Capstone Project

Biodiversity for the National Parks

March 2018

Completed Code in the Learning Environment

Joao Oliveira

Exploring Species dataset

The dataset species_info.csv contains information about the different species in the National Parks. It's also divided into the following columns:

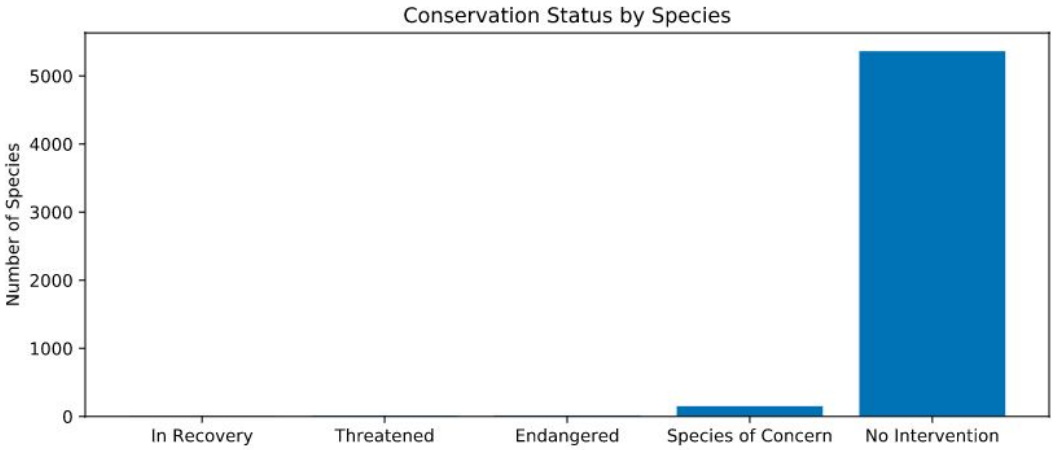
- Category - category of the specie (e.g. Reptile) as a string
- Scientific_name - scientific name of the specie (e.g. Bos bison) as a string (this is the unique id)
- Common_names - common name for the specie (e.g. American Bison, Bison) as a string
- Conservation_status - current conservation status of the specie (e.g. Endangered) as a string

From the analysis, there's also to highlight that:

a) Table includes a total of 5541 unique species
b) Column 'Category' includes the types: 'Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant' and 'Nonvascular Plant'.
c) Column 'Conservation_status includes option: 'nan', 'Species of Concern', 'Endangered', 'Threatened' and 'In Recovery'.

Exploring Species dataset

d) After cleaning the dataset to replace conservation_status ‘nan’ by ‘No Intervention’, it’s important to notice that 5363 of those species are actually ‘No Intervention’ (96.78% of total)



	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

Significance calculations

As a next step we grouped the data to investigate if certain types of species were more likely to be endangered.

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

We used chi-square to compare the following categories (for both the null hypothesis is that differences happened by chance):

Mammals vs Birds: where pval = 0.687 so the difference does not seem significant (pval > 0.05)
Reptile vs Mammal: where pval = 0.038 so the difference is significant (pval < 0.05)

Recommendation for conservationists

From the analysis described in the previous slide, we can conclude there are species were more likely to be endangered.

Therefore from the table and the significance calculations, we can conclude that Birds and Mammals are the types of species where more resources need to be allocated (eg. more regular check-ups or tracking). This will reduce the chances of species of these categories to be endangered.

Sample Size Determination

The Problem: “Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park. The scientists want to test whether or not this program is working. They want to be able to detect reductions of at least 5 percentage points. For instance, if 10% of sheep in Yellowstone have foot and mouth disease, they'd like to be able to know this, with confidence.”

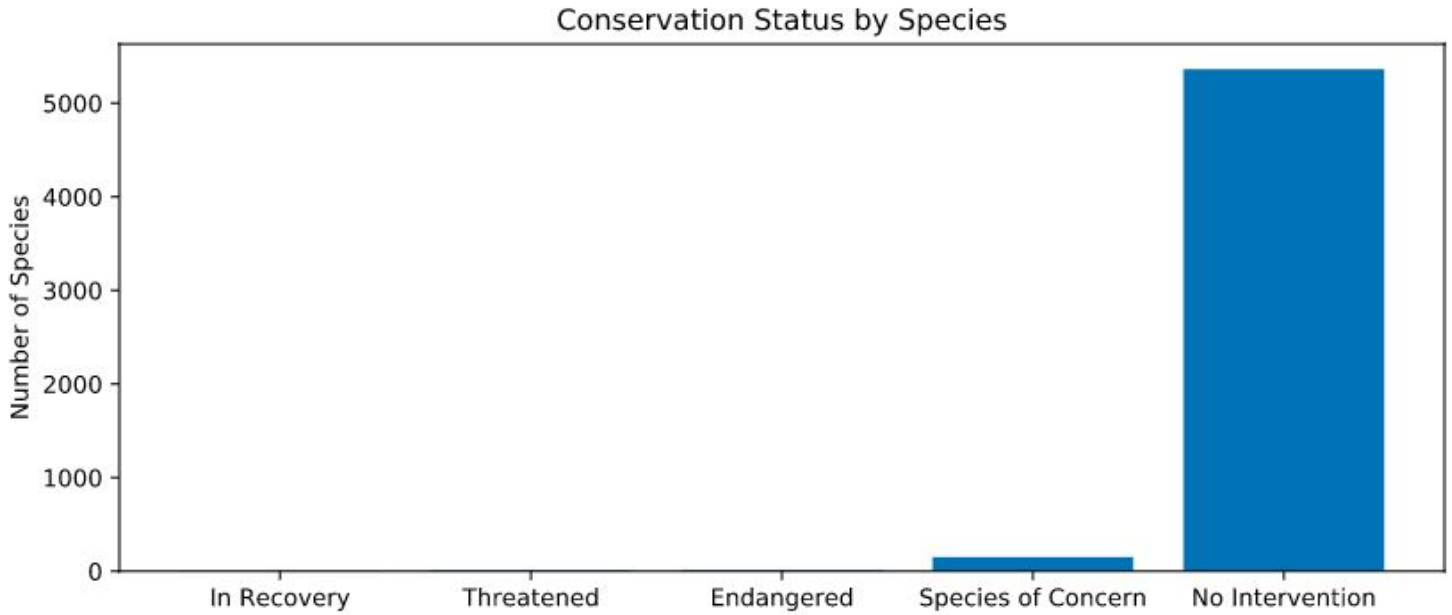
Steps: we needed to determine the number of sheep that they would need to observe from each park to make sure their foot and mouth percentages are significant to analyze the disease spread. The following inputs were used:

- **Baseline =** 15% (the percentage of sheeps with the disease at Bryce National Park last year)
- **Minimum Detectable Effect** = 33%, being ((15%-10%))/15% = 0.333x100 = 33%
- **Significance** = 90% (recommended in the instructions)

Based on those values, the online calculator provided 890 as the sample size needed.

To analyze how long (in weeks) would take to complete reach that number of observations, we just needed to divide the sample size by the number of weekly observations in a given park.

Appendix - Graphs



Appendix - Graphs