Biodiversity for the National Parks

Introduction to Data Analysis: Capstone Project 2

By Brock Kusmick

The Data

The species_info.csv file on it's surface, contains a list of many different species of animals as well as their "Conservation Status," or, in other words, how endangered that particular species is. There are 4 columns of data: Category, Scientific Name, Common Name, and Conservation Status.

- The "category" column contains data on the type of animal, such as "mammal," "bird," "reptile," etc. There are 7 total categories.
- The "scientific name" column contains the scientific names for each species. There are roughly 5500 unique names in the data.
- The "common name" column contains the common names for each species.
- The "conservation status" column contains the level of endangerment of each species. There are 5 possible statuses. "Species of concern," "threatened," "endangered," "in recovery," or blank/null, meaning the species is doing well.

The Data

In addition to the type of data contained within the file, it is also important to understand the content of that data. For instance, there are roughly 5550 unique scientific names within the data. However, the file itself contains more than 5800 rows. Are these duplicates the result of carelessness, or perhaps some other reason. It is important to make sure our data is clean before we begin analyzing it. For our purposes it is sufficient to work with the 5550 unique scientific names.

Other interesting things to note are the number of species within each category. Vascular plants make up most of the data with over 4000 records, while amphibians and reptiles make up the least. It is also interesting that vascular plants have the lowest percentage of protected species.

Finally, despite there being roughly 5550 species, only 15 are endangered. That's less than .3%! Of course any amount of species being endangered is tragic, but it goes to show we are doing a pretty good job of protecting the other species that live here.

Statistical Significance

Are certain types of species more likely to be endangered than other?

To answer this question we can use a hypothesis test to determine the statistical significance of a hypothesis. For instance, we see in the data that mammals and birds have a similar proportion of species which are protected. Mammals about 17% and birds about 15%. Is this slight difference due to chance? In this case we hypothesize that this difference IS the result of chance, also called our null hypothesis.

We then run a Chi-Squared test on mammals and birds. We choose Chi-Squared because this type of test is best for categorical data. We then get a p-value from this test which determines whether or not we will reject the null hypothesis. A value of less than 0.05 means we will reject the null. In other words, we are at least 95% certain that difference is significant. In this case we received a p-value of about 0.68. Very high! So we can't be certain that the difference between mammals and birds is not due to chance, and cannot reject the null.

Statistical Significance

We then test if there is a significant difference in the percentage of protected mammals vs. reptiles. Our null hypothesis will again be that the observed difference is a result of chance. When we run the same chi-squared test from our first hypothesis test we receive a p-value of about 0.033. Clearly this is less than 0.05. In this case we can conclude that we are at least 95% certain that the observed difference is not a result of chance. In other words, there is a significant difference in the percentage of protected mammals and reptiles.

Thus, we conclude that some types of species are more likely to be protected than others.

Recommendations

I’m sure it is heartbreaking for any conservationist to have to hear about endangered species. However, the fact of the matter is there will likely always be endangered species given the growing human population. There is no one correct solution. We can only inform ourselves and do the best we can with the information we are given. That is where data analysis comes in. It allows us to make the most informed choices possible.

Given the data collected, my recommendation is for conservationists to concentrate their efforts on saving the protected birds and mammals. They have a much higher percentage of protected species. Birds in particular have the highest in absolute quantity of protected species as well at 75. In my opinion, this is likely due to the fact that birds and mammals are more likely to be hunted than other species, and are most affected by deforestation. By concentrating their efforts on these species they will be able to save the maximum amount of protected species possible.

Sample Sizes

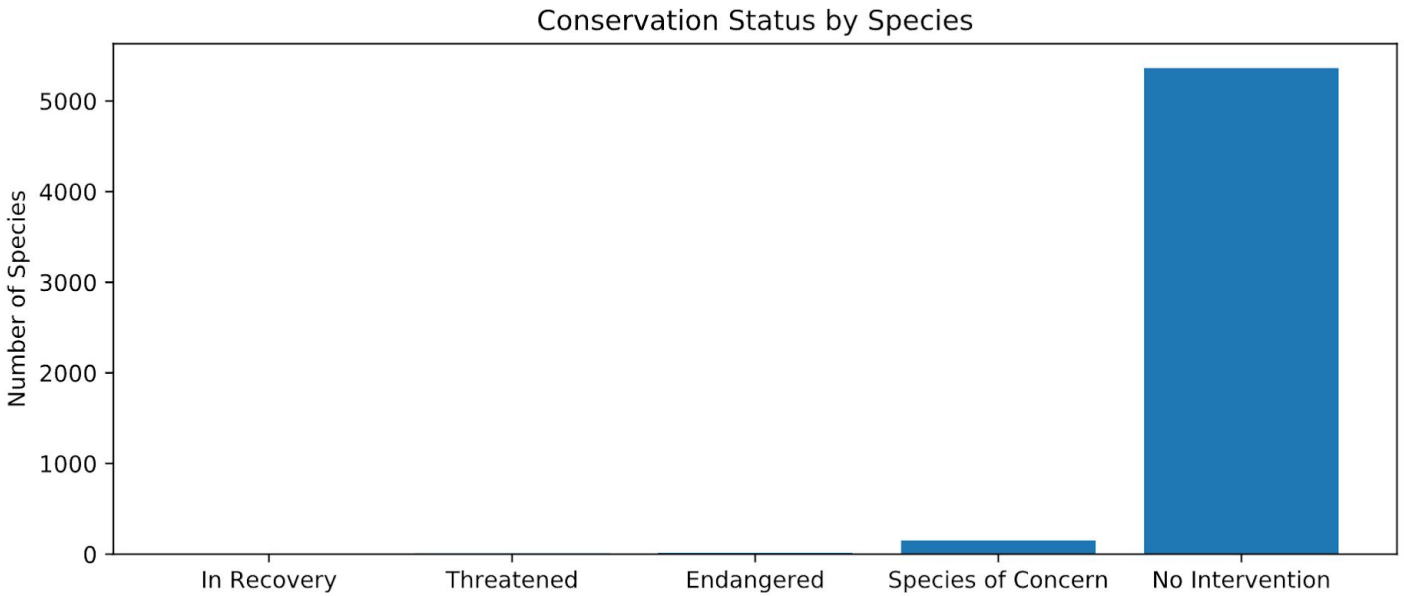
Determining sample sizes can be tricky. On one hand, we want to have as many samples as possible so we can get the best representation of the entire population. On the other hand, we want to minimize the amount of samples we need in order to cut down on other factors, such as file sizes, time it takes to collect the samples, etc. If we can get an accurate representation with 1000 samples, for instance, it's pointless for us to collect 5000 samples.

That's where the sample size calculator comes in. It gives the minimum number of samples needed in order to obtain the desired observable effect at the desired confidence level.

Samples Sizes

To use the calculator, we first need a baseline percentage. In our case, it was observed that last year 15% of sheep in Bryce Park had foot and mouth disease. This is our baseline. This year, we would like to be able to detect at least a 5% difference in the occurrence of foot and mouth and a desired confidence level, which we choose to be 90%. A 5% drop from 15% is a change of 33.33%. Therefore, we want to be able detect a change of at least 33.33%, the “minimum detectable effect.” From here we simply plug our numbers into the calculator. 15% - baseline, 90% - significance, 33% - detectable effect, and voila! We need 890 samples to achieve this.

From here we can then easily calculate the time it will take to obtain these samples using our weekly sheep observation data. At Yellowstone it will take a little under two weeks since we can collect 507 samples per week. And at Bryce it will take about 3.5 weeks since we can collect 250 samples per week there. Simply divide the required samples (890) by the number of sheep observed per week (507 and 250, respectively)



Type URL here