BIODIVERSITY FOR THE NATIONAL PARKS

Codecademy Capstone Project

Maria Sarafi

Description of the ‘species_info.csv’ DataFrame

The National Parks Service has provided the CSV file ‘species_info.csv’ with data about different species in the National Parks. The DataFrame includes:

1. The species category
2. The scientific name of each species
3. The common names of each species
4. The species conservation status

The names of the columns of the DataFrame are: category, scientific name, common name and conservation status.
There are 5541 different species that belong in 7 categories: ‘Mammal', ‘Bird', ‘Reptile', ‘Amphibian', ‘Fish', 'Vascular Plant’, 'Nonvascular Plant’.
There are 5 different values of the conservation status: ‘No Intervention’(5363 species), 'Species of Concern’(151 species), ‘Endangered’(15 species), ‘Threatened’(10 species) and 'In Recovery’(4 species).

The majority of the species has conservation status equal to ‘No Intervention’(NaN) and only a small number of species are categorised as needing some sort of protection.

Significance Calculations

Examining the data of the DataFrame, I should perform a sequence of significance calculations between different categories of endangered species in order to decide whether certain types of species are more likely to be endangered. Considering the data, it seemed like mammals are more likely to be endangered than birds. I should investigate if there is a significant difference. I've performed a significance test to see if there is a significant difference or this difference is due to chance.

First, I've performed a significance test between mammals and birds using the values for protected and not-protected species.
The result of the test has proved that the difference is not significant.

Second, I've performed a same significance test between reptiles and mammals using the values for protected and not-protected species.
In this case, the result of the test has proved that the difference is significant.

Recommendation

We can safely conclude that certain types of species are more likely to be endangered than others.
The significance calculations have proven that there is a larger number of mammals that is in danger and that is not a result of chance.

Therefore, the conservationists should focus more on the protection of mammals than on the protection of reptiles.

Sample Size Determination

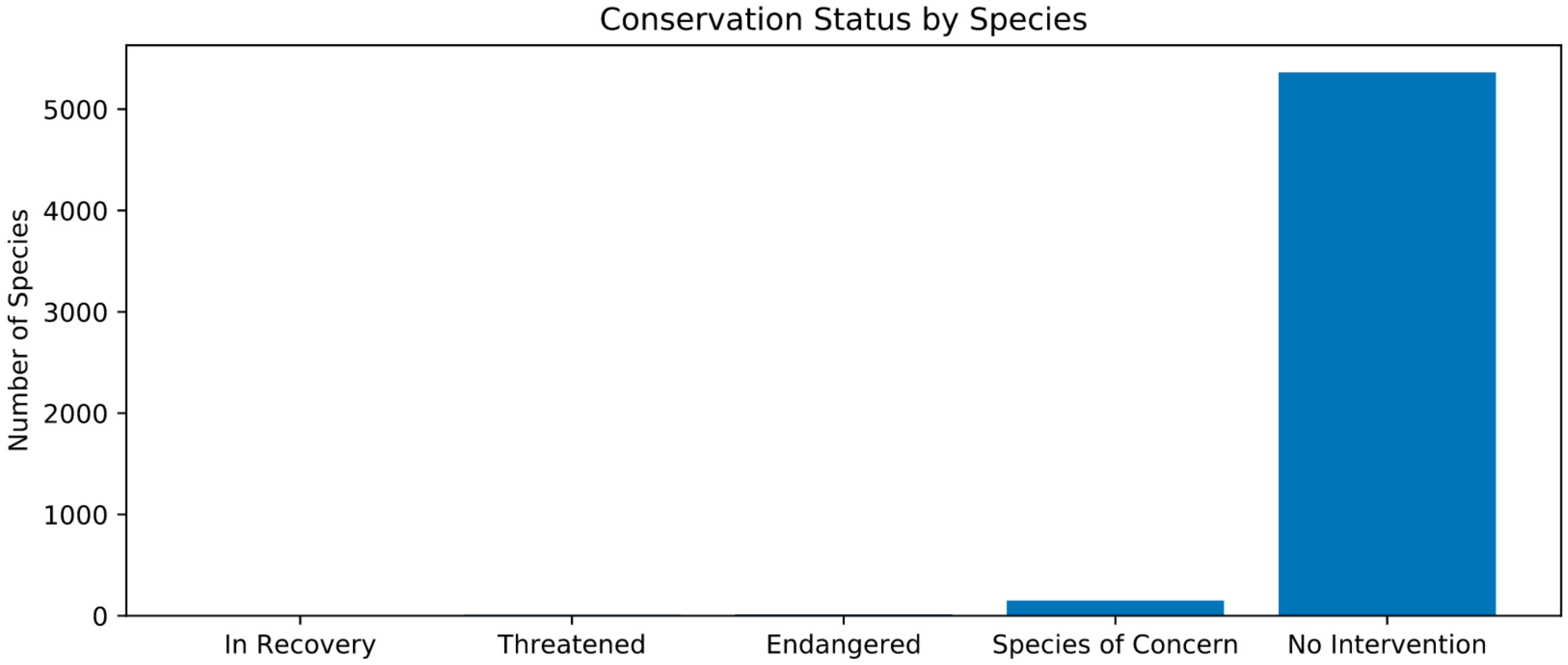
For the sample size determination for the foot and mouth disease study we need to know:

1. The baseline percentage
2. The ‘Minimum Detectable Effect
3. The level of Statistical Significance

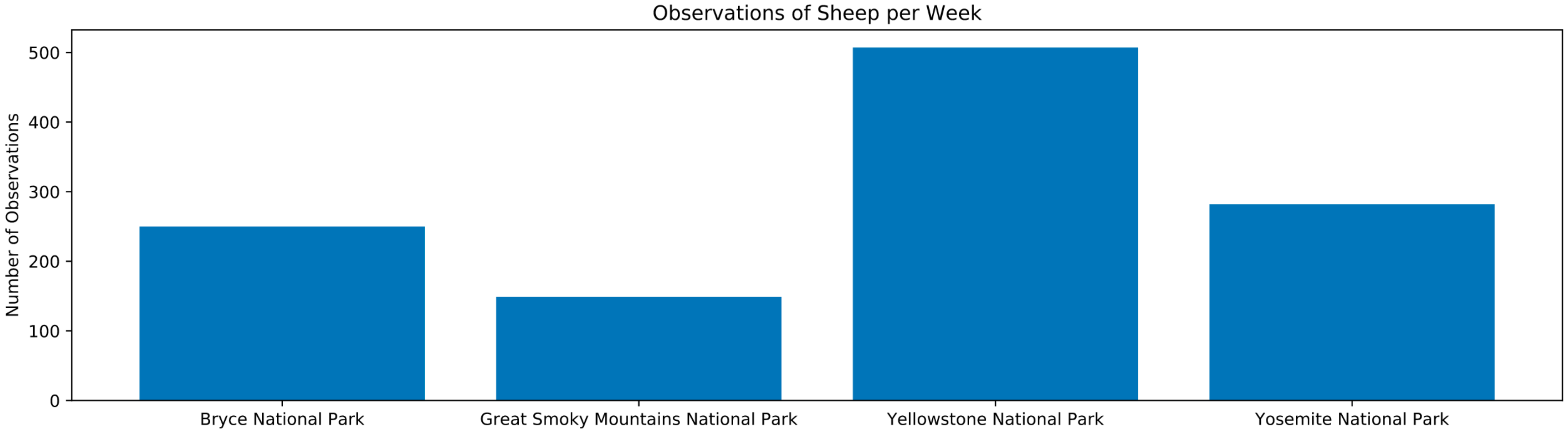
The scientists currently have recorded that last year 15% of sheep at Bryce National Park have foot and mouth disease, therefore the baseline percentage is 15.
The "Minimum Detectable Effect" is a percent of the baseline, so if the scientists want to observe a 5% change with confidence, the minimum detectable effect would be equal to 100 * 5 / baseline (= 33.33%).
For our calculations we use the default level of significance (90%).

Using these values into the sample size calculator, I’ve calculated that the sample size consists of 870 sheep.
The scientists need to spend 1 week at Yellowstone National Park and 3 weeks at Bryce National Park in order to observe enough sheep.

Plotting Conservation Status by Species



Plotting Sheep Sightings


Observations of Sheep per Week
Number of Observations
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park