



Biodiversity for the National Parks

07.02.2018

Submitted by:
Nandini Deka

Overview

The National Parks Service would like to perform some data analysis on the conservation statuses of endangered species across different National Parks and to investigate if there are any patterns or themes to the types of species that become endangered.

Objective

To determine if certain types of species are more likely to be endangered than others.

Species Data

A CSV is provided based on data from the National Park Services - species_info.csv, with the following information in it:

- The category in which the species belongs - category
- The scientific name of each species - scientific_name
- The common names of each species - common_names
- The species conservation status - conservation_status

Key Observations:

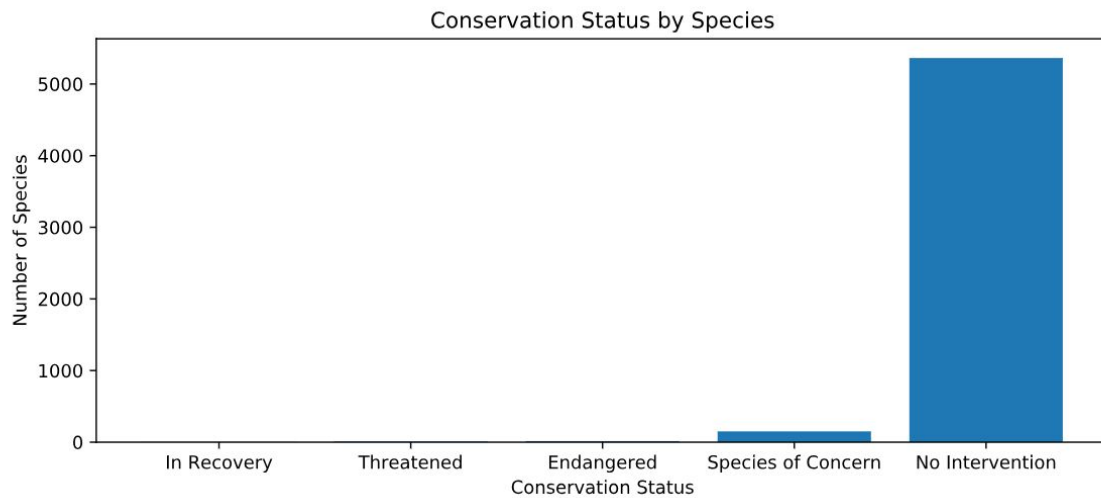
1. Total No of different species : 5541
2. Category of species in the data : ['Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant', 'Nonvascular Plant']
3. Conservation Statuses: ['Species of Concern', 'Threatened', 'Endangered', 'In Recovery']

Conservation Status:

The conservation status of a species has been defined as one of the following options:

- Species of Concern: declining population or appears to be in need of conservation.
- Threatened: vulnerable to endangerment in the near future.
- Endangered: seriously at risk of extinction.
- In Recovery: formerly Endangered, but currently not in danger of extinction throughout all or a significant portion of its inhabitable range.
- No Intervention: has been assigned as the status for all species without any conservation status, i.e. to None/NaN values.

Segmenting the data by the assigned conservation status, we get the following plot:



Because of the size of the last bar, the rest of them are hardly visible. Here is the division of Species by Statuses:

Conservation Status	# of Species
In Recovery	4
Threatened	10
Endangered	15
Species of Concern	151
No Intervention	5363

Protected Vs Not-Protected:

Further analysing the data by grouping every Category into Protected (Conservation Status is not 'No Intervention') and Not-Protected (Conservation Status is 'No Intervention') we can look at the Percentage Protected for any Category and works towards our initial question - are certain types of species more likely to be endangered?

	category	not_protected	protected	percent_protected
0	Amphibian	73	7	8.750000%
1	Bird	442	79	15.163148%
2	Fish	116	11	8.661417%
3	Mammal	176	38	17.757009%
4	Nonvascular Plant	328	5	1.501502%
5	Reptile	74	5	6.329114%
6	Vascular Plant	4424	46	1.029083%

From the above table it looks like Mammals are more likely to be endangered than Birds. But is this difference in percent_protected significant enough to make this claim? In order to determine if the difference is significant, we perform a chi-squared test.

Chi-Squared Hypothesis Testing:

Performing Chi-Squared test on a couple of combinations of categories, we get the following results:

- Pvalue for Mammal vs Bird is : 0.687594809666
- Pvalue for Mammal vs Reptile is : 0.0383555902297

The p-value for the chi-squared test for the slight difference in percentages protected for Mammals and Birds is ~ 0.688 . Thus we can conclude that the difference is not significant and it is a result of chance. Hence we can't claim that Mammals are more likely to be endangered than Birds.

But the p-value for the chi-squared test for the slight difference in percentages protected for Mammals and Reptiles is ~ 0.038 - which is significant. Hence Mammals are definitely more likely to be endangered than Reptiles.

Conclusion and Recommendation:

Thus we can conclude that certain species are more likely to be endangered than the others. In order to determine which are more likely we will have to conduct hypothesis testing for different combinations of categories and work with the ones which are most likely to be endangered than the rest of the categories.

In this case we found Mammals to be more likely to be endangered than Reptiles and conservationists can start looking at the different species of Mammals which are present and how to work towards making them less endangered over time.

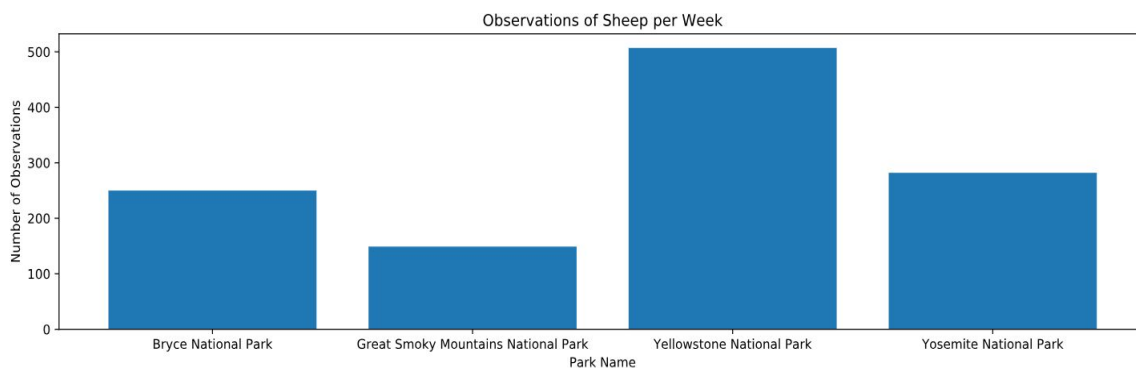
Foot and Mouth Disease Reduction Effort

Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease in sheep at that park. Scientists want to test if the program has been successful, which is quantified as 5% points decrease in the disease.

Data:

There is a dataset of sightings of various species in different National Parks per week. But this dataset only has the scientific names. Hence we have to use our Species dataset to find the observations for any type of sheep sighted. From the Species dataset, filtering out for Category as Mammal and Common Name containing 'Sheep' we get a subset of data for all types of Sheep in our dataset. Combining this with the Observations data we get the total observations of sheep in any National Park.


National Park	Observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282



Sample Size Determination:

The only information available to scientists is that in the previous year 15% of sheep have Foot and Mouth disease in Bryce National Park. Taking this as the 'baseline conversion rate' and using 5% as our desired conversion rate, we can calculate the minimum detectable effect as:

$$\text{Minimum Detectable Effect} = 100 * 5 / 15 = 33.33\%$$



Using the default level of significance as 90% and the above two values for Baseline Conversion Rate and Minimum Detectable Effect we can use the sample size calculator to get the Sample Size per Variant, which turns out to be 870. Hence if the scientists wanted to be sure that a >5% drop in observed cases of foot and mouth disease in the sheep at Yellowstone was significant they would have to observe at least 870 sheep.

Thus, the time required to collect observe the required sample size in Yellowstone National Park is approximately 2 weeks, with weekly observations of 507. And it is approximately 4 weeks in Bryce National Park, with weekly observations of 250.