Biodiversity for the National Parks

Andrei Karlsson

Key Lines of Investigation

- To see if there are any patterns or themes to the types of species that become endangered.

- In particular, to see if certain species are more likely to be endangered than others.

- To assist scientists in tracking sheep locations.

- To assist Park Rangers at Yellowstone National Park to see if their program to reduce the rate of foot and mouth disease at the park is working.

Species Data

▶ The data obtained from the National Parks Service ('species_info.csv') contained only the category and scientific name of each species in the National Parks, their associated common names and conservation status.

▶ Since our investigation focused on the species that are or likely to become endangered, it was natural to group the data by conservation status.

▶ Because the species data frame did not have conservation status data entries for many of the species, these null values were assigned a 'No Intervention' conservation status.

▶ The results of this can be seen below, and are graphed in Figure 1 (slide 9):

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

Significance Calculations

▶ With a view of testing the differences in conservation statuses of the species for significance, the table was manipulated so that for each category of species, it was easy to see how many of those species were protected or not.

▶ This involved using Python calls for creating and renaming new columns and pivoting the table:

Category	Protected	Not Protected	Percent Protected
Amphibian	7	72	0.088608
Fish	75	413	0.153689
Bird	11	115	0.087302
Mammal	30	146	0.170455
Nonvascular Plant	5	328	0.015015
Reptile	5	73	0.064103
Vascular Plant	46	4216	0.010793

- Given categorical data among multiple sources of data, it was clear significance calculations would be done by Chi-squared hypothesis testing.

- After coding the appropriate contingency tables, with the chi2_contingency call imported, p-values of the hypothesis tests were obtained to observe whether the difference between the protection status of two categories of species was significant.

- The p-values of 0.05 or less indicate a significant difference between the two categories and 0.05 indicate no significant difference, or in other words, difference as a result of chance. Some results are shown below:

Categories Tested	P-value	Difference
Fish and Birds	0.68759481	Not significant
Reptiles and Mammals	0.03835559	Significant
Vascular and Non-Vascular Plants	0.66234195	Not significant
Amphibians and Mammals	0.12757670	Not significant
Birds and Reptiles	0.05313542	Not significant
Birds and Non-Vascular Plants	4.61226803e-79	Significant

Key Recommendations

- Our results show that while many differences in the percentage of some species that fall into a protected category was down to chance, there were some species who's difference in protection status was statistically different.

- These included, for example, the difference between Mammals and Reptiles, and Birds and Vascular Plants.

- So the findings show that certain types of species are more likely to be endangered than others.

- The key recommendations to follow up from this project include careful analysis of the full set of results of which species are more likely to be endangered, in order to see what programs in the National Parks should be introduced or prioritised.

- Storing data which directly conveys which species are protected or not will be helpful for future analysis.

- It would also be recommended that analysis be carried out further down the classification scale, in order to see if there are differences among those species that are protected.

Foot and Mouth Disease

▶ Recorded sightings of different species at several national parks for the past seven days were obtained from the National Parks Service in a data frame called ('observations.csv'). This contained the scientific name of the species, the park name of where they were sighted and the number of observations.

▶ Park Rangers at Yellowstone National Park had been running a program to reduce the rate of foot and mouth disease at that park, and last year it was recorded that 15% of sheep at Bryce National Park had foot and mouth disease. Analysis was carried out to detect reductions at Yellowstone of at least five percentage points.

▶ Firstly the data were manipulated by the creation of a new column 'is_sheep' in the species data frame using a lambda function and row selection to mark sheep species. This was then merged with the observations data frame, and manipulated again by grouping the data by the various park names, so the sightings of sheep could be easily seen. This is shown in Figure 2 (slide 10).
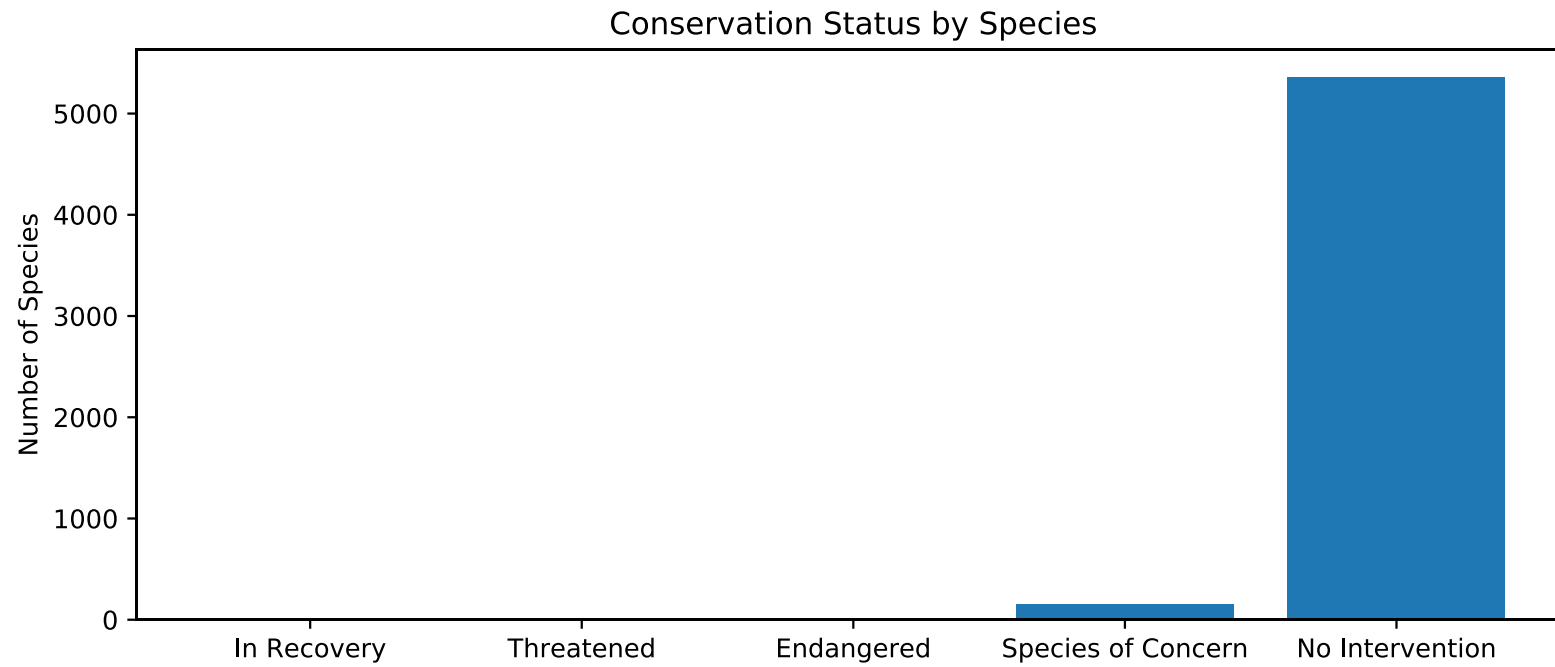
Foot and Mouth Disease: Sample Size Determination

▶ In order to determine with confidence what percentage of sheep have foot and mouth disease at the four parks, the number of sheep that would need to be observed was calculated.

▶ The baseline percentage was 0.15, given the 15% prevalence at Bryce.

▶ The minimum detectable effect was calculated as 0.33.

▶ Given the standard statistical significance of 90%, the sample size per variant of sheep was calculated as 870.

▶ Therefore, the following table shows for how long observations must be made in each park in order to determine foot and mouth prevalence with confidence.

National Park	Weeks Observing
Bryce National Park	3 weeks 4 days
Great Smoky Mountains National Park	5 weeks 6 days
Yellowstone National Park	1 week 6 days
Yosemite National Park	3 weeks 1 day

Figures

► Figure 1

Conservation Status by Species

▶ Figure 2

Observations of Sheep per Week