Biodiversity in US National Parks

Data analysis on the conservation statuses of endangered species

by: Santiago Jaramillo

Note: English is my second language. I apologize for misspelling words or making any mistake.

Data in Species.csv and observations.csv

Data in species.csv

- We were provided with a file “species_info.csv” with data about different species in US National Parks which include information such as category, scientific name, common names and conservation status of each species.

- Total 5.824 species.

- We converted csv file to a DataFrame.

- First 5 rows.

category	scientific_name	common_names	conservation_status
Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	NaN
Mammal	Bos bison	American Bison, Bison	NaN
Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	NaN
Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
Mammal	Cervus elaphus	Wapiti Or Elk	NaN

Data in species.csv

- Values of category: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, and Nonvascular plant.

- 76.7% are Vascular Plants

category	scientific_name
Reptile	79
Amphibian	80
Fish	127
Mammal	214
Nonvascular Plant	333
Bird	521
Vascular Plant	4470

Data in species.csv

- Values of conservation status: Species of concern, Threatened, Endangered, In recovery and No Intervention.

- 3.27% (191 species) require protection.

conservation_status	scientific_name
In Recovery	4
Threatened	10
Endangered	16
Species of Concern	161
No Intervention	5633

Data in observations.csv

- Conservationists have been recording sightings of different species at several national parks. File observations.csv include information such as scientific name, park name and number of observations.

- Total 23.296 observations of different species

- Recorded observations from 4 national parks: Bryce National Park, Great Smoky Mountains National park, Yellowstone National Park and Yosemite National Park.

- We converted csv file to a DataFrame.

- First 5 rows.

scientific_name	park_name	observations
Vicia benghalensis	Great Smoky Mountains National Park	68
Neovison vison	Great Smoky Mountains National Park	77
Prunus subcordata	Yosemite National Park	138
Abutilon theophrasti	Bryce National Park	84
Githopsis specularioides	Great Smoky Mountains National Park	85

Significance Calculations in species

Significance calculations

- In order to identify the types of species more likely to be endangered we performed the following calculations:

    1. Created a column called “is_protected” to identify the which species require protection.
        - True if conservation_status is not equal to No Intervention, and False otherwise.

    2. Grouped “category” and “is_protected” to know how many species of each category required protection.

Significance calculations

- Performed a calculation to determine the percentage of species that are protected of the total number of species in each category.

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

- Conclusion based on results: species in category “mammal” are more likely to be endangered than species in “bird”.

Significance calculations

- In order to validate if the previous statement was true, we performed a chi squared test with categorical data (protected and not protected) and two datasets (mammal and bird).

- Due that the p-value (0.6875948096661336 2) is higher than 0.05, we accepted the null hypothesis which stated that there's no significant difference between the datasets (“mammals” and “birds”).

Significance calculations

- We performed another test comparing “mammal” and “reptiles” categories.

- We reject the null hypothesis, and state that there is a significant difference between two of the datasets (“mammals” and “reptiles”) because we got a p-value (0.03835559022969897 7) less than 0.05.

Significance Calculations with Species and Observations

Significance calculations

- In order to determine how many total sheep observations (across all three species) were made at each national park we performed the following calculations.

1. Created a new column called “is_sheep” in Data Frame SPECIES to identify which is species contain in the word “sheep” in their common_name.

    - True if the common_names contains “Sheep” and False otherwise.

Significance calculations

2. Created a DataFrame called “sheep_species” in which we selected the rows of species where “is_sheep” is True and category “mammal”.

3. We merged two DataFrames sheep_species with observations in order to get a DataFrame called sheep_observations.

4. We performed a function in which we grouped the park name with observations to get the sum of observations for each park_name.

Significance calculations

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

Recommendations

Recommendations

- There are 16 species seriously at risk of extinction that require immediate attention.
- There are 10 species vulnerable to endangerment in the near future.
- It looks like species in category “Mammal” are more likely to be endangered than species in “Reptile” category.

Sample Size Determination

Sample Size Determination

- 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.
    - One of the first steps to designing a successful experiment is determining the number of samples that you need in order to have confidence in the results.
- In order to determine the sample size necessary for determine, a sample size calculator requires three numbers:
    - The Baseline conversion rate.
    - The Minimum detectable effect (lift)
    - The Statistical significance
- We calculate a baseline by looking at historical data for the option that we're currently using. In this case our scientists knew that 15% of sheep at Bryce National Park have foot and mouth disease.

Sample Size Determination

- As the Lift is generally expressed as a percent of the baseline conversion rate. Our scientists know that 15% of sheep at Bryce National Park have foot and mouth disease. They want to be able to detect reductions of at least 5 percentage point. That means that scientists want to increase the conversions by 0.5%.

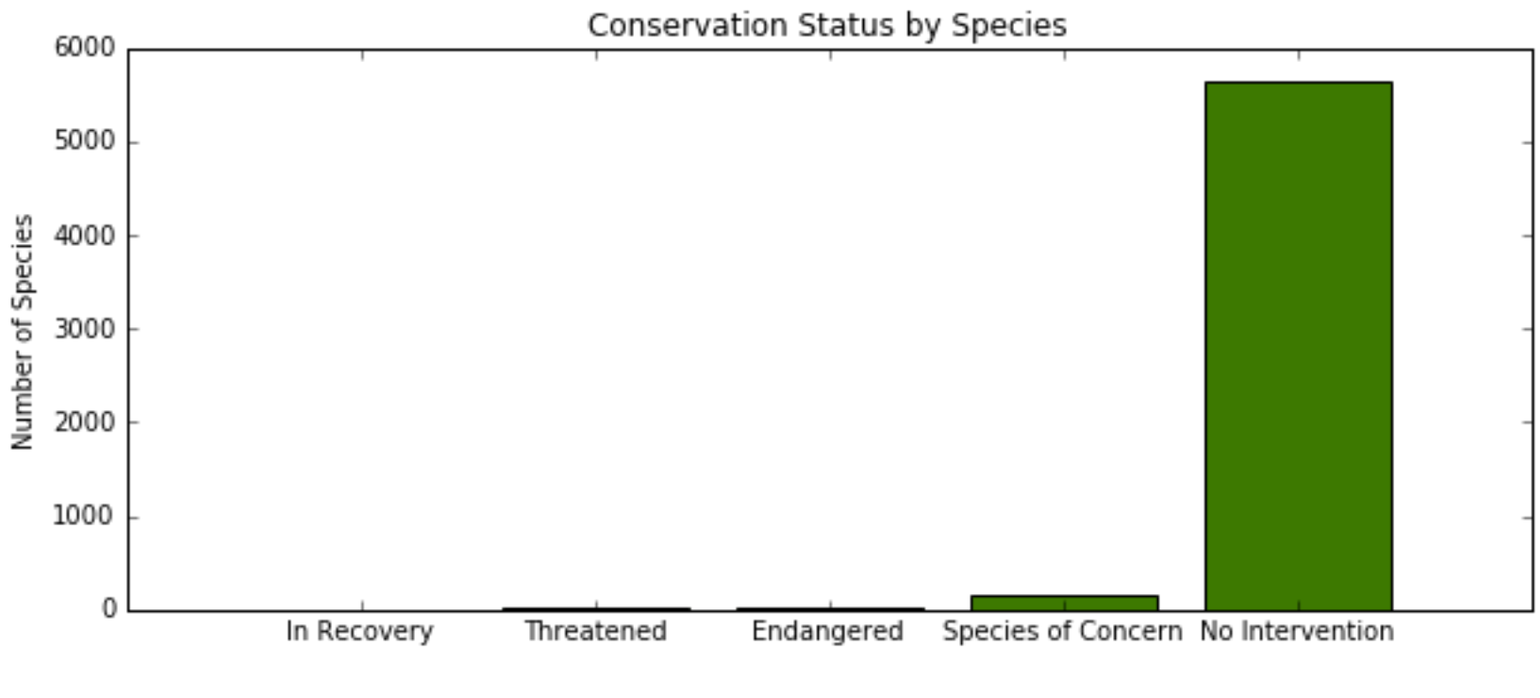
- The default level of significance (90%).

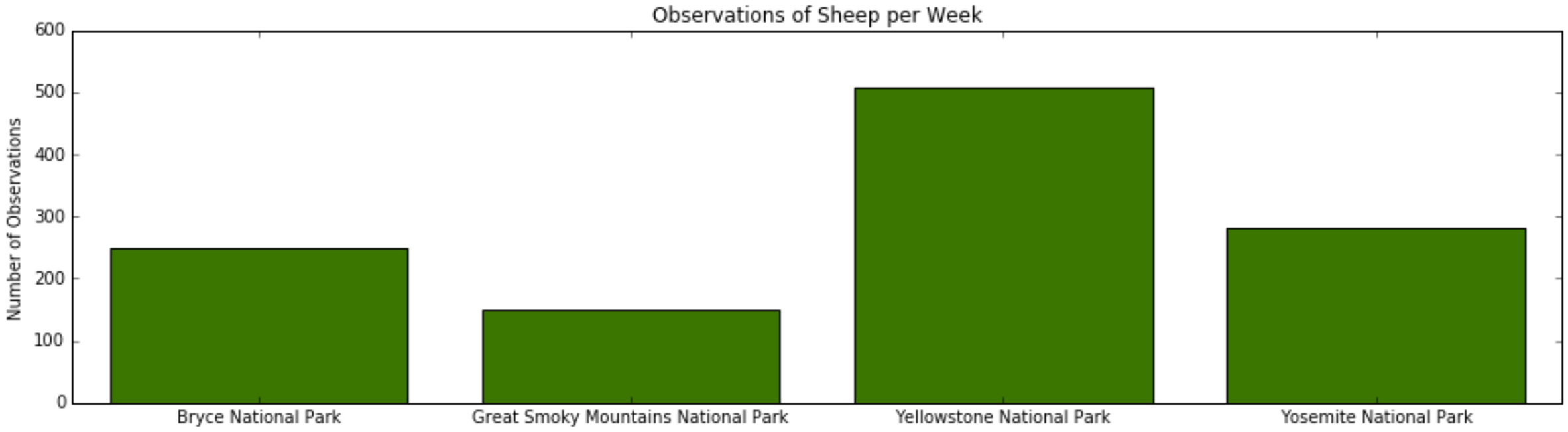
Sample Size Determination

- Baseline = 15
- lift = 100 * 0.05 / 0.15
- level_significance = 90
- sample_size_determination = 520 sheep need to observe from each park.

Graphs

Graphs


Conservation Status by Species
Number of Species
6000
5000
4000
3000
2000
1000
0
In Recovery
Threatened
Endangered
Species of Concern
No Intervention


Observations of Sheep per Week
Number of Observations
600
500
400
300
200
100
0
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park