STATISTICS TEACHER
SUPPORTING THE TEACHING AND LEARNING OF STATISTICS
ASA NCTM
STatistics Education Web
STEW
online journal of K-12 statistics lesson plans

Bayesian Inference For Proportion Of Water On Earth

Jason Cleveland, Jacksonville State University

Published: March 2023

Overview of Lesson

In this lesson students will practice data collection through simulation, performing estimation, and providing support for or against a claim. The focus will be on what proportion of earth is covered in water. To accomplish this, the students will employ a simple technique for binary data and use web-apps that visually display some of the harder concepts they can learn later.

Type of Data

- One categorical variable
- Data generated or collected as a class

Learning Objectives

After completing this activity, students will be able to:

- Describe how to collect data through simulation
- Develop estimates for the proposed question of interest
- Explain introductory Bayesian reasoning

Audience

- Students would benefit with some knowledge of probability, but conceptual understanding of means and proportions should suffice. Grades 9 - 12+
- *Prerequisites:* Prior to this lesson, students should have experience with computing means and proportions, fractions, and conceptual understanding of weighted averages.

Time Required

One class period (50 minutes) and 15 minutes at the beginning of the next class period if the teacher decides to incorporate the suggested problem at the end.

Technology and Other Materials

- *Technology:*
    - o Data collection tool – RANDOM.ORG web applet
    - o Statistical analysis tool – Matt Bogner's Probability Distribution web applet

  (Links for both are included in the "How to use _____" document for each tool)
- Student copies of the Bayesian Inference Worksheet
- Teacher reference documents of Weight Average and Beta Distribution Briefly Explained if needed

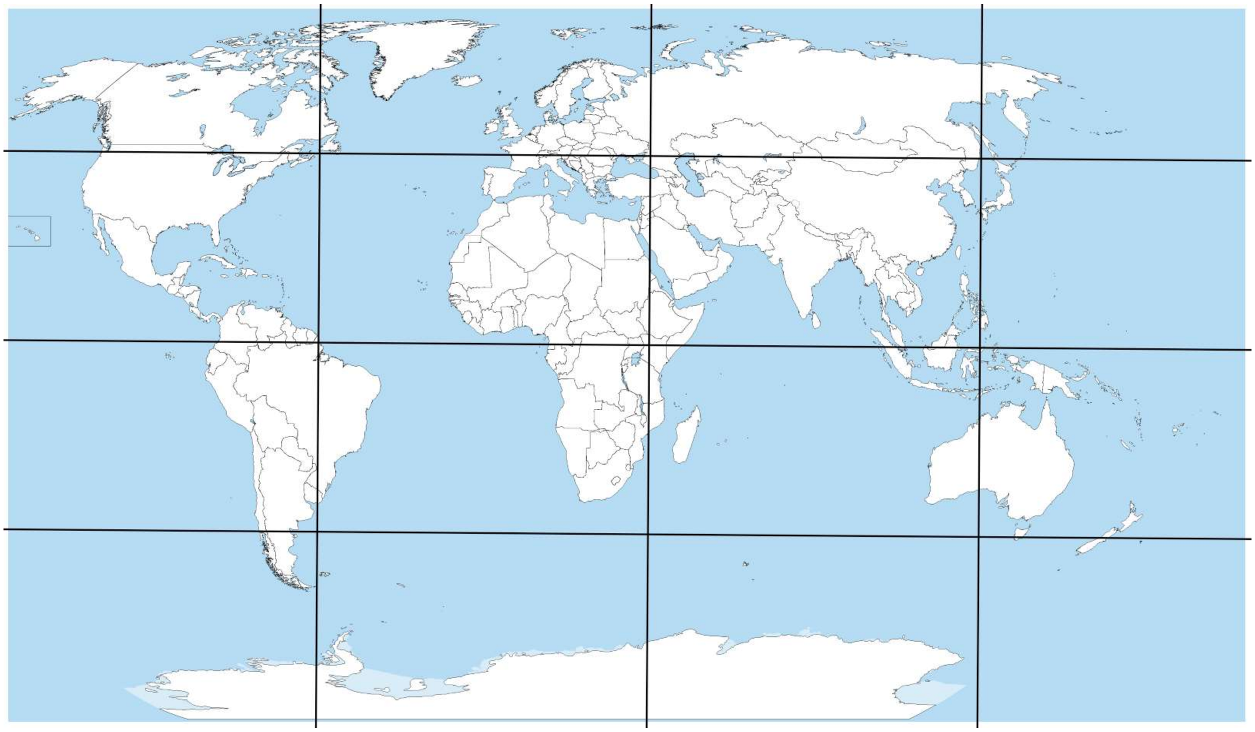
Lesson Plan

This lesson plan aims to showcase how a simple methodology of combining information can yield inference.

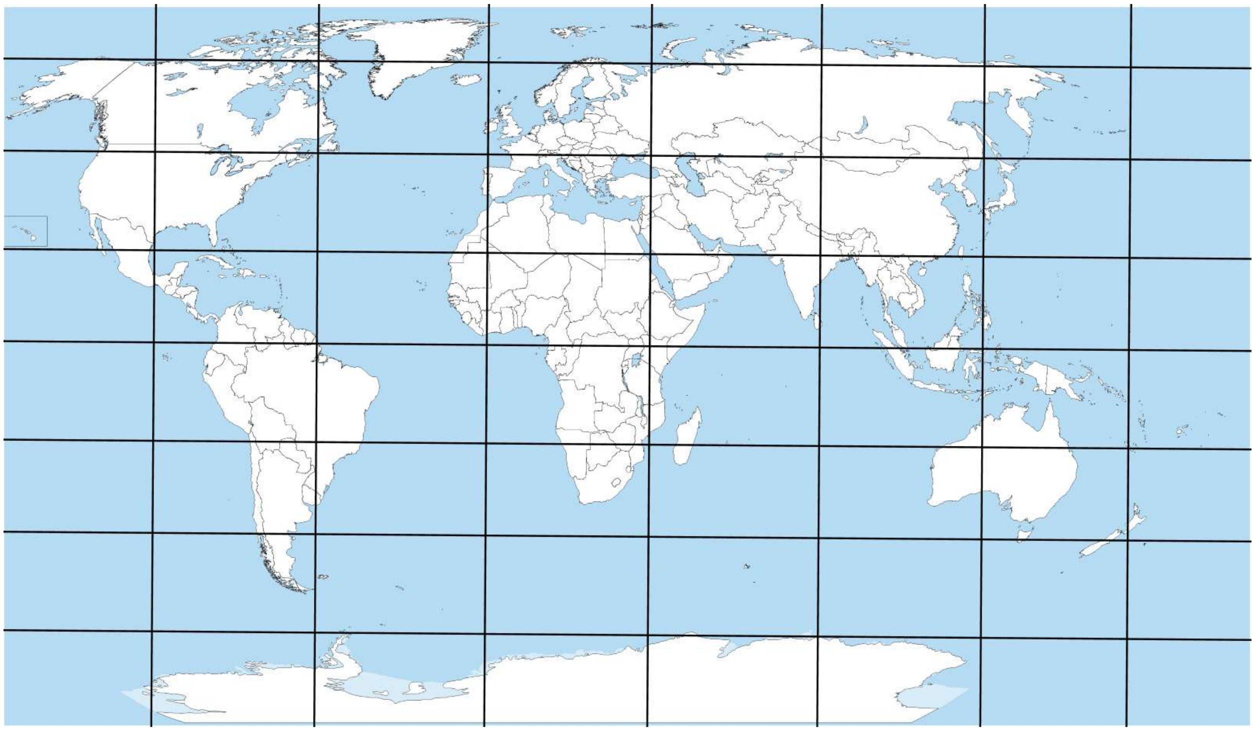
Suppose we wanted to show our students how to infer what proportion of water covers the earth. We could investigate this by spinning a globe, randomly placing our fingers on a spot as we stop the spinning globe and repeat this process many times. However, that would be quite time consuming. Also, we might not be prone to exactly random locations as the globe since it might have an underside or topside with accessories that would get in our way of choosing some of those spots. What are we to do?

Well, we can use a web-app that informs us that it randomly selects a location on earth via latitude and longitude. But now a new skeptical thought arises "can we trust that web-app?"

How about a grid overlaying a map of earth? Could we use this? Sure! However, this can lead to the same line of questioning as before "can we trust the map to be an accurate representation?" Also, how precise is our grid. To demonstrate this, we will give a few visual examples. Take note of the four-by-four grid presented below.



Could we accurately estimate the proportion of water on earth from this? I do not think we could. Well, what about an eight-by-eight?



I imagine this would be like the four-by-four grid. In fact, we could make this grid larger and larger and have the same issue of a time intensive process just like with spinning the globe.

This lesson plan will focus on combing two sources of uncertainty like those presented above to provide inference. We will take some information from a grid and some from the random observations that the web-app produces for us (which takes the place of having to physically spin a globe ourselves).

Formulate a Question

There are two main questions we want to formulate for this lesson plan.

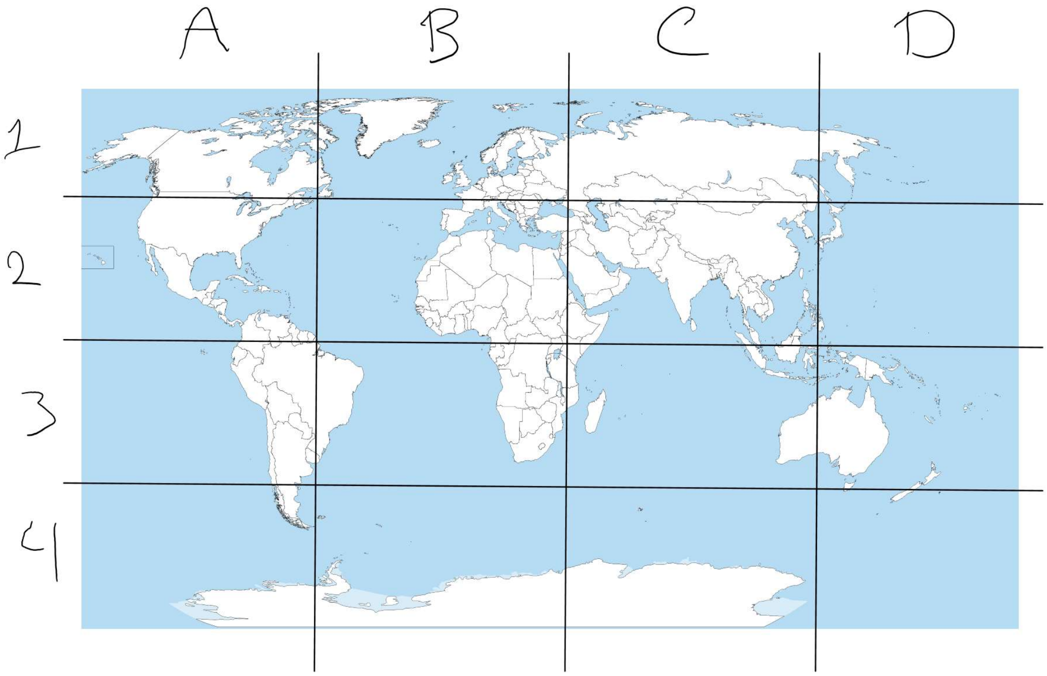
1. What proportion of earth is covered in water?
2. Can we attempt to test certain claims about the proportion of water on earth? For example, "given our data and modeling assumptions, is it reasonable to believe that the proportion of water on earth is less than 50%?"

For formulating question 1:

If the teacher is looking to engage students with reasons for learning Bayesian reasoning, then there are a few real-life examples that can be made as interesting as the teacher wants them to be. Three such examples will be presented here. During the Second World War, a form of Bayesian reasoning was used to help break the Enigma machine. In 1968, another form of Bayesian reasoning was utilized to help locate the nuclear-powered submarine USS Scorpion. In 1988, the same Bayesian reasoning was used to find SS Central America which contained a lot of gold but sank because of a hurricane. The teacher can always spice these up more, but honestly espionage, heroic journeys to recover possible dangerous but valuable items, and literal treasure hunting are all already pretty spiced up!

The teacher should start off by telling the students why we will not spin a globe numerous times nor create a highly segmented grid (either exactly as presented above or in their own words). After doing this, inform the students that they will be using a web-app to collect simulated locations on earth (instead of spinning a globe). Also let them know that they will combine their sample information with information they get from a four-by-four grid approximation. To assist the teacher with this, they should follow through with what is about to be presented.

My suggestion for starting is to present the four-by-four grid overlaying the map (provided above or below) and give each grid location a label (e.g. columns have alphabetical labels and rows have numerical labels). Then the teacher and students should go through the process of determining if each referenced grid is mostly land or mostly water (recall that a more refined grid could help with this but then we could always argue that another refined grid would do better). An example of what might be concluded upon by the teacher and students is provided below.

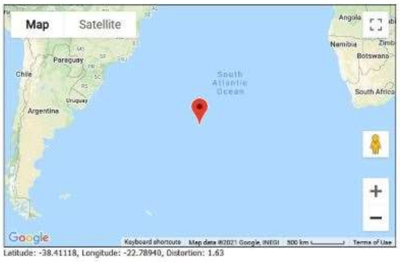


A1 - L
A2 - L
A3 - W
A4 - W
B1 - W
B2 - L
B3 - W
B4 - W
C1 - L
C2 - L
C3 - W
C4 - W
D1 - W
D2 - W
D3 - W
D4 - W

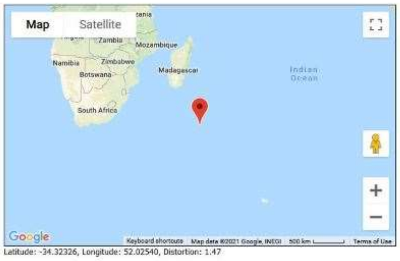
Keep the counts of land and water (labeled in the above as L and W, respectively) from the grid approximation as they will be used later. Next, pull up the data collection web-app, RANDOM.org (https://www.random.org/geographic-coordinates/) (see How to use RANDOM.org document).

Inform the students that this will play the role of randomly spinning the globe. The teacher should inform them of how to use the web-app. The teacher should collect enough observations to outline how to combine both results. An example of a few observations from the web-app are provided below.

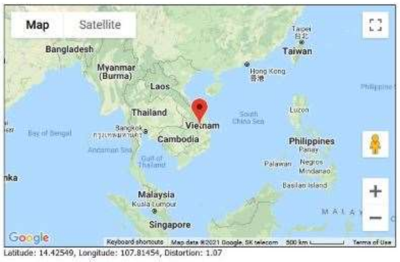
For example:



Latitude: -38.41118, Longitude: -22.78940, Distortion: 1.63



Latitude: -34.32326, Longitude: 52.02540, Distortion: 1.47



Latitude: 14.42549, Longitude: 107.81454, Distortion: 1.07

The teacher should record the number of observed lands and waters, sum them up for the total number of observations, and obtain the proportion for water by taking the count of water out of the total number of observations. The teacher should communicate this structure to the students by making something like the chart provided below.

Based off ten random simulations I ran; I would do the following:

	Count of Land	**Count of Water**	**Total Observations**	**Proportion Estimate**
Simulation	3	7	10	7/10

With that in mind, go back and get the counts from the grid approach. The teacher should alter the tabular display to provide the counts for both approaches.

Based off the above student response, we would do the following:

	Count of Land	**Count of Water**	**Total Observations**	**Proportion Estimate**
Grid	5	11	16	11/16
Simulation	3	7	10	7/10

The last step necessary to provide the student with what they need to analyze and interpret results for question one above is combining these results.

Following along with the example provided:

	Count of Land	**Count of Water**	**Total Observations**	**Proportion Estimate**
Grid	5	11	16	11/16
Simulation	3	7	10	7/10
Combined	8	18	26	18/26

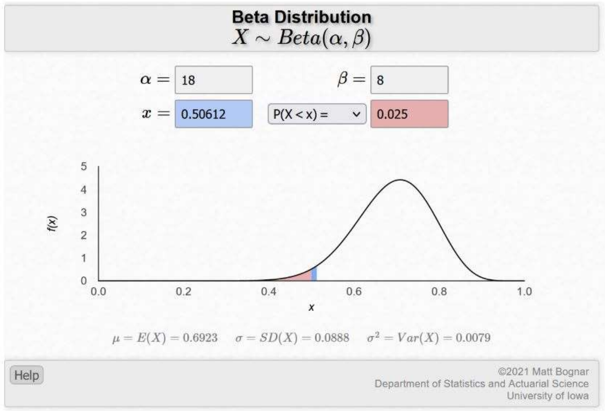
All that happened in the table above was combining results through addition for the Combined row. That is, add the count of lands, water, and total observations together for Grid and Simulation rows. The estimate is still computed in the same manner as the other rows. With all this information conveyed to the students, they should be able to provide a point estimate for question one. However, we want to provide some region of plausible values. Therefore, one last step is necessary for creating an interval estimate.

The teacher should read through the Beta Distribution Briefly Explained document beforehand. This way they can provide some of the basics for conceptually understanding parameter values in the Beta distribution.

To provide an interval estimate, the teacher should demonstrate Matt Bognar's web-app to them. They can even use the visualization of the web-app to demonstrate what changing α and β parameters do to the distribution. Instructions for how to navigate this web-app are also provided for this lesson plan (see How to use Matt Bognar's Probability Distribution Web Applet document).

The teacher should inform the students of how to obtain a 95% interval estimate (the idea that the middle 95% of the distribution is covered or that we obtain the lower bound by looking for the 2.5% boundary and the upper bound by looking for the 97.5% boundary).

Using the numbers presented in the tabular display above, the teacher should have an interval along the lines of $(0.506\,, 0.851)$. See image below for visual representation.

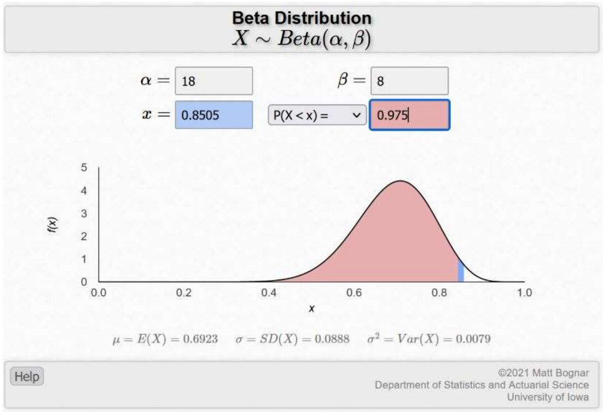


Beta Distribution
$X \sim Beta(\alpha, \beta)$

$\alpha =$ 18 $\beta =$ 8
$x =$ 0.50612 P(X < x) = 0.025

$\mu = E(X) = 0.6923$ $\sigma = SD(X) = 0.0888$ $\sigma^2 = Var(X) = 0.0079$

©2021 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa



Beta Distribution
$X \sim Beta(\alpha, \beta)$

$\alpha =$ 18 $\beta =$ 8
$x =$ 0.8505 P(X < x) = 0.975

$\mu = E(X) = 0.6923$ $\sigma = SD(X) = 0.0888$ $\sigma^2 = Var(X) = 0.0079$

©2021 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

The teacher can discuss how with only 26 combined data points the inference should be somewhat uncertain. This wide interval estimate should demonstrate that. The teacher can also discuss ways to remedy this issue, mainly making the grid more refined, collecting more simulated data, or both in tandem. The teacher can also inform the students that they will be doing the second option (collecting more simulated data) to improve their inferences in this assignment.

For formulating question 2:
Now we will focus on the second question. The teacher can choose a different region of interest (other than less than 50%), but I will continue with it here to provide the necessary groundwork. The teacher should familiarize themselves with the Weighted Average Reference document provided for this lesson plan and demonstrate how to compute this to the students. The reason this is important is for us, or our audience we present our findings to, to know how much of the inference comes from the grid approach and how much comes from the sampling approach.

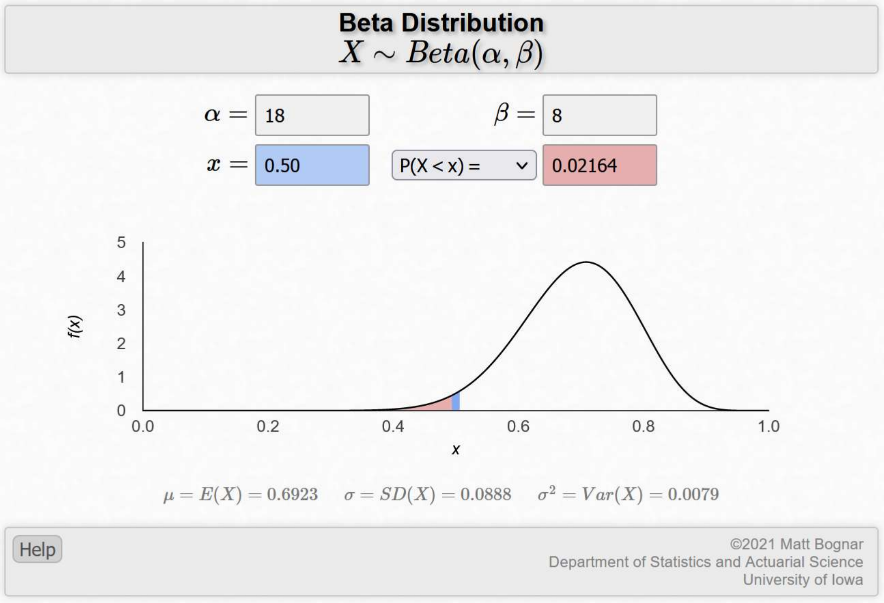
The teacher should make sure that the students understand how to compute how much weight is given to the grid and how much weight is given to the simulated data. After the teacher has provided this explanation, they should demonstrate Matt Bognar's web-app. Let us use our example from before to determine the support for the claim. That is, we are going to assess the probability of water being less than 50% given our combined data and model assumptions. To remind you, the example table is provided again here:

	Count of Land	Count of Water	Total Observations	Proportion Estimate
Grid	5	11	16	11/16
Simulation	3	7	10	7/10
Combined	8	18	26	18/26

In the web-app, we will be asked for a few values. Those values are as follows: $\alpha, \beta, x,$ and $P(X > x)$ [or $P(X < x)$ depending on which one we choose in the drop-down menu]. We have some of these values and those will allow the web-app to compute the ones we do not have.

Supply the α space with the Combined count of water value in your table, the β space with the Combined count of land value in your table, choose the $P(X < x)$ option since we want the probability that it is less than 50% for our question, and provide the x space with the value of 0.50 as the decimal representation of 50%.

You should have something like the following image after entering these in correctly.



Beta Distribution

$X \sim Beta(\alpha, \beta)$

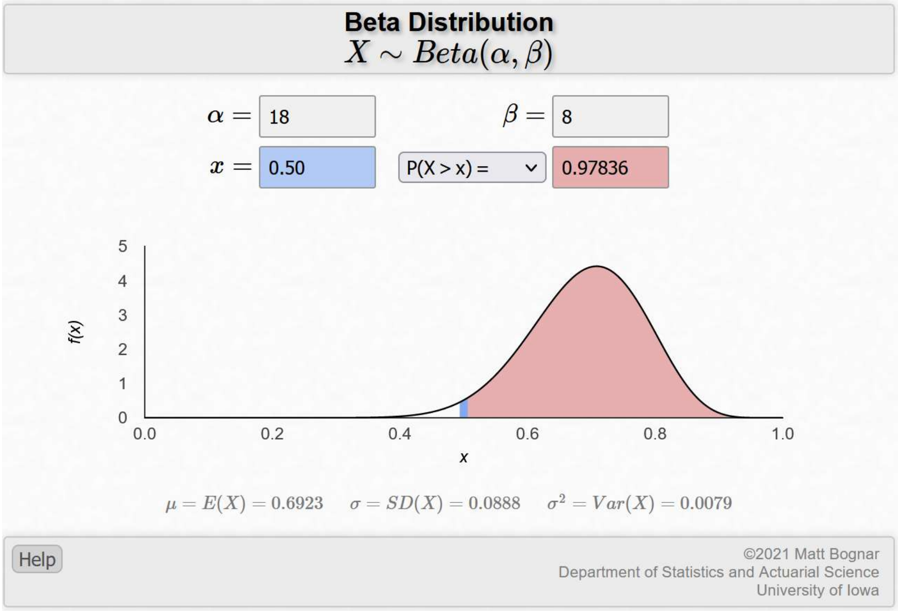
$\alpha =$ 18 $\beta =$ 8

$x =$ 0.50 P(X < x) = 0.02164

$\mu = E(X) = 0.6923$ $\sigma = SD(X) = 0.0888$ $\sigma^2 = Var(X) = 0.0079$

©2021 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

For comparison's sake, we can flip the probability option in the drop-down menu to see what the probability of water on earth being greater than 50% is. An image of that is provided below:



Beta Distribution

$X \sim Beta(\alpha, \beta)$

$\alpha =$ 18 $\beta =$ 8

$x =$ 0.50 P(X > x) = 0.97836

$\mu = E(X) = 0.6923$ $\sigma = SD(X) = 0.0888$ $\sigma^2 = Var(X) = 0.0079$

Help

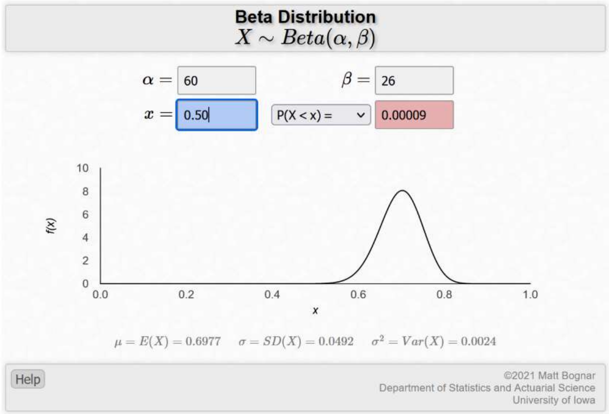
©2021 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

The teacher can spend some time discussing how both probabilities are not certain but that, based off what our combined data and modeling assumptions were, it seems more likely the earth has more than 50% water on it (in comparison to less than 50% water on it). Let the students talk through these ideas, which claim seems more likely, for a little bit and then direct them towards an important concept they should see in the assignment. How can we make reasonable inferences?

The teacher can create a hypothetical where they collected numerous data values such that their table looks like the one below:

	Count of Land	**Count of Water**	**Total Observations**	**Proportion Estimate**
Grid	5	11	16	11/16
Simulation	21	49	70	49/70
Combined	26	60	86	60/86

Point out through the web-app that this leads to the following probabilities:



Beta Distribution

$X \sim Beta(\alpha, \beta)$

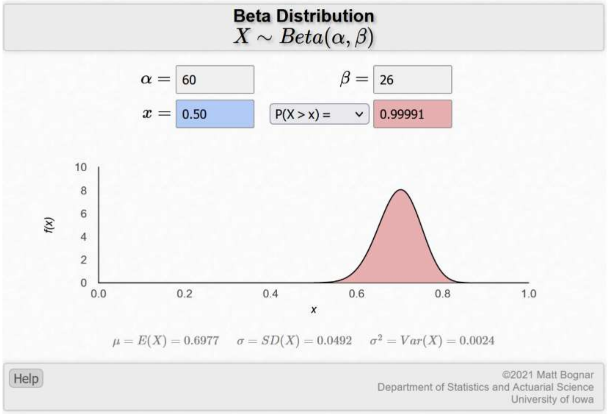
$\alpha =$ 60 $\beta =$ 26

$x =$ 0.50 P(X < x) = 0.00009

$\mu = E(X) = 0.6977$ $\sigma = SD(X) = 0.0492$ $\sigma^2 = Var(X) = 0.0024$

Help

©2021 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa



Beta Distribution

$X \sim Beta(\alpha, \beta)$

$\alpha =$ 60 $\beta =$ 26

$x =$ 0.50 P(X > x) = 0.99991

$\mu = E(X) = 0.6977$ $\sigma = SD(X) = 0.0492$ $\sigma^2 = Var(X) = 0.0024$

Help

©2021 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

The choice of which hypothesis is more likely seems obvious when we compare these two probabilities. Discuss how these two tables have vastly different probabilities to support the claim of more than 50% water on earth (97.836% compared to 99.991%).

At this point the teacher can reveal to the students that they have performed a Bayesian analysis. They have incorporated prior information, the grid from the map approach which did not take simulated data but did still convey information necessary for our inference, and information from the simulated data together. In fact, the Bayesian framework allows us to balance between these two forms of information precisely through our weighted average approach. Want to use a more refined grid because you do not trust the web-app's 'randomness'? Then refine the grid and put more weight on that part of the model! Want to use more random simulated data to control for the possible imprecise map that you do not trust? Then take more samples and put more weight on that part of the model!

If the teacher wishes to point out the main themes of Bayesian analysis, then they should follow through this paragraph. The grid approach using the segmented map would be what we call the "prior." It is information that does not come from the sample itself but is informed through a variety of methods (such as prior data, subject matter expertise, subjective beliefs, etc.). Here though, we specifically used a map of the world with the hopes that it accurately reflects the true proportion of what we wanted to estimate. The data plays the same role the teacher might be familiar with in the orthodox approach. The combined row plays a new role that is called the "posterior." It is, as we have labeled it here, a combination of the prior and the data meant to provide inference. These are the main components of a Bayesian analysis and they have been demonstrated here in a simplistic manner to showcase that all we are doing is combining two pieces of information together for inference.

Collect Data *(estimated time: 10 – 20 minutes)*

After the teacher has explained the concepts above, they should put the students into groups and pass out the worksheet. The worksheet will direct the students to the same website that the teacher used for data collection. Each group will be prompted to collect fifty observations. If the students need help, then the teacher should be willing to aid them in collecting the data.

Analyze Data *(estimated time: 10 – 20 minutes)*

The teacher should walk around the room and answer questions from students as they arise. Questions could cover clarifications on what the teacher presented to them earlier (the above information, in relation to their own values). The analysis section of the worksheet is simply following the process the teacher laid out above. The worksheet will walk the students through this process as well, but if they need assistance then provide it.

The teacher should be on the lookout for questions from students about directives six, nine, and ten. Questions could range from those seeking clarification on computations to clarification on how to use the web-app for analysis. These directives cover the analysis of the two main questions posed. Other directives might pose difficulties too, but they are mostly setting up for these directives (the analysis of the data).

The teacher knows their students best. I can offer general suggestions, like above, but the teacher will always know what works or does not work for their students. Make sure to check if students look lost or confused. If they appear to be, then make sure to ask them where they got lost and present the material to them again (in a way that makes sense to them). Be calm and patient and make sure to hear them out. Knowing your students and presenting alternative versions of the same approach but tailored to them might help them.

Interpret Results *(estimated time: 10 – 20 minutes)*

The teacher should let the students write their own interpretation prompted to them in the worksheet but go over it in the last part of class. The students will provide a discussion of how strong (or weak) the support for their claim is, how much influence the two approaches played a part in the analysis, and if they would personally believe the claim anymore given all this information.

If the teacher is willing to combine all the results to demonstrate another concept to the students, then follow through the optional problem mentioned below. Otherwise, let the students discuss these topics and guide them (as necessary).

Optional Concept of Combining Every Groups Data:

An interesting aspect of Bayesian inference is the ability to update probabilities upon seeing new data. This might not sound interesting, but it is quite a useful ability. Supposing that no two groups in the class copied each other's random observations, each group brings to the table a new set of observations that were randomly generated. Thus, we can consider each group's data row as a multiple step process (or a singular step process) of updating.

To demonstrate, let us suppose that we have five groups in the classroom and that their data entries are those presented below:

	Count of Land	**Count of Water**	**Total Observations**	**Proportion Estimate**
Group 1 Data	14	36	50	36/50
Group 2 Data	15	35	50	35/50
Group 3 Data	18	32	50	32/50
Group 4 Data	12	38	50	38/50
Group 5 Data	15	35	50	35/50

If the assumptions stated above (randomly generated and not copied) hold, then we can combine all this information into a singular Data row and use all the Data for the analysis process.

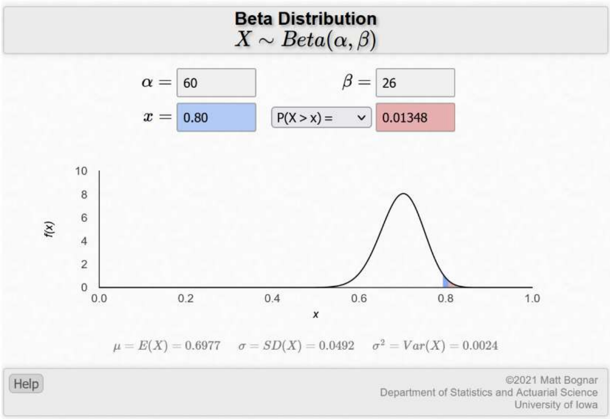
For example:

	Count of Land	**Count of Water**	**Total Observations**	**Proportion Estimate**
Grid	5	11	16	11/16
Simulation	74	176	250	176/250
Combined	79	187	266	187/266

The teacher can explain how this phenomenon could be useful as follows: the more data we can collect, the better our inference will be. Since we can combine every group's data, we are making better inference. The teacher can re-run through answering questions 1 and 2 from above with this new data to highlight this to the students.

<u>Possible extension (changing the region being tested):</u>
The teacher might wish to analyze a different region than "less than 50% water on earth." This is fine! There is nothing stopping you or your students from changing the region of investigation! Say you want to investigate the claim of "more than 80% water on earth." The parameters in the web-app would remain the same, but you would need a new x value to compute the probability to provide support for or against the claim.

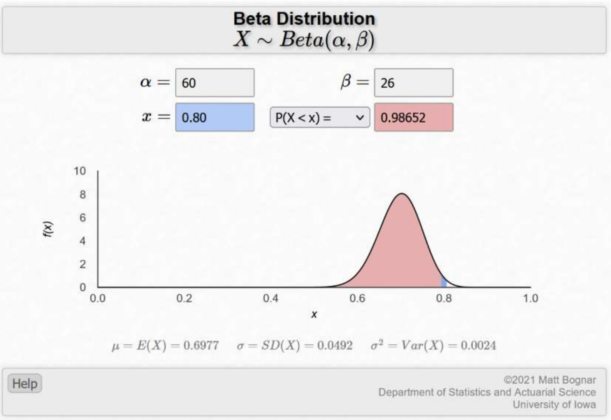
Using the parameters obtained through the hypothetical that the teacher might present before allowing the students to work on the assignment themselves, we obtained the following probabilities for and against (in that order) the new claim:



Beta Distribution
$X \sim Beta(\alpha, \beta)$

$\alpha =$ 60 $\beta =$ 26
$x =$ 0.80 P(X > x) = 0.01348

$\mu = E(X) = 0.6977$ $\sigma = SD(X) = 0.0492$ $\sigma^2 = Var(X) = 0.0024$



Beta Distribution
$X \sim Beta(\alpha, \beta)$

$\alpha =$ 60 $\beta =$ 26
$x =$ 0.80 P(X < x) = 0.98652

$\mu = E(X) = 0.6977$ $\sigma = SD(X) = 0.0492$ $\sigma^2 = Var(X) = 0.0024$

Take note of the 1.348% probability in favor of the claim and the 98.652% probability against the claim. The evidence required to analyze the "less than 50%" claim did not need a lot of data, but it seems that this new claim could require more data (depending on how sure you feel 98.652% is in comparison to 1.348%). Sometimes new claims require more data to support or refute!

Attached Materials

Listed below are materials that have been discussed in the lesson plan above:

- Weighted Average Reference document
- How to use Matt Bognar Probability Distribution Web Applet document
- How to use RANDOM.ORG document
- Beta Distribution Briefly Explained document
- Bayesian Proportion Inference Worksheet document

References

Bognar, M. (2021, October 20). *Matt Bognar Homepage*. Retrieved from Matt Bognar Homepage Web site: https://homepage.divms.uiowa.edu/~mbognar/

Randomness and Integrity Services Ltd. (2021, October 20). *Random Geographic Coordinates*. Retrieved from RANDOM.ORG: https://www.random.org/geographic-coordinates/