BINOMIAL POPULATION

For this worksheet, we will deal with a binomial population. This means that there are we will be keeping track of events, and at each trial, the event can either be a success or a failure, thus having only two possible outcomes and being binary. Below, you will see how a uniform prior distribution for the possible values of p, the true proportion of successes, is updated through the collection of more data.

OVERVIEW OF THE BASIC IDEAS

1. First, start by opening the Discrete Bayesian Statistics Applet. Click the button beside "Binomial" under the "Population Distribution to Pick From" header in the side panel, if it is not already clicked.

2. Set the Uniform Prior Distribution to go from a lower value of 0 to an upper value of 1. Also, set the number of values to be 21. This will make the possible values for the true proportions be in the set, {0, 0.05, 0.10, … , 0.95, 1}.

3. Looking at the plot in the top left, what do you notice about the probabilities of each proportion in the prior distribution? What type of distribution is this called?

   ____

4. Set the "True Probability of Success" to 0.60, and note how the true Population Distribution for 10 trials changes (plot in the top right). At what number of successes is the peak in the graph? What proportion is this of the "Number of Trials"?

5. Now, looking at the Prior Distribution plot again, approximate the probability that the true proportion of successes is 0.60, based on our uniform prior distribution. Report that value. How does it relate to the prior probability of other proportions of success?

   Big Idea: Under the Uniform Prior Distribution, we assume that all proportions of success are equally likely.

6. Next, assume we are going to take a sample of 13 data points. Change the "Number of Trials" input to 13 (keeping the true proportion of successes at 0.60) and notice how the "True Population Distribution" plot in the top right changes. What is the most likely number of successes for 13 trials? About what probability do we have of seeing that number of successes?

   ____

7. Suppose, in our sample, we received 9 successes and 4 failures. Plug in these numbers into the appropriate inputs on the bottom left of the screen. Which plot(s) change after you input these values?

   ____

8. Now, looking at the Prior vs Posterior Distributions plot, how did the addition of new data (the 9 successes and 4 failures) update what we believe the true proportion of successes is? Is it still uniform? Where is the posterior distribution centered?

9. Still observing the Prior vs Posterior Distribution Plot, what range of proportion of successes do we now believe are more probable than in the prior distribution? (*Hint: At what proportions of success are the probabilities higher than in the uniform prior distribution?*)

   ____

10. After collecting data and updated our beliefs, approximately what would we now say that the probability of the true proportion of successes being 0.60 is? How does this value compare to the value you found in exercise 5?

11. Lastly, we must look at a credible interval of what the true proportion of success can be, based on our sample. Use the buttons labelled "Size of Credible Interval," and the displayed graph on the bottom right and text to fill in the blanks in the statement below.

    There is a 95% probability that the true proportion of successes is between _________ and ___________ .

    Big Idea: In Bayesian statistics, when estimating what the true proportion of successes is for a binomial population, we start with a prior distribution, and update that prior distribution based on new data to get a better estimate of probabilities, known as the posterior distribution. By adding up the posterior distributions' probabilities from highest to lowest until it goes over a threshold, we can also make claims that there is a certain chance that the true parameter value falls between two values.

BATTING AVERAGE EXAMPLE

Suppose we want to estimate the true batting average of the Houston Astros second baseman, Jose Altuve. This would be estimating the true proportion of hits out of at-bats that he gets. At the end of the 2015 season, Altuve got a hit in .313 of his at-bats (had a batting average of .313). Set this to be the true probability of success in the applet.
Also for this example, start by clearing the Number of Successes and Number of Failures in a sample, and then follow the directions below.

12. First, we must set up a prior distribution of the batting averages that we think can be possible for Altuve. A good prior estimate is that his true batting average is uniform between .200 and .400. Set up the prior distribution to be the discrete uniform distribution from .200 to .400 by .001 ({.200,.201,.202,...,.399,.400}). (*Hint: You must choose the appropriate "Number of Values" to set the prior distribution to be by .001*)

13. Now, suppose that we watch Jose Altuve play 5 games live, which can be assumed to be as good as random. In those games, he gets 18 at-bats. Change the Number of Trials to 18 and observe the plot on the top right. What are the two most probable number of hits that Altuve would get in 18 at-bats?

    ____

14. In those 5 games, suppose Altuve got 6 hits in his 18 at-bats. Is this an unexpected number of hits? About what is the probability he got exactly 6 hits in those 18 "random" at-bats? Would this value surprise you?

    ____

15. Enter the appropriate number of successes and failures into the inputs for our sample in the applet. What can you see now about the relationship between the prior and posterior distributions? What batting averages do we now believe to have a lower probability of being the true batting average than in the prior distribution?

16. At approximately what proportion of successes (hits) does the posterior distribution reach its maximum value?

    ____

17. After the sample, did the likelihood of Altuve's true batting average being .313 go up or down? Is this what you expected? Explain.

18. Create a 90% Credible Interval for the true batting average (proportion of successes for Jose Altuve). Is his true batting average, 0.313, in our interval? Report the upper and lower values of the credible interval.

    ____________ , ____________

    Big Idea: After a sample, we can update our beliefs about the parameter we are estimating, thus transforming our prior distribution into our posterior distribution. The posterior distribution will have a maximum at the proportion of successes found in the sample. The updated probability of that proportion being the true value of the parameter though is also dependent upon the prior distribution, such as the width and the number of values, which you will discover in the next section of this worksheet.

19. Now, what if we didn't have as good of an understanding of what Altuve's possible batting average could have been prior to watching him play? In this case, change the lower value of the prior to 0 and the upper value to 1, and the number of values to 1001. (This allows us to capture the entire set: {.000,.001,.002,...,.999,1.000})

20. Now, given the new prior, but the same information as before (6 hits out of 18 at-bats), what differences do you see in the posterior distribution compared to the posterior distribution when the limit of values was narrower, as above?

21. Approximately, at what batting averages do we now have a lower probability of being the true batting average than in the prior? How does this answer differ from that found in exercise 14?

22. Also create a 90% Credible Interval and report the upper and lower values of this interval. Is this interval wider or narrower than what you found in exercise 18? Why are these two intervals different?

23. Now, suppose we were able to see 10 times as many at-bats, and actually saw Altuve get 180 at-bats, in which he got 60 hits. (Note that this is the same proportion of hits as in our smaller sample previously.) Change the Number of Trials to 180 and roughly

estimate the probability of Altuve getting 60 hits in 180 at-bats if his true batting average is in fact 0.313. Report this value.

____

24. Now update the number of successes and failures in the sample to appropriately represent our new observation of 180 at-bats. Examine the Prior vs Posterior Distributions plot and estimate what batting averages now have a lower probability of being the true batting average than in the prior? How do these values compare to the intervals found in exercise 19?

25. Lastly, create a 90% credible interval. Based on this interval would you accept someone telling you that Jose Altuve's true batting average is 0.313? What evidence do you have, for or against this claim?

Big Idea: When our prior distribution is more broad, with a wider interval that it covers, the posterior distribution will cover a wider interval and the interval of proportion of successes that increase in probability from the prior will be wider, comparatively. Another relationship that can be found, is that the interval of proportion of successes whereas the probability of being the true proportion increases is smaller when the sample size is greater. This supports the basic understanding of statistics that we get a more accurate estimate of a parameter with a larger sample size. Thus in Bayesian statistics we have an understanding that the probability of certain values increases, narrowing the values we think could be the parameter, making the estimate more accurate.

Explore more different inputs and types of changes that can occur based on the uniform prior distribution for binomial data. Take into account how the posterior distribution and how the credible interval intervals are affected.