Posterior sampling, prediction, and model checking

Math 315, Adam Loy

This example was taken, with permission, from Statistical Rethinking.

The `birthorder` data set contains data from a sample of 100 two-child families and was provided in *Statistical Rethinking*. The data set has two columns:

- `first`: the biological sex of the first-born child (male = 0, female = 1)
- `second`: the biological sex of the second-born child (male = 0, female = 1)

Use this data set to complete the following questions.

```
birthorder <- read.csv("https://raw.githubusercontent.com/aloy/math315-fall2019/master/data/birthorder.
```

1.

Using grid approximation, compute the posterior distribution for the probability of a birth being a girl. Assume a uniform prior probability. Which parameter value maximizes the posterior probability?

In order to use grid approximation to compute the posterior distribution for the probability of a birth being a boy we must do the following:

- define the grid of parameter values to consider
- compute the prior density at each parameter value on the grid
- compute the likelihood at each parameter value on the grid
- compute prior × likelihood for each value on the grid
- normalized (standardize) the posterior plausibilities to obtain posterior probabilities

Before beginning all of this computation, let's count the number of boy births out of 200 which is needed for the likelihood:

```
ngirls <- sum(birthorder$first, birthorder$second)
ngirls
```

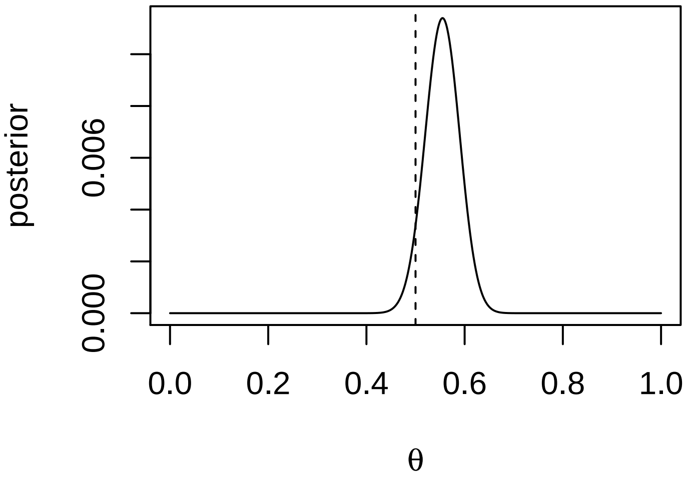
```
## [1] 111
```

You should find that there are 111 girls.

```
theta <- seq(0, 1, by = 0.001)
prior <- dunif(theta, 0, 1)
likelihood <- dbinom(ngirls, size = 200, prob = theta)
unstd.posterior <- prior * likelihood
posterior <- unstd.posterior / sum(unstd.posterior)
```

Note: It was not required to plot the posterior, but it's typically a nice idea to check your work:

```
plot(x = theta, y = posterior, type = "l", xlab = expression(theta))
abline(v = 0.5, lty = 2)
```



There are numerous ways to find the value of `p` that maximizes the posterior, one way is to use the `which.max()` function along with indexing:

```
theta[which.max(posterior)]
```

```
## [1] 0.555
```

2.

Using the sample function, draw 10,000 random parameter values from the posterior distribution you calculated above. Use these samples to estimate the 50%, 89%, and 97% highest posterior density intervals.

To sample 10,000 random parameter values from the posterior, we use the `sample()` command. Then it's just a matter of passing those samples to `hid()`:

```
theta_draws <- sample(theta, size = 1e4, prob = posterior, replace = TRUE)
library(HDInterval)
hdi(theta_draws, credMass = .5)
```

```
## lower upper
## 0.530 0.577
## attr(,"credMass")
## [1] 0.5
```

```
hdi(theta_draws, credMass = .89)
```

```
## lower upper
## 0.495 0.606
## attr(,"credMass")
## [1] 0.89
```

```
hdi(theta_draws, credMass = .97)
```

```
## lower upper
## 0.479 0.629
## attr(,"credMass")
## [1] 0.97
```

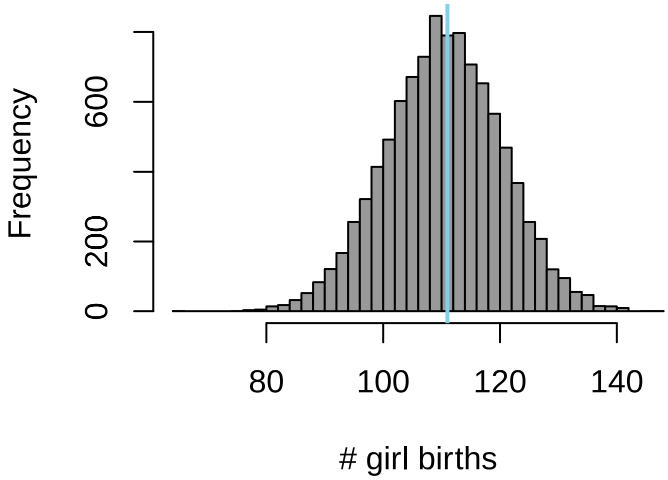
Each of these intervals is the narrowest range of parameter values that contains the specified probability mass.

##3. Use `rbinom(n = 10000, size = 200, prob = ___)` to simulate 10,000 replicates of 200 births. (Note: you will need to fill in the blank with the name of your sample from #2.) You should end up with

10,000 numbers, each one a count of girls out of 200 births. Compare the distribution of predicted numbers of girls to the actual count in the data (111 girls out of 200 births). There are many good ways to visualize the simulations, but density plots and histograms are probably the easiest. Does it look like the model fits the data well? That is, does the distribution of predictions include the actual observation as a central, likely outcome?

You can use the samples generated from the posterior in #2 to simulate 10-thousand samples of 200 predicted births. Note that this is the posterior predictive distribution.

```r
girl_sim <- rbinom(1e4, size = 200, prob = theta_draws)
hist(girl_sim, xlab = "# girl births", main = NULL, breaks = 50, col = "gray60")
abline(v = ngirls, col = "skyblue", lwd = 2)
```



The histogram is the simulated counts of female births, out of 200 maximum. The sky blue vertical line is the observed count in the data. It looks like the model does a good job of predicting the data. In a sense, this isn't surprising, because the model was fit to this exact aspect of the data, the total count of boys.

###4. *Now compare 10,000 counts of girls from 100 simulated first borns only to the number of girls in the first births. How does the model look in this light?*

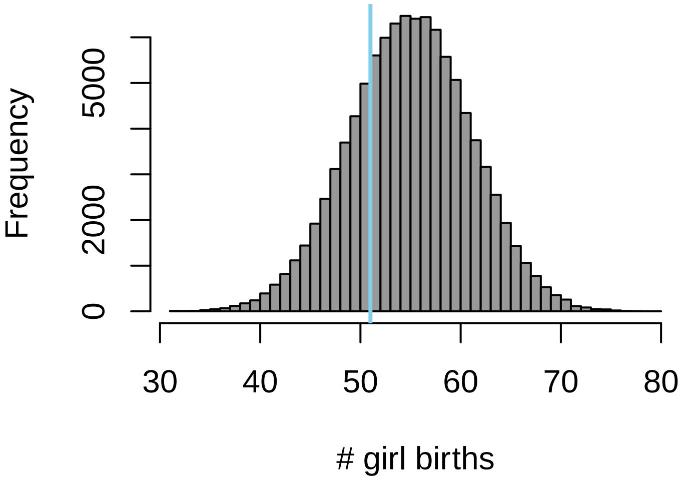
The model can be threatened a little more by asking if it can predict other, subtler aspects of the data. How does it do predicting just first borns?

The code to answer this question looks very similar, but now there are only 100 births to predict. The posterior samples are the same as before.

```r
birth1_sim <- rbinom(1e5, size = 100, prob = theta_draws)
```

The posterior predictive distribution is plotted below:

```r
hist(birth1_sim, xlab = "# girl births", main = NULL, breaks = 50, col = "gray60")
abline(v=sum(birthorder$first), col = "skyblue", lwd = 2)
```



There is still general agreement between the observed data and the predictions made by the model. (Notice that the observed data are not right on center now.) The frequency of girls among first borns is a little less than the model tends to predict. The model doesn't have a problem accounting for this variation though, as the sky blue line is well within the distribution of the simulated data.

5.

The model assumes that sex of first and second births are independent. To check this assumption, focus now on second births that followed male first borns. Compare 10,000 simulated counts of girls to only those second births that followed boys. To do this correctly, you need to count the number of first borns who were boys and simulate that many births, 10,000 times. Compare the counts of boys in your simulations to the actual observed count of girls following boys. How does the model look in this light? Any guesses what is going on in these data?

Now an even harder test of the model. How does it do predicting second births that follow girl births? The only tricky part of this is counting up boys born after girls. There are numerous ways to do this. In my opinion, the easiest approach is to create a data frame and use the `count()` function:

```r
with(birthorder, table(first, second))
```

```
##      second
## first  0  1
##     0 10 39
##     1 30 21
```

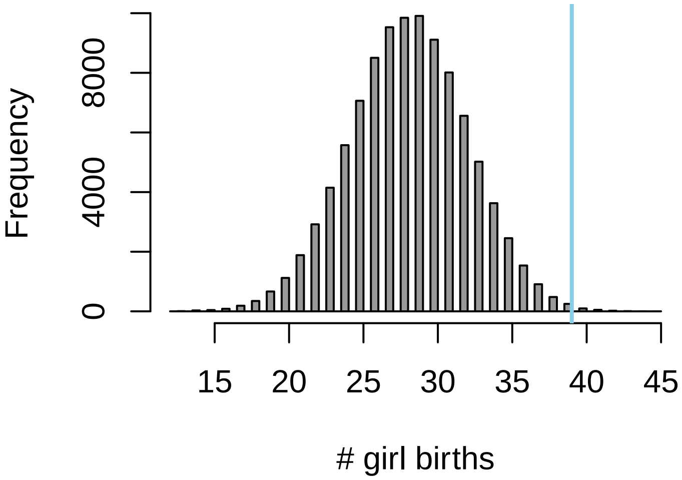
We find that there are 39 girls born after boys in our data set, and 51 first-born girls.

Now, we create the posterior predictive distribution for the number of girls following boy births:

```r
b1_g2_sim <- rbinom(1e5, size = 51, prob = theta_draws)
```

The resulting distribution is plotted below:

```r
hist(b1_g2_sim, xlab = "# girl births", main = NULL, breaks = 50, col = "gray60")
abline(v = 39, col = "skyblue", lwd = 2)
```



These predictions are pretty terrible. The observed number of girls who follow boys is far in excess of what the model predicts. This might indicate that the first and second births are not independent.