6: Model Checking

10/07/19

Introduction

Now we finally admit some of the uncertainty we have about our choices of likelihood and prior distribution. Instead of asking "is our model true," we ask "are our inferences being substantially affected by the model's deficiencies?"

We are either making inferences on unobservable θ with $p(\theta \mid y)$, or *potentially* observable data with $p(\tilde{y} \mid y)$. This means it's generally easier to check inferences on the latter!

Notation Clarification

There is now a difference between y^{rep} and $\tilde{y}$!

1. y: the data we have observed and have made inferences using
2. y^{rep}: data simulated from the posterior predictive distribution (ppd)
3. $\tilde{y}$: actual future data, possibly coming from the true model we aren't using
4. $p(y^{\text{rep}} \mid y)$ our "working" ppd that we are examining and are unsure about

External Validation

The gold standard for evaluating the posterior predictive distribution is **external validation**, which is when you compare actual future data $\tilde{y}$ with your predictions coming from $p(y^{\text{rep}} \mid y)$.

In general, we might want to predict $T(\tilde{y})$, where T is some arbitrary test function of a new data (set) $\tilde{y}$.

In a time series context: predict, wait for new data to arrive, compare.

In a non-time series context: predict, wait for new data to be collected, compare.

Posterior Predictive Checks

When we can't/won't wait for new data to arrive, we can use our existing data set y by calculating a **posterior predictive p-value** p_B.

$$p_B = P(T(y^{\text{rep}}) > T(y) \mid y) = \int_{\{T(y^{\text{rep}}):T(y^{\text{rep}})>T(y)\}} p(y^{\text{rep}} \mid y) \mathrm{d}y^{\text{rep}}.$$

1. if $p_B = .5$, the median of $p(T(y^{\text{rep}}) \mid y)$ is exactly equal to $T(y)$.
2. if $p_B > .5$, the median of $p(T(y^{\text{rep}}) \mid y)$ is greater than $T(y)$.
3. if $p_B < .5$, the median of $p(T(y^{\text{rep}}) \mid y)$ is less than $T(y)$

Posterior Predictive Checks

$p_B = .5$ does not prove your model is good!

1. Maybe you're using an "easy" test quantity (e.g. a sufficient statistic)
2. Maybe it does poorly for other test quantities

Similarly, $p_B < \epsilon$ or $> 1 - \epsilon$ for some small ϵ does not prove your model is bad!

1. The feature of data, characterized by $T(y)$, is not addressed by the model
2. Maybe it does much better for other test quantities

Posterior Predictive Checks

We can extend this a bit further:

$$p_B = P(T(y^{\text{rep}}, \theta) > T(y, \theta) \mid y) = \iint_A p(y^{\text{rep}}, \theta \mid y) \mathrm{d}y^{\text{rep}} \mathrm{d}\theta$$

where $A = \{(y^{\text{rep}}, \theta) : T(y^{\text{rep}}, \theta) > T(y, \theta)\}$

Notation

We use superscripts for draws, and subscripts for indexes in a particular data set.

Example: $y^{i,\text{rep}}$ is the ith replicated data set

Example: y_i is the ith element of one data set

Posterior Predictive Checks

When we can't/won't evaluate this integral, we can use Monte Carlo!

$$
\begin{aligned}
p_B &= \iint_A p(y^{\text{rep}}, \theta \mid y) \mathrm{d}y^{\text{rep}} \mathrm{d}\theta \\
&= E\left[\mathbf{1}((y^{\text{rep}}, \theta) \in A) \mid y\right] \\
&\leftarrow S^{-1} \sum_{i=1}^{S} \mathbf{1}((y^{i,\text{rep}}, \theta^i) \in A)
\end{aligned}
$$

where

If we can't simulate directly from the ppd, then we can do the following.

For $i = 1, \ldots, S$

1. Simulate $\theta^i \sim p(\theta \mid y)$
2. Simulate $y^{i,\text{rep}} \sim p(y \mid \theta^i)$
3. Calculate the number of times $(y^{i,\text{rep}}, \theta^i) \in A$ divided by S

Example: Speed of Light Measurements

A histogram of real/observed/historical y

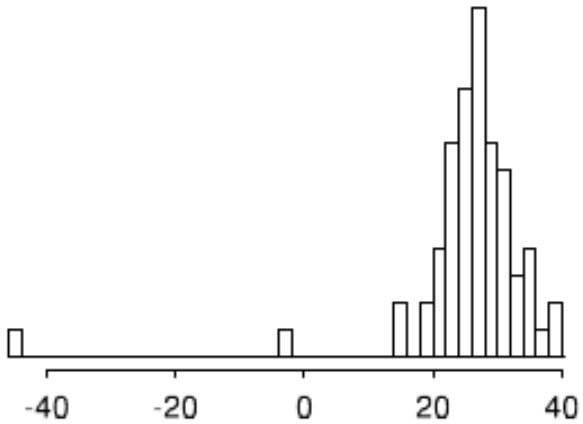


Figure 3.1 *Histogram of Simon Newcomb's measurements for estimating the speed of light, from Stigler (1977). The data are recorded as deviations from* 24,800 *nanoseconds.*

Example: Speed of Light Measurements

1. y: 66 univariate measurements
2. $p(y \mid \mu, \sigma^2) = \prod_{i=1}^{66} \text{Normal}(y_i \mid \mu, \sigma^2)$
3. $p(\mu, \sigma^2) \propto \sigma^{-2}$
4. $j = 1, \ldots, 20$ simulations
5. each $y^{i,\text{rep}}$ is a data set of size 66 simulated from the ppd

Example: Speed of Light Measurements

Let's simulate a data set $y^{\text{rep}} \sim p(y^{\text{rep}} \mid y)$

Recall from chapter 3:

$$y_i^{\text{rep}} \mid y \sim t_{n-1}(\bar{y}, s^2(1 + 1/n)).$$

```
n <- length(y)
s <- sd(y)
my <- mean(y)
sampt20 <- replicate(20, rt(n, n-1)*sqrt(1+1/n)*s+my) %>%
  as.data.frame()
dim(sampt20)
[1]   66 20
```

http://avehtari.github.io/BDA_R_demos/demos_ch6/demo6_1.html

Example: Speed of Light Measurements

Each one of these simulated data sets produces one univariate $T(y^{\text{rep}})$

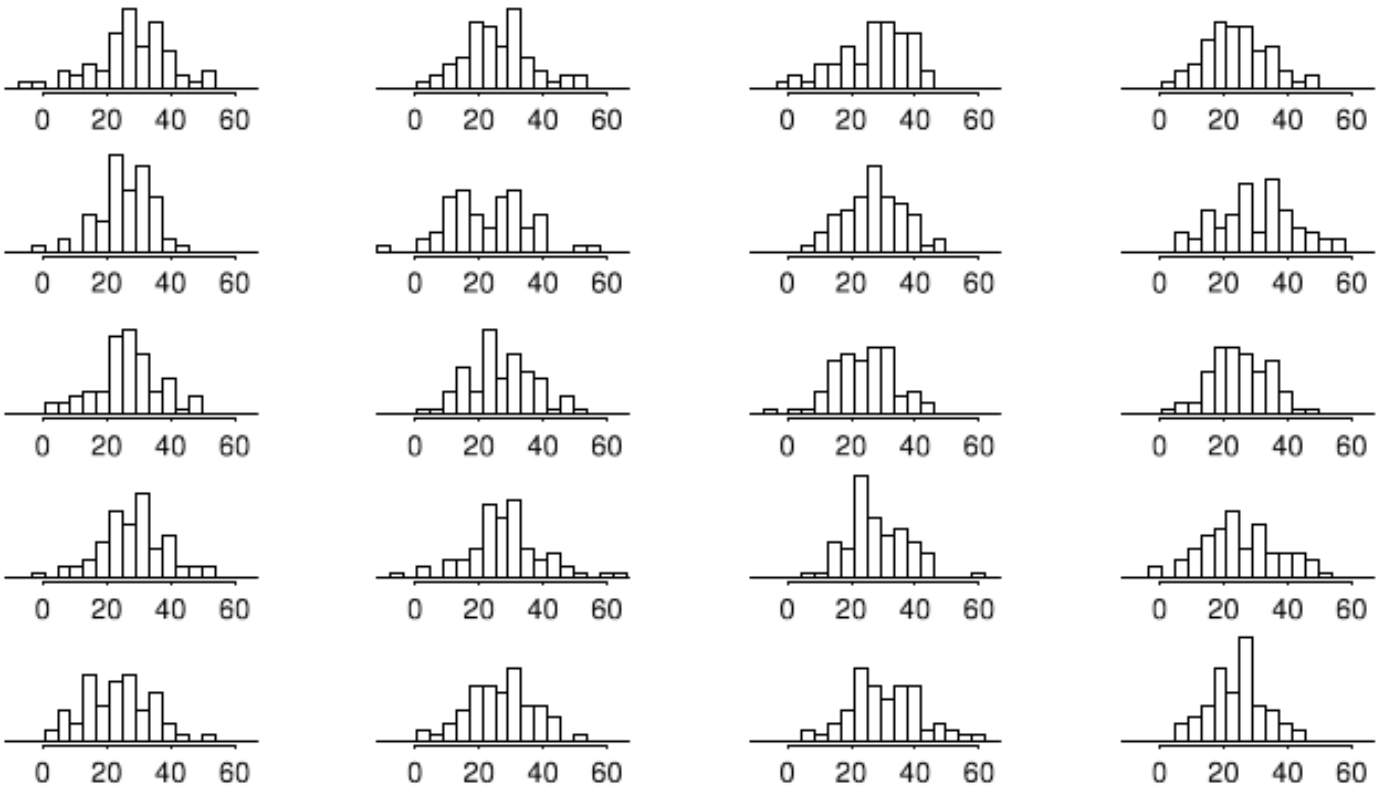


Figure 6.2 *Twenty replications, y^{rep}, of the speed of light data from the posterior predictive distribution, $p(y^{\text{rep}}|y)$; compare to observed data, y, in Figure 3.1. Each histogram displays the result of drawing 66 independent values $\tilde{y}_i$ from a common normal distribution with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

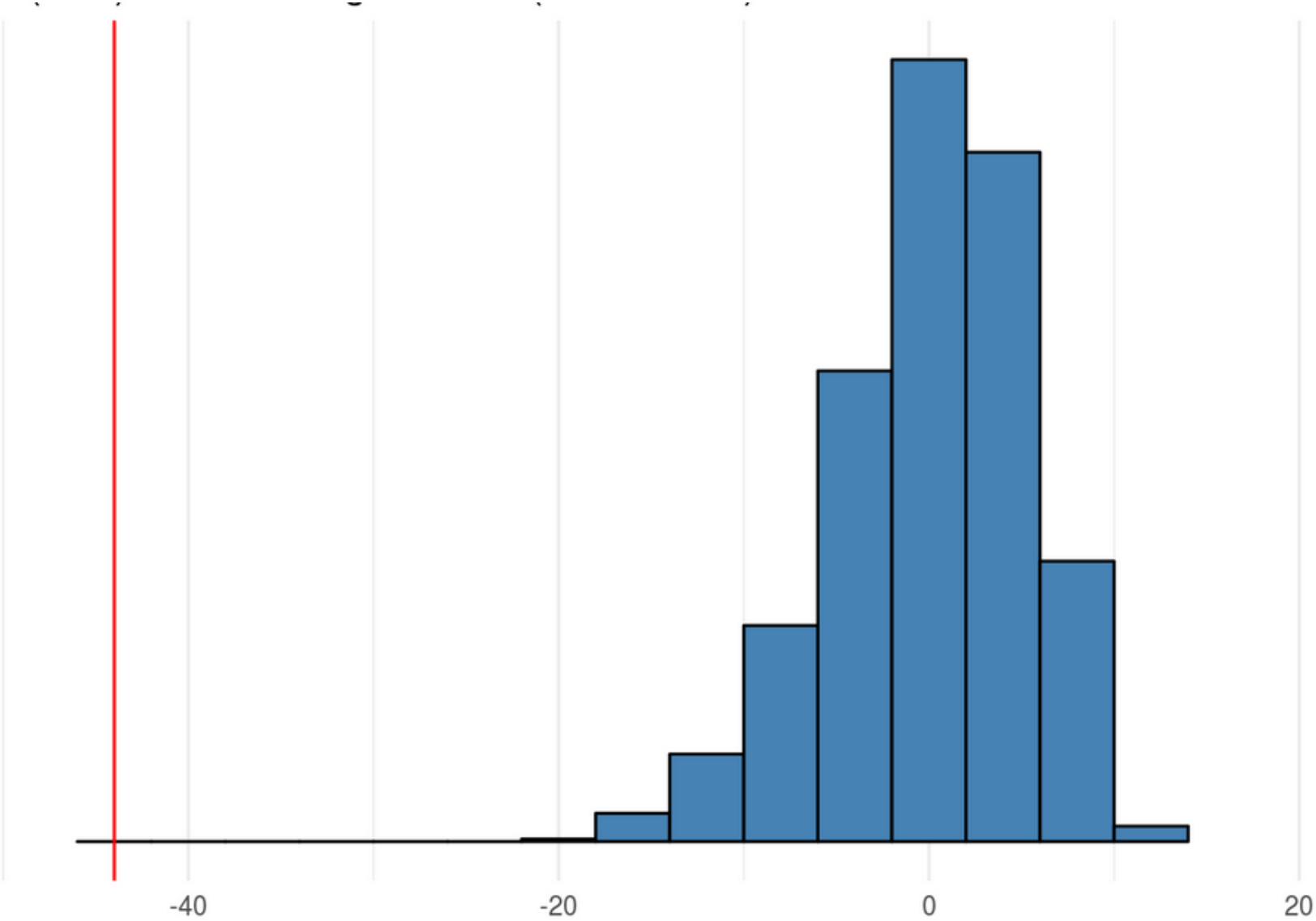
Example: Speed of Light Measurements

Let $T(y^{j,\text{rep}}) = \min(y_1^{j,\text{rep}}, \ldots, y_n^{j,\text{rep}})$ and $T(y) = \min(y_1, \ldots, y_n)$. Then

$$P(T(y^{\text{rep}}) > T(y) \mid y) \approx 1000^{-1} \sum_{j=1}^{1000} \mathbf{1}\left(T(y^{j,\text{rep}}) > T(y)\right)$$

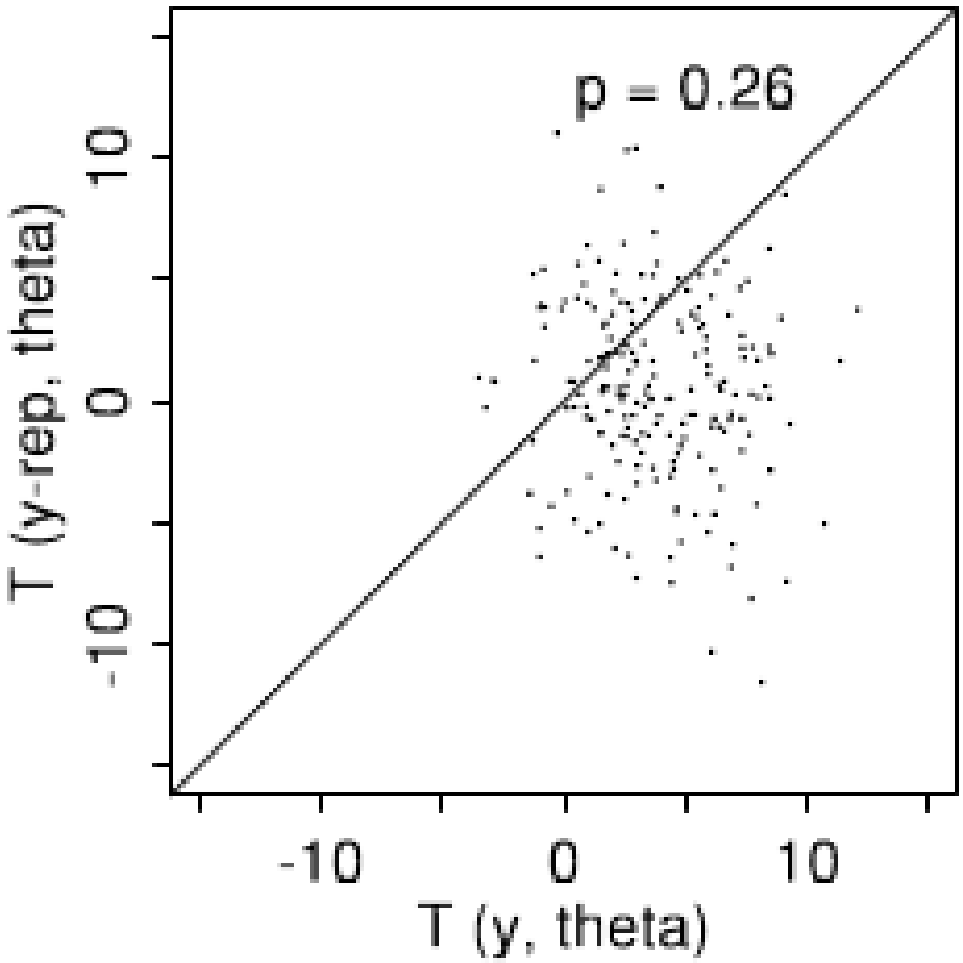
```
sampt1000 <- replicate(1000, rt(n, n-1)*sqrt(1+1/n)*s+my) %>%
  as.data.frame()
mean(sapply(sampt1000, min) > min(y))
[1] 1
```

Example: Speed of Light Measurements



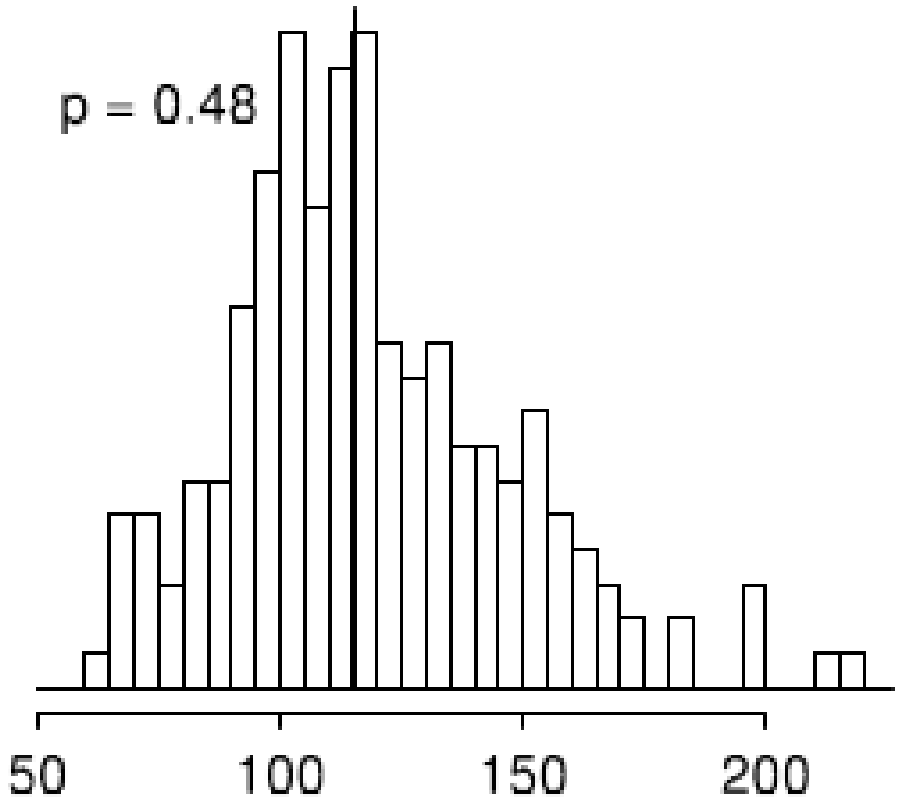
Example: Speed of Light Measurements

Let $T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$



Example: Speed of Light Measurements

Let $T(y) = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$



Bayesian sufficiency implies "predictive sufficiency"
$P(T(y^{\text{rep}}) > T(y) \mid y) = P(T(y^{\text{rep}}) > T(y) \mid \bar{y}, T(y))$

Example 2: Are our Bernoulli rvs Correlated?

From chapter 2:

1. $y_1, \ldots, y_n \mid \theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$
2. $\theta \sim \text{Uniform}(0, 1)$
3. $\theta \mid y \sim \text{Beta}(\sum_i y_i + 1, n - \sum_i y_i + 1)$

Example 2: Are our Bernoulli rvs Correlated?

From chapter 2:

1. $y_1, \ldots, y_n \mid \theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$
2. $\theta \sim \text{Uniform}(0, 1)$
3. $\theta \mid y \sim \text{Beta}(\sum_i y_i + 1, n - \sum_i y_i + 1)$

$$y = (1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0)$$

Example 2: Are our Bernoulli rvs Correlated?

From chapter 2:

1. $y_1, \ldots, y_n \mid \theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$
2. $\theta \sim \text{Uniform}(0, 1)$
3. $\theta \mid y \sim \text{Beta}(\sum_i y_i + 1, n - \sum_i y_i + 1)$

$$y = (1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0)$$

Let $T(y^{\text{rep}}) = \sum_{i=2}^{n} |y_i^{\text{rep}} - y_{i-1}^{\text{rep}}|$ be the number of switches.
Note that $T(y) = 3$

Example 2: Are our Bernoulli rvs Correlated?

Problem: $p(y^{\text{rep}} \mid y)$ is not closed-form.

For $i = 1, \ldots, 10^4$:

1. draw $\theta^i \sim p(\theta \mid y)$
2. draw $y^{i,\text{rep}} \sim p(y \mid \theta^i) = \prod_{j=1}^{n} \text{Bernoulli}(\theta)$
3. return $T(y^{\text{rep}})$

Example 2: Are our Bernoulli rvs Correlated?

Problem: $p(y^{\text{rep}} \mid y)$ is not closed-form.

For $i = 1, \ldots, 10^4$:

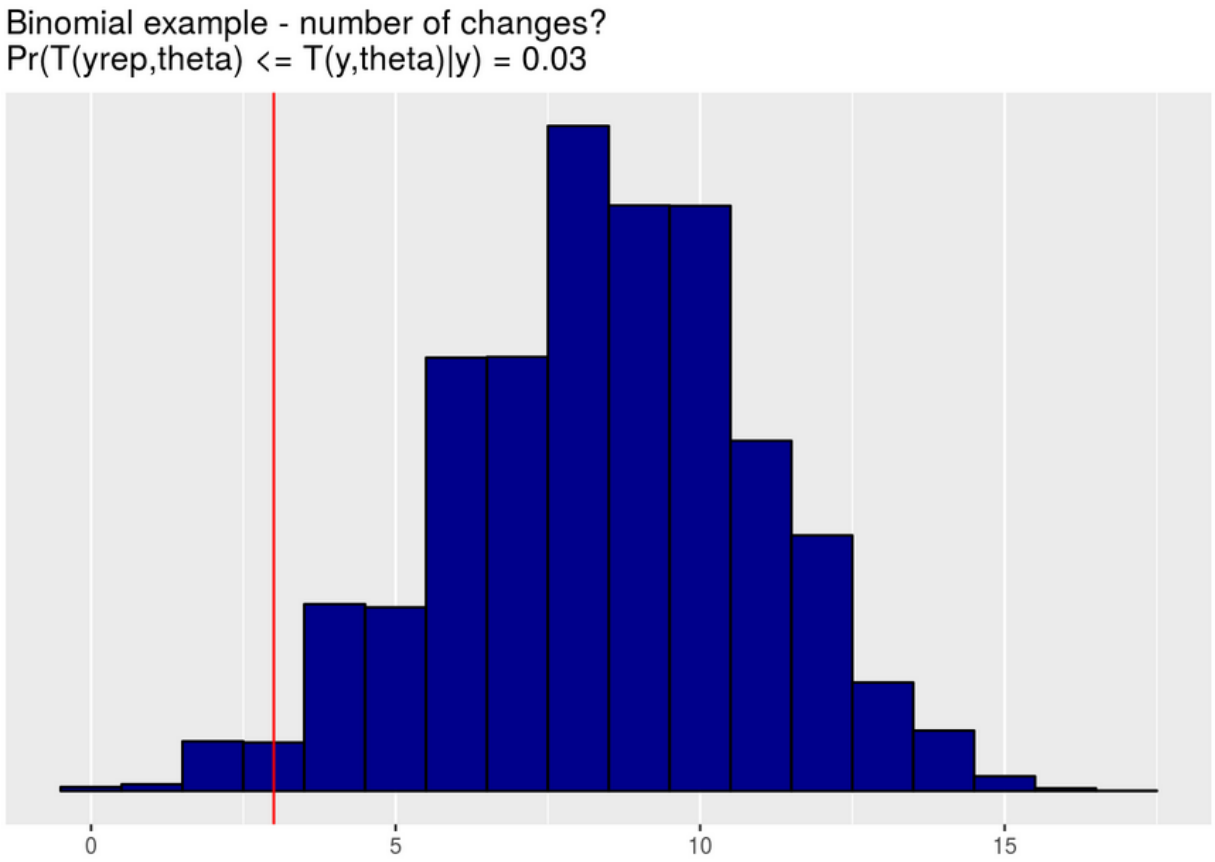
1. draw $\theta^i \sim p(\theta \mid y)$
2. draw $y^{i,\text{rep}} \sim p(y \mid \theta^i) = \prod_{j=1}^{n} \text{Bernoulli}(\theta)$
3. return $T(y^{\text{rep}})$

```
n <- length(y)
s <- sum(y)
rb <- function(s, n) {
  p <- rbeta(1, s+1, n-s+1)
  yr <- rbinom(n, 1, p)
  sum(diff(yr) != 0) + 0.0
}
Tyr <- data.frame(x = replicate(10000, rb(s, n)))
```

http://avehtari.github.io/BDA_R_demos/demos_ch6/demo6_2.html

Example 2: Are our Bernoulli rvs Correlated?

$$P(T(y^{\text{rep}}) > T(y) \mid y) \approx .97$$



p-values and u-values

If θ is perfectly estimated,

$$P(T(y^{\text{rep}}) > T(y) \mid y) = P(T(y^{\text{rep}}) > T(y) \mid \theta, y) \sim \text{Uniform}(0, 1)$$

This is related to the "CDF transformation." For $X, \tilde{X} \mid \theta \overset{\text{iid}}{\sim} F$:

$$F_X(X) = P(X \leq x \mid \tilde{X} = x, \theta) \sim \text{Uniform}(0, 1)$$

In our case, they are saying that

$$P(T(y^{\text{rep}}) > T(y) \mid y) = P(T(y^{\text{rep}}) > T(y) \mid \theta, y),$$

or in other words

$$P(T(y^{\text{rep}}) \leq T(y) \mid y) = P(T(y^{\text{rep}}) \leq T(y) \mid \theta, y)$$

p-values and u-values

Generally $P(T(y^{\text{rep}}, \theta) \leq T(y, \theta) \mid y)$ is "stochastically less variable" than uniform, which means p-value is more likely to be near 0.5 than near 0 or 1.

(This is left as an in-class discussion due Oct 14. Give your arguments through simulations.)

A u-value is defined as a function of the data y that has a $U(0, 1)$ sampling distribution. Therefore, Bayesian p-values are not always u-values.