BINOMIAL POPULATION

For this worksheet, we will deal with a binomial population. When an event is binary, that means that each trial can have only two possible outcomes, either a success or a failure. The true proportion of successes, p, can be estimated using statistical practices. Below, you will see how a uniform prior distribution for the possible values of p is updated through the collection of relevant data.

THE PRIOR DISTRIBUTION

1. First, start by opening the Discrete Bayesian Statistics Applet for Non-Uniform Prior Distributions. Then on the left side panel, select the button labelled “Binomial” as the population distribution we will be picking from.

2. Next, click the “Choose File” button in the top left to pick a prior distribution. After clicking the button, pick the CSV file named “coinFlipPrior.csv,” which you should have downloaded and saved prior to beginning this worksheet. Observe both the plot of the prior distribution, and the table at the bottom of the page. Describe the shape of the prior distribution.

Big Idea: As you have learned before, a uniform prior distribution can be useful in Bayesian statistics when you don’t have much knowledge of the true probability of success beyond a simple range of values. We are not simply limited to uniform prior distributions though, especially if we have prior intuition about what the true proportion of successes is.

COIN FLIPPING EXAMPLE

3. Now, suppose this distribution is our prior distribution for our beliefs of the possible true proportion of heads for a coin that we have. We want to estimate the true proportion of heads for this coin using Bayesian statistics.

4. Imagine that we toss this coin 10 times. In these 10 tosses, we get 6 heads and 4 tails. Input this data into the correct places in the side panel. Observe the posterior distribution after we have this new added knowledge about our coin. What values appear to be higher now than in the prior distribution? Explain how the data changed our beliefs about the coin.

Big Idea: The posterior distribution is obtained from a combination the prior distribution and the data that was collected. The proportions of success that are closer to the sample proportion of success will increase in probability from the prior to the posterior probabilities.

5. Report the 95% credible interval for the true proportion of heads for this coin after our first ten flips.

   ____

6. Instead, suppose that we flipped the coin 100 times, but got the same sample proportion of successes. Change the input values to have 60 successes and 40 failures. Observe the posterior distribution. What are the highest probable values for the true proportion of heads of this coin? What do you notice about the proportions closer to the sample success rate versus those farther away?

7. Report the 95% credible interval after we have flipped the coin 100 times. How does this interval compare to the interval you found in exercise 5?

8. Once again, increase the number of coin flips that we have performed. Suppose, on this set of trials, we obtained 300 heads and 200 tails. This is also a sample success rate of 0.60, but a sample size 5 times the size as in exercise 6. Observe the posterior distribution: report the highest values, and comment on how the values closer to the sample success rate compare to those farther away.

9. Once again, report the 95% credible interval of the true proportion of heads for our coin. How does this interval compare to those found in exercises 5 and 7?

10. From our experimentation with different sample sizes above, what seems to have a larger effect on the posterior distribution, the sample size or the prior distribution? Explain your reasoning. (*Hint: Think back to the when we used a uniform prior distribution, and how the posterior distribution was affected by the sample sizes then.*)

Big Idea: When performing discrete Bayesian statistics, the prior distribution has a larger effect on the posterior distribution than the sample size. It takes a very large sample size to "pull" the posterior distribution toward the results found in the sample. Thus, if we have a general idea of which values are most likely for the population parameter to be, we need a lot of data against our presumptions to sway us from the prior distribution. This is a place where we must be cautious because we do not want to affect the posterior distribution with the prior distribution unless we are sure of our beliefs prior to collecting data.

NON-SYMMETRIC PRIOR DISTRIBUTION

In many cases, symmetric prior distributions are very good to find the true population parameter, but we are not limited to just symmetric prior distributions. Instead of using a uniform or symmetric non-uniform prior distribution, the example below will use a non-symmetric prior distribution. We will discover how the prior distribution affects our beliefs regarding the posterior distribution and true population parameter.

11. Click the "Choose File" button and find the CSV file named "safariTripPrior.csv." After selecting this file, observe the prior distribution. Use both the graph at the top of the page and the table at the bottom to confirm that the prior distribution is non-symmetric.

12. For the following problems, suppose this distribution represents the possible true proportions of seeing a lion on a one-day safari trip and their probabilities. From the prior distribution, what would you say is the most likely true probability of seeing a lion on a one-day safari trip?

____

13. Now, imagine that we go out on a safari trip on 15 seemingly random days. On those 15 days, we spot a lion of 5 of those days. How does this data affect the posterior distribution? Input the data into the applet and comment about the posterior distribution.

14. Report the 95% credible interval of the probability of seeing a lion on a one-day safari trip.

15. Keep increasing the sample size in increments of 15, each time adding 5 successes and 10 failures. This will be the same sample proportion as in the first set of 15. Watch what happens to the posterior distribution. At what sample size does the posterior distribution become somewhat symmetrical, whereas it has only one peak?

16. From your answer in exercise 15, what does this mean about the strength of the sample size versus the prior distribution in affecting the posterior distribution? Explain your reasoning.

17. Lastly, report the 95% credible interval after performing enough sample trials to have a somewhat symmetrical posterior distribution (the number of trials found in exercise 15).

Big Idea: When we have a non-symmetrical prior distribution of the possible values for the true population proportion of successes, once again, the prior distribution has more influence on the posterior than the sample size. Although a non-symmetric posterior distribution may seem unnatural, it is actually very useful. When performing Bayesian statistics starting with a symmetric prior distribution, after collecting data and creating the posterior distribution, the posterior distribution may become non-symmetric. You may have noticed this in exercises 4 and 6 above. We can then use that posterior as a new prior distribution for new data. Non-symmetric prior distributions are used less when first starting experiments, but are used in powerful ways throughout the entirety of the data collection and the posterior distribution updating process.

I encourage you to create your own prior distributions and experiment with how the posterior distribution changes based on different sample inputs. This will help you to grasp how we start with a prior belief that is updated through the collection of more and more relevant data.