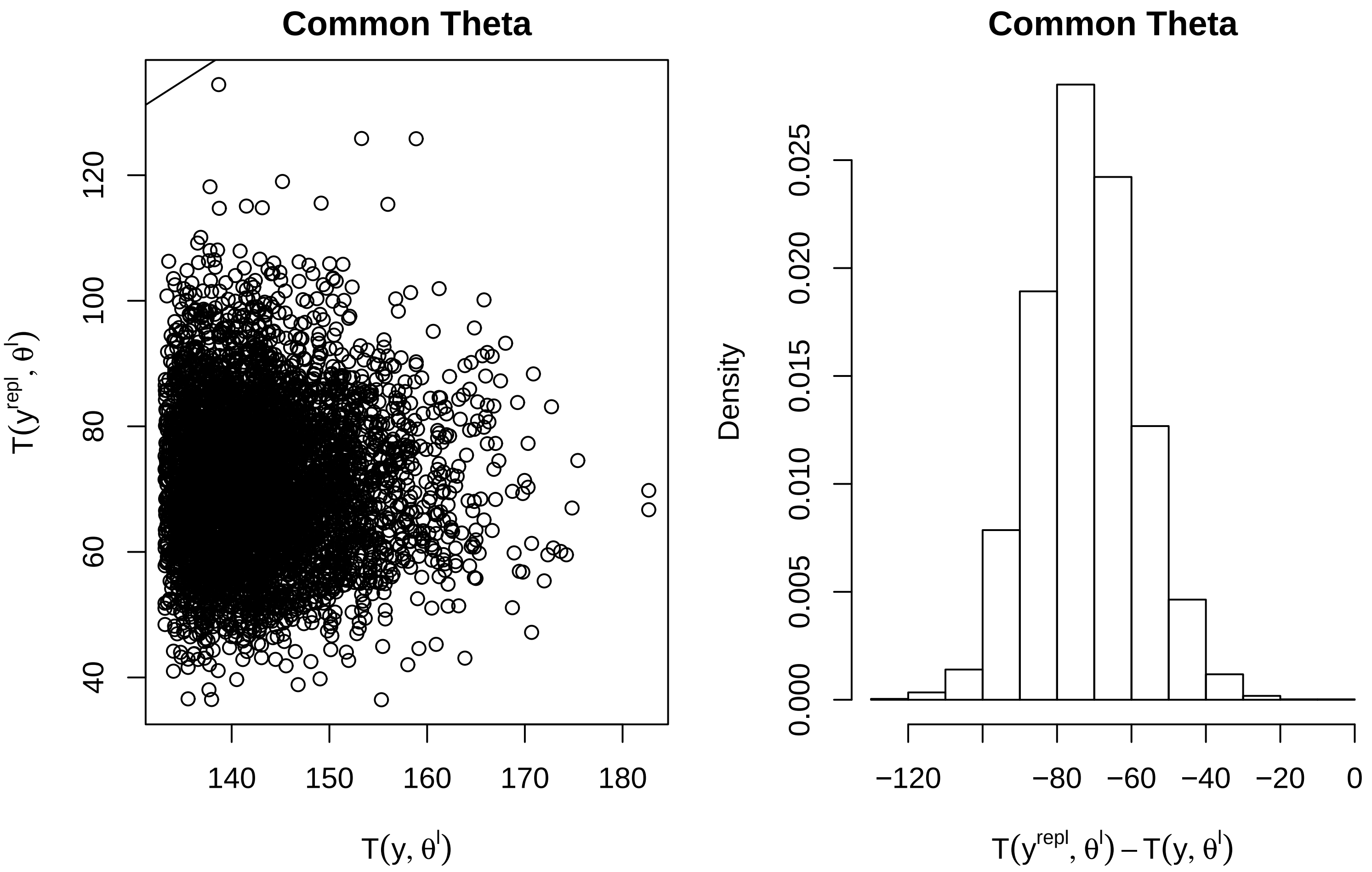
$$\begin{aligned} y_i|\theta &\overset{ind}{\sim} Bin(n_i, \theta) \\ \theta &\sim Beta(\alpha, \beta) \\ p(\alpha, \beta) &\propto \frac{1}{(\alpha+\beta)^{5/2}} \end{aligned}$$





$$\hat{p}_B = 0.4752$$

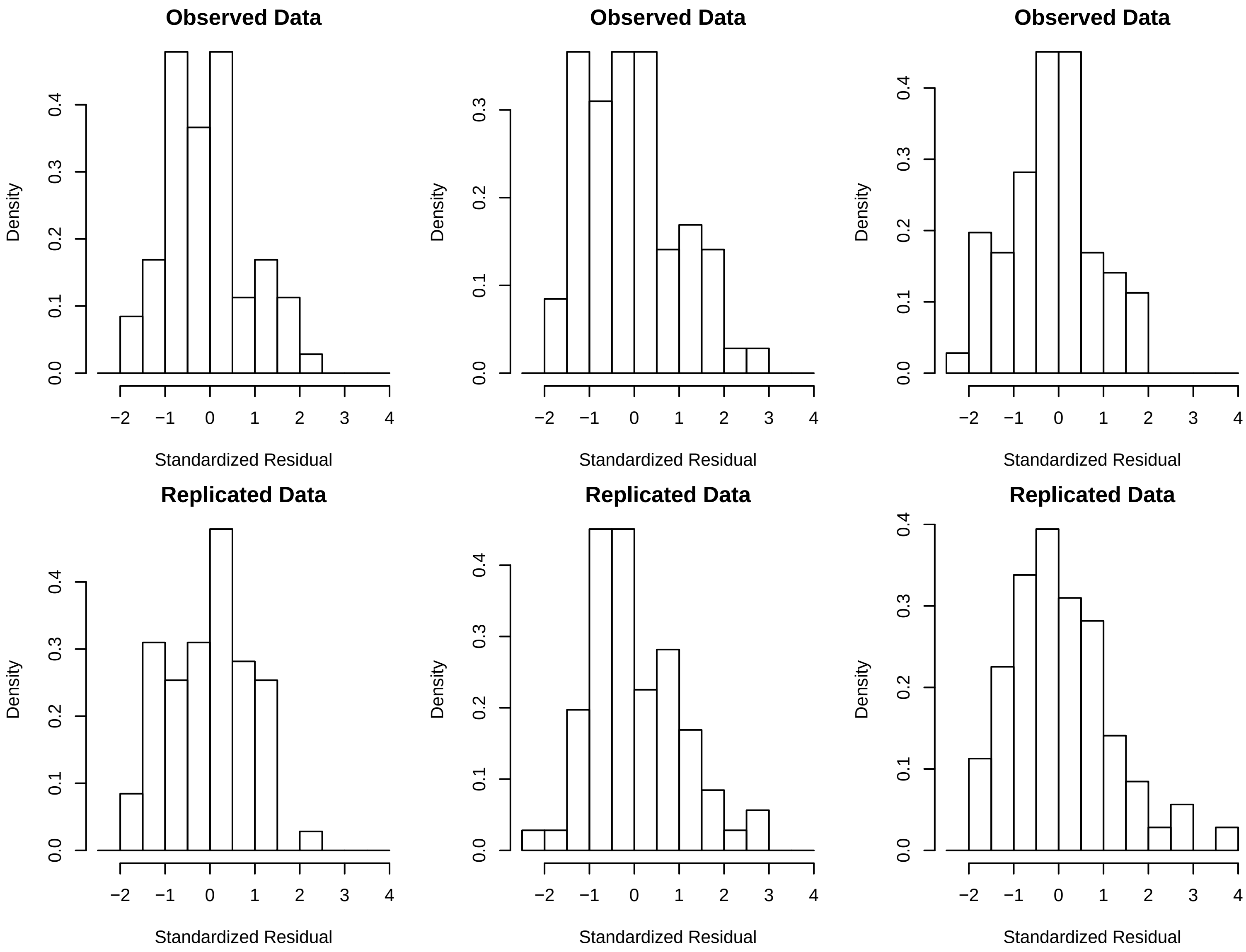
$$\hat{E}[y^{repl}, \theta^l] = 71.28 \qquad \hat{E}[y, \theta^l] = 72.09$$



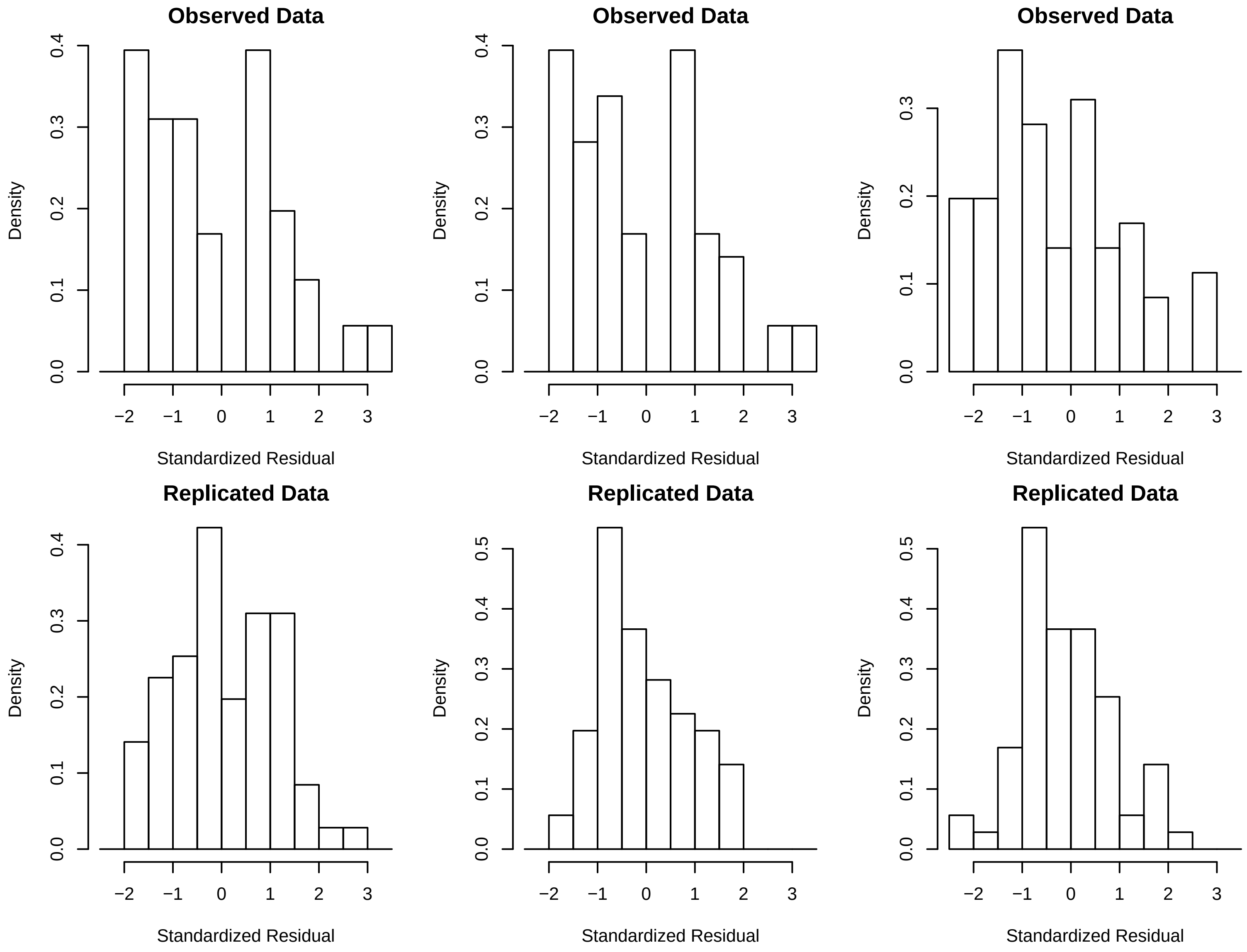
$$\hat{p}_B = 0$$

$$\hat{E}[y^{repl}, \theta^l] = 70.89 \qquad \hat{E}[y, \theta^l] = 143.09$$

Different Thetas



Common Theta


Observed Data

Observed Data

Observed Data

Replicated Data

Replicated Data

Replicated Data

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 | 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.

Model Expansion

Adding Parameters to a Model

While there are many ways of coming up with new models when the data doesn't seem to fit, adding parameters to a model is a common approach.

1. To deal with lack of fit or missing prior knowledge about the data, process, or parameters.
2. To get around questionable modelling assumption or ones with no justification.
3. If two (or more) possible models are under consideration, it may be possible to consider them as special cases or a more general model.

For example, for the two normal based models

$$\begin{aligned} y_{ij}|\theta_j &\overset{ind}{\sim} N(\theta_j, \sigma^2) \\ \theta_j &\overset{iid}{\sim} N(\mu, 10) \\ p(\mu) &\propto 1 \end{aligned}$$

and

$$\begin{aligned} y_{ij}|\mu &\overset{ind}{\sim} N(\mu, \sigma^2) \\ p(\mu) &\propto 1 \end{aligned}$$

are special cases of the model

$$\begin{aligned} y_{ij}|\theta_j &\overset{ind}{\sim} N(\theta_j, \sigma^2) \\ \theta_j &\overset{iid}{\sim} N(\mu, \tau^2) \\ p(\mu) &\propto 1 \\ \tau &\sim p(\tau) \end{aligned}$$

The first is $\tau^2 = 10$ and the second is $\tau^2 = 0$.

4. A model can be expanded to include additional data. One way of handling this is fitting the current piece into a hierarchical model.

   Or the model can be modified to handle different data structures. For example, allowing for missing data.

In these cases, the current model $p(y, \theta)$ is embedded into the model $p(y, \theta, \phi)$ or $p(y, y^*, \theta, \phi)$

This approach also gives a way to examine sensitivity to the choice prior. We could examine $p(\theta|y)$ directly under different fixed priors on θ (e.g. for different values of ϕ plugged into $p(\theta|\phi)$. Or we could put a prior on ϕ and look at

1. conditional posterior $p(\theta|y, \phi)$ for different choices of ϕ

2. marginal posterior $p(\theta|y)$

$$p(\theta|y) \propto \int p(\theta|y, \phi) p(\phi|y) d\phi$$

3. marginal posterior $p(\phi|y)$