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.csv")
```

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?
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.
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?
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?
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?