codecademy

Codecademy Capstone Project

Biodiversity for the National Parks

Michael Sendik

February 22nd, 2019

Analysis of Species' Conservation Status

Dataset Characteristics

What specifically did I notice about the dataset, species_info.csv

Data Set Purpose

- Dataset to be used to perform analysis on the conservation status of species and to investigate if there are any patterns or themes to the types of species that become endangered

Key Characteristics

- 5824 records
- Four types of data captured: specie category, scientific name, common name, and conservation status
- Seven unique species, 5541 unique scientific names, and 5504 unique common names (implies that there is some redundancy in the labeling of common names and must therefore use scientific names category for analysis)
- Conservation status was defined as one of four options: Endangered, In Recovery, Species of Concern, and Threatened
- Only 191 records had a conservation status assigned

Other Characteristics

- Initial analysis of the dataset illustrates that the majority of records are for vascular plant species (~77%)
- Of records with Conservation Status approximately 41% are for Birds
- Majority of species with Conservation Status are protected as “Species of Concern”; i.e., declining population or appears to be in need of conservation

Species Category	% of Records	% Records w/ Status
Vascular Plant	76.8%	24.1%
Reptile	1.4%	2.6%
Nonvascular Plant	5.7%	2.6%
Mammal	3.7%	19.9%
Fish	2.2%	5.8%
Bird	8.9%	41.4%
Amphibian	1.4%	3.7%
Total	**100.0%**	**100.0%**

Conservation Category	No. Records	% of Total
No Intervention	5363	96.72%
Species of Concern	151	2.76%
Endangered	15	0.27%
Threatened	10	0.17%
In Recovery	4	0.07%
Total	**5541**	**100.00%**

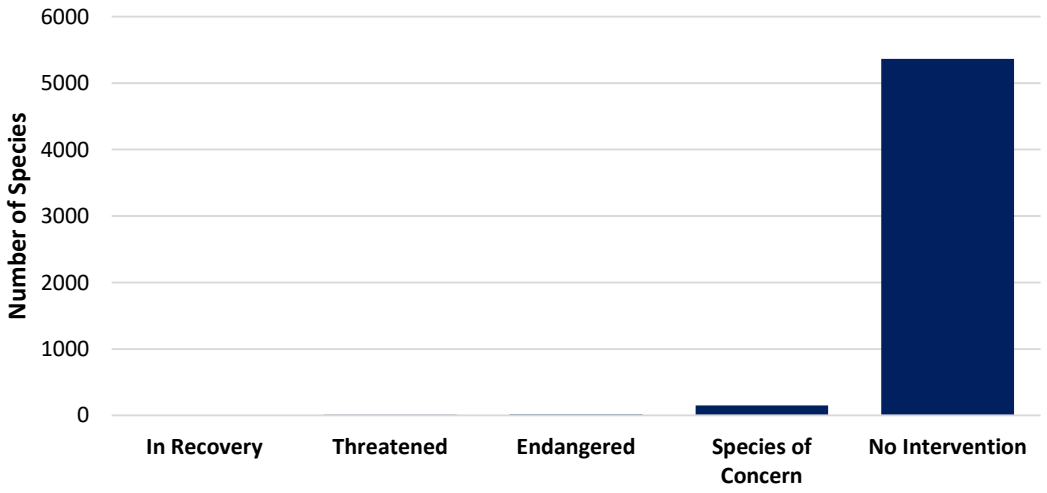
Dataset Characteristics, continued

Graphical and pivot analysis of conservation status

Specie Protection by Category

Species Category	Number of Species		
	Protected	Not Protected	% Protected
Amphibian	7	72	8.86%
Bird	75	413	15.37%
Fish	11	115	8.73%
Mammal	30	146	17.05%
Nonvascular Plant	5	328	1.50%
Reptile	5	73	6.41%
Vascular Plant	46	4216	1.08%
Grand Total	**179**	**5363**	3.23%

Conservation Status by Species



Commentary

- Consistent with our earlier findings, Birds are the species that currently have the greatest level of protection; however,
- Mammals have the greatest amount of protection as a percentage of total species, which may imply they are more likely to be endangered than birds ***[hypothesis to be tested in next section]***
- Of those species that have protection, the vast majority of that protection is categorized as “Species of Concern”
- Most of all the species listed within the dataset have no conservation / specie protection status
- *Note: although the bar chart below was requested as part of the analysis exercise, I feel it does little to further the analysis. I think it would be more useful to exclude the “No Intervention Category”. Doing so would greatly reduce the scale of the y-axis and allow readers to see observations identified within other conservation categories*

Significance Calculations

Testing endangered status between different categories of species

Chi-Squared Analysis

Contingency Table (Mammal vs Bird)		
Specie	**Protected**	**Not Protected**
Mammal	30	146
Bird	75	413
pvalue = 0.687594		

Contingency Table (Reptile vs Mammal)		
Specie	**Protected**	**Not Protected**
Reptile	5	73
Mammal	30	146
pvalue = 0.03835		

- Purpose of the analysis was to determine if certain types of species were more likely to be endangered than others
- In each test-case, the null-hypothesis determined whether or not those differences were significant
- Because the data to be compared was categorical and in multiple pieces, Chi-Squared analysis was chosen to test each null-hypothesis
- ***Mammal vs Bird*** - we can conclude that the difference between the percentages of protected birds and mammals is not significant and is a result of chance; i.e., pvalue = 0.68759
- ***Reptile vs Mammal*** - we can conclude that the difference between the percentages of protected reptiles and mammals is significant and is not a result of chance; i.e., pvalue = 0.03835

Based on our initial Chi-Squared analysis, we can conclude that certain types of species are more likely to be endangered than others; e.g., mammals

Recommendations

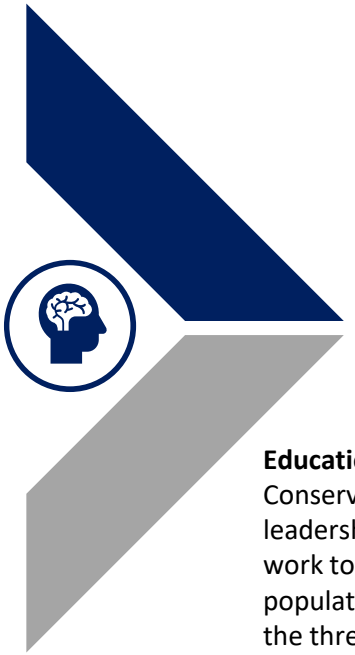
What should conservationists do about endangered species?

1

Application
As we have seen that some species are more likely to be endangered, conservationists should reduce thresholds required for certain species to obtain such status; e.g., mammals

2

Tracking
Conservationists must actively track specie population and adjust status / thresholds of protection as appropriate; this will ensure species are protected especially as populations fluctuate year-over-year



3

Education
Conservationists must take a leadership role and actively work to educate the population to help reverse the threat suffered by some species; e.g., educate the population on climate change, pollution, etc.

Supplemental Data Exercise

(Species Tracking and Foot & Mouth Sample Size Determination)

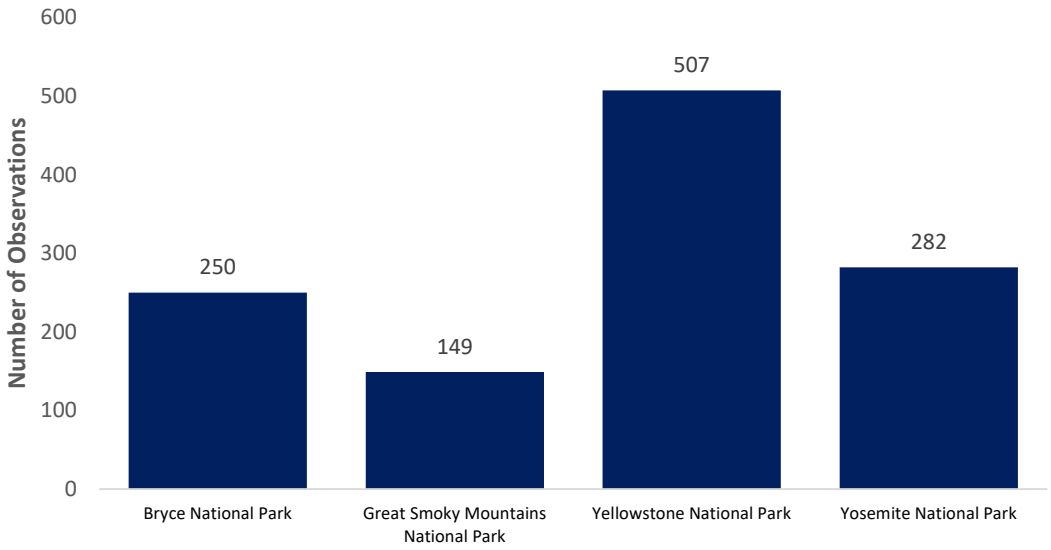
Species Tracking

National Park Service provided supplemental data from several parks for 7-day period

Data Set Purpose

A team of ruminant-enthused scientists has been tracking the movements of various species of sheep across different national parks and have asked for assistance in analyzing the observation and species to help track sheep locations.

Observations of Sheep Per Week



Commentary

- Leveraged a lambda function to add “Is_sheep” supplemental column to all records that contained “sheep” in the common-name column (note: all data was filtered on “mammal”
- Supplemental dataset was then merged with the initial dataset, “species_info” on the string “scientific name” to determine the total number of observed sheep per week in each park
- From the bar chart, it is easy to see that Yellowstone National Park has many more observations of sheep each week

Foot & Mouth Reduction Effort

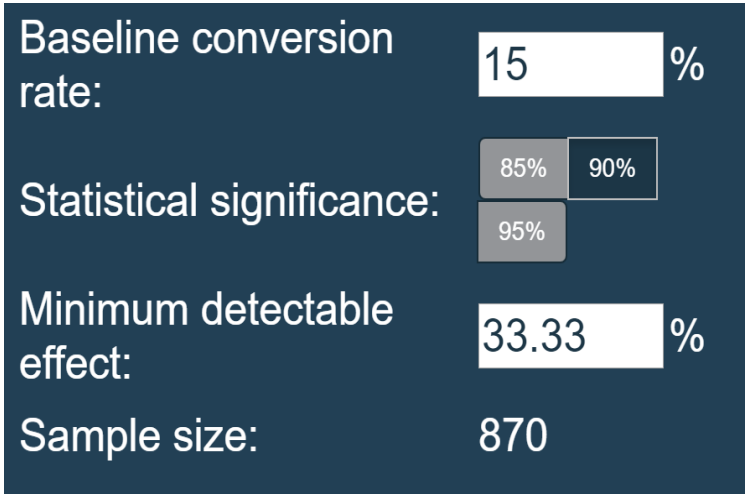
Sample Size Determinization

Problem Background

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.

The only information that the scientists currently have is that last year it was recorded that 15% of sheep at Bryce National Park have foot and mouth disease. Using this value and an online size calculator, determine the number of sheep that they would need to observe from each park to make sure their foot and mouth percentages are significant, using a level of significance of 90%

Sample Size Calculator



Commentary

- **Baseline Conversation Rate –** given as the recorded observation from the previous year (15% of sheep at Bryce National Park)
- **Statistical Significance –** level of certainty demanded by the National Park Service
- **Minimum Detectable Effect** – Equal to the desired amount of reductions detected divided by the baseline conversation rate. Note: figured multiplied by 100 to normalize for percentage
- **Sample Size** – Based on the calculations above, it will be necessary to sample 870 sheep to ensure results with a 90% confidence level
- **Sample Period Duration** – To determine the sample period duration in each park, the required sample size was divided by the weekly observations of sheep at each park
  - Yellowstone National Park = 870 / 507 = **1.71 weeks**
  - Bryce National Park = 870 / 250 = **3.48 weeks**
  - Great Smokey National Park = 870 / 149 **= 5.83 weeks**
  - Yosemite National Park = 870 / 282 **= 3.08 weeks**