

Biodiversity for the National Parks

Introduction to Data Analysis Capstone Project

Julian Morones

Washington D.C. August 2018

Index

- Introduction
- Species Conservation Status
- Foot and mouth Reduction Effort

Introduction

Biodiversity is defined as “the variability among living organisms from all sources, including terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species, and of ecosystems.”

Biodiversity is the foundation of ecosystem services to which human well-being is intimately linked. No feature of Earth is more complex, dynamic, and varied than the layer of living organisms that occupy its surfaces and its seas.

As part of its role of preserving the ecological and historical integrity of the places entrusted to its management; the **National Park Service** (NPS) has asked us to help in:

- The analysis of endangered species from several different parks.
- The analysis of recording sightings of sheep and the effectiveness of a program to reduce foot and mouth disease.

Index

➤ Introduction

➤ Species Conservation Status

- Overview
- Exploratory analysis
- Endangered species – Contingency table
- Endangered species – Frequency table
- Endangered species – Chi Square testing

➤ Foot and mouth Reduction Effort

Species Conservation Status

Overview

The purpose of this section is to perform some data analysis on the NPS's `species_info.csv` file relating to the conservation statuses of the different species living in the parks and to investigate if there are any patterns or themes to the types of species that become endangered.

The file contains the following information about 5541* unique species currently living in our parks:

- Category or classification of each species (amphibian, bird, fish, mammal, vascular & nonvascular plants and reptile)
- The scientific name of each species
- The common names of each species
- The species conservation status, mainly:
 - 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.

5 * Database contains a total of 5824 records. 283 of these records are duplicates that contain additional common names by which the species are known

Species Conservation Status

Exploratory analysis (1/2)

A first glance on our data shows that:

- 83% of the biodiversity of our national parks is made up by vascular and nonvascular plants.
- Remaining 17% is made up by animalia.
- Only 3% of species have some type of conservation status.

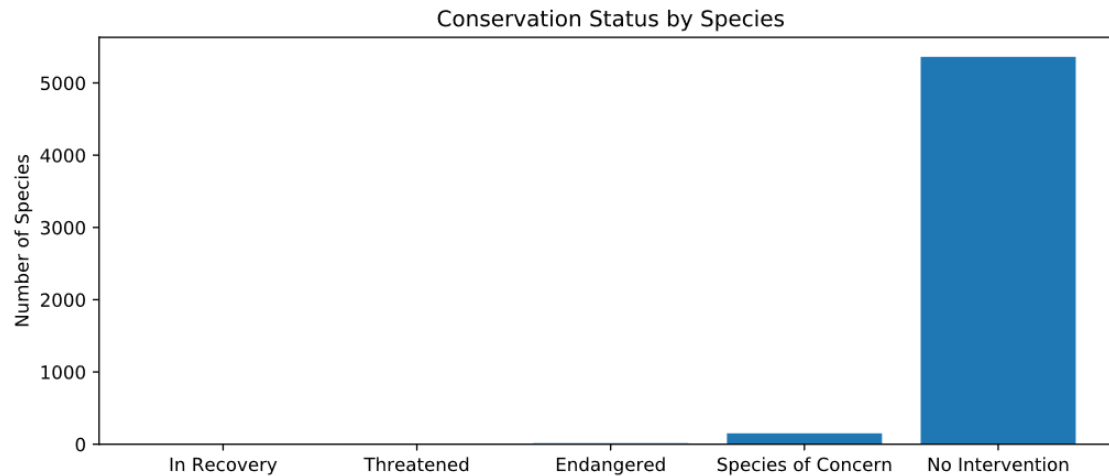
Note: There is a difference between the grand totals for these two variables (5541 vs. 5543) that is unaccounted for!!!

Category	Total	% of total
Amphibian	79	1%
Bird	488	9%
Fish	125	2%
Mammal	176	3%
Reptile	78	1%
Nonvascular Plant	333	6%
Vascular Plant	4262	77%
Grand Total	5541	100%

Conservation status	Total	% of total
Endangered	15	0%
In Recovery	4	0%
Species of Concern	151	3%
Threatened	10	0%
No Intervention	5363	97%
Grand Total	5543	100%

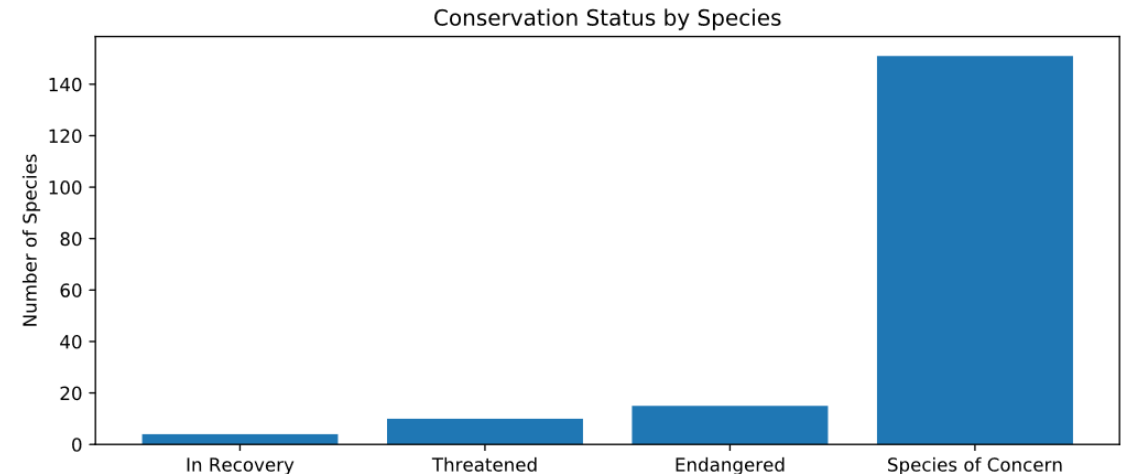
Species Conservation Status

Exploratory analysis (2/2)



Graph 1 shows the distribution of the species among the different conservation status.

Graph 2 also shows the distribution of the species among the different conservation status excluding those with “No Intervention” status.



Species Conservation Status

Endangered species – Contingency table

We are interested in knowing if there are certain types of species more likely to be endangered than others.

- To answer this question let's begin by building the contingency table of our variables `category` and `conservation_status`.
- This table simply displays the (multivariate) frequency distribution of our variables of interest.
- In this form the table is not very useful to make any assumptions. Let's hence construct the corresponding frequency table.

Category	Protected	Not protected	Grand Total
Amphibian	7	72	79
Bird	75	413	488
Fish	11	115	126
Mammal	30	146	176
Reptile	5	73	78
Nonvascular Plant	5	328	333
Vascular Plant	46	4216	4262
Grand Total	179	5363	5542

Species Conservation Status

Endangered species – Frequency table

The frequency table by category is obtained by dividing each row by it's row total.

From here we can observe that:

- Overall, only 3% of the species have a protection status.
- Between categories there seems to be some differences, for example:
 - Birds and mammals appear to be more protected than plants or other animals like reptiles.
 - Overall, plants seem less protected than animals.

Category	Protected	Not protected
Amphibian	9%	91%
Bird	15%	85%
Fish	9%	91%
Mammal	17%	83%
Reptile	6%	94%
Nonvascular Plant	2%	98%
Vascular Plant	1%	99%
Grand Total	3%	97%

Category	Protected	Not protected
Animalia	14%	86%
Plants	1%	99%
Grand Total	3%	97%

Species Conservation Status

Endangered species – Chi Square testing (1/2)

In this section we will use Chi-square testing to test the null hypothesis that there is no significant difference between the species categories. We will reject this hypothesis if the resulting p-value is less than 0.05.

Results are presented in the following table:

	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile
Bird	0.176					
Fish	0.825	0.077				
Mammal	0.128	0.688	0.056			
Nonvascular Plant	0.002	0.000	0.000	0.000		
Reptile	0.781	0.053	0.741	0.038	0.034	
Vascular Plant	0.000	0.000	0.000	0.000	0.662	0.000

From this table we can observe that there are significant differences for all p-values in red. For example, in the case of Amphibians and Nonvascular Plants we can conclude that the difference in endangered species is not due to randomness.

Species Conservation Status

Endangered species – Chi Square testing (2/2)

Finally, we would like to test whether there is a significant difference between animalia categories and the vegetabilia.

Contingency and frequency tables are presented below:

Category	Protected	Not protected
Animalia	128	819
Plants	51	4544
Grand Total	179	5363

Category	Protected	Not protected
Animalia	14%	86%
Plants	1%	99%
Grand Total	3%	97%

Chi-square test returns a p-value of **3.20E-85** which implies that this difference is not due to chance and hence, that it would be logical to conclude that animal species are more endangered than plants or that the NPS puts more effort in protecting the animals.

Index

- Introduction
- Species Conservation Status
- Foot and mouth Reduction Effort
 - Overview
 - Sheep sightings
 - Sample size determination

Foot and mouth Reduction Effort

Overview

Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease of sheep at that park. The goal in this section is to determine the effectiveness of this program. We will be working with a file called `observations.csv` which contains the recordings of different species at several national parks for the past 7 days.

The file contains 23,296 observations, the data been recording is:

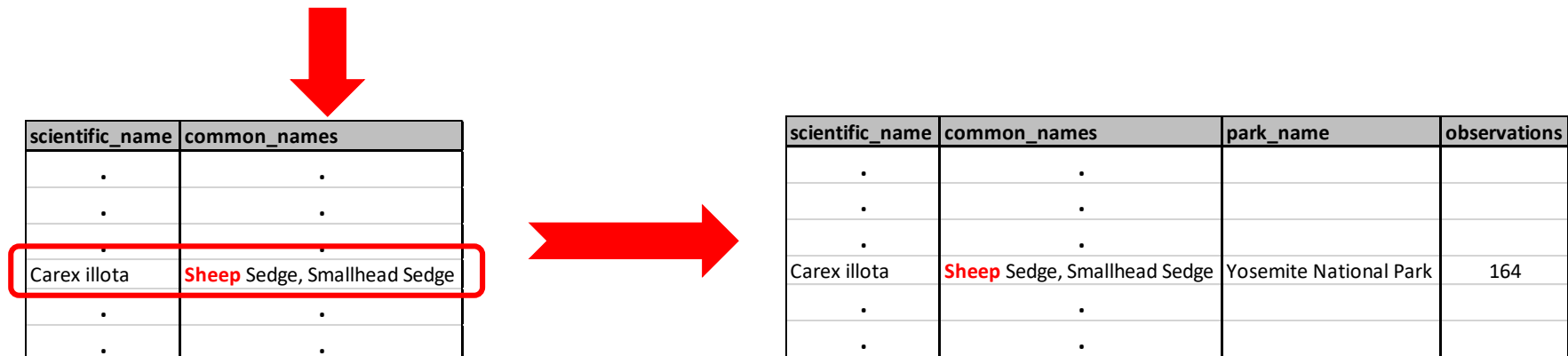
- The scientific name of each species
- The name of the park
 - Bryce National Park
 - Great Smoky Mountains National Park
 - Yellowstone National Park
 - Yosemite National Park
- The total animal sightings

Foot and mouth Reduction Effort

Sheep sightings (1/2)

observation.csv only contains the scientific name of the different species which is not useful to identify the different kind of sheep's. **species_info.csv** on the other hand contains both the scientific and common names for each species.

In order to determine the sheep observations, we will look in **species_info.csv** for the cases where the common name contains the string “sheep” and the species category is mammal. For those cases, we will look for the scientific name in both files and return the name of the park and the total observations from **observation.csv**.

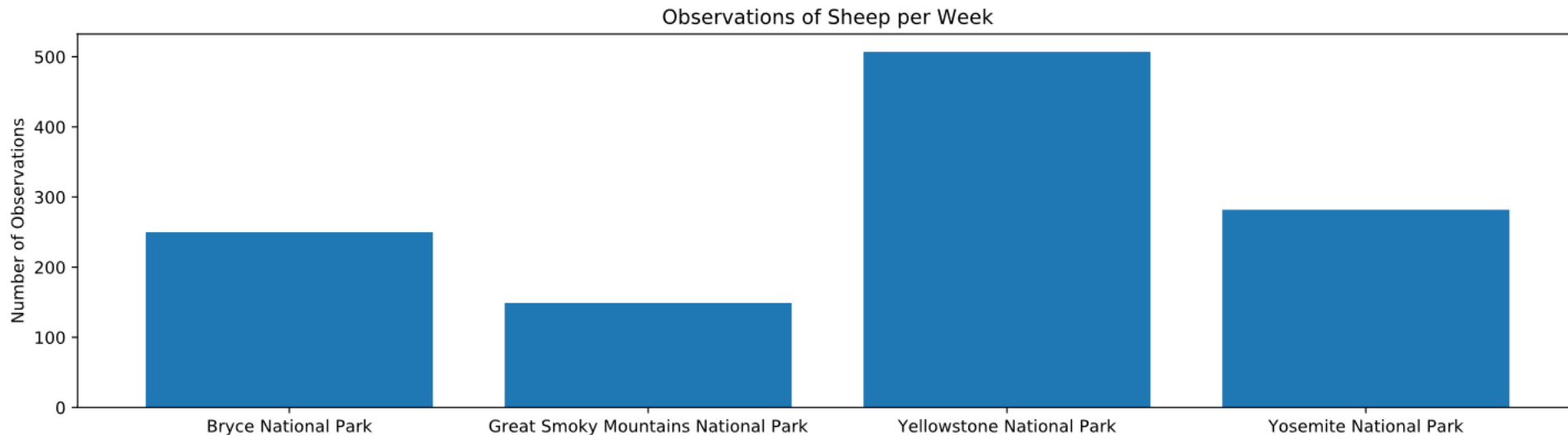


Foot and mouth Reduction Effort

Sheep sightings (2/2)

Now that all relevant data is in one DataFrame; we can have a look at the number of sheep observations per park:

- In the past week, there were 1,188 sheep observations
- Yellowstone registered the max number of observations 507
- Great Smokey Mountains registered the min number of observations 149



Foot and mouth Reduction Effort

Sample Size Determination

The NPS want to test whether or not their foot and mouth program is working. They want to be able to detect reductions of at least 5 percentage points.

The only information we have is that last year it was recorded that 15% of sheep at Bryce National Park had foot and mouth disease. We will need to calculate the number of sheep that we need to observe from each park to make sure their foot and mouth percentages are significant with 90% level of significance.

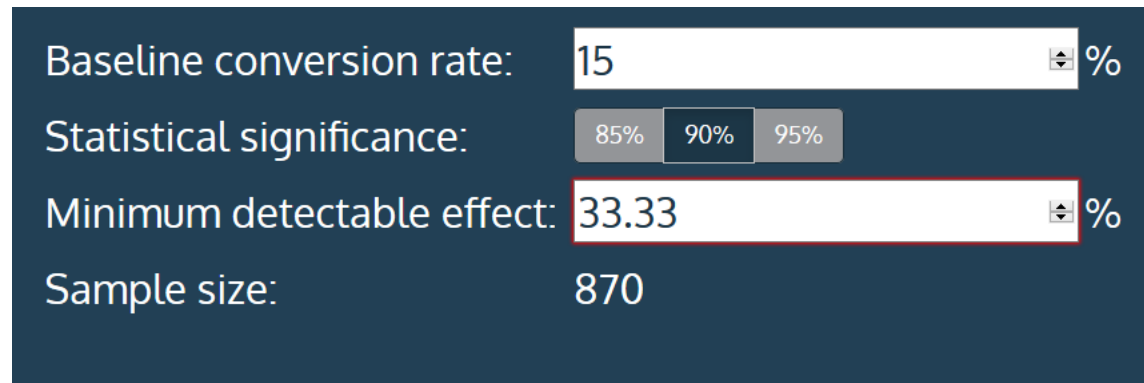
In order to do this we need to determine the following parameters to input in a sample size calculator:

- The Baseline conversion rate => 15
- The Minimum detectable effect => $100 * (5/15)$
- The Statistical significance => 90%

Foot and mouth Reduction Effort

Sample Size Determination (2/2)

With the data from the previous section and Codecademy's sample calculator we estimated that we would need to 870 observations to be able to determine with 90% confidence level whether the foot and mouth effort is effective or not.



A screenshot of a sample size calculator interface with a dark blue background and white text. It features four rows of input fields and a row of buttons. The first row has a label 'Baseline conversion rate:', a text input with '15', and a percentage sign. The second row has a label 'Statistical significance:', three buttons labeled '85%', '90%', and '95%', with '90%' being the selected option. The third row has a label 'Minimum detectable effect:', a text input with '33.33', and a percentage sign. The fourth row has a label 'Sample size:', and a text input with '870'.

Baseline conversion rate:	15	%
Statistical significance:	85% 90% 95%	
Minimum detectable effect:	33.33	%
Sample size:	870	

Since we only registered 507 observations last week, it would take us two weeks to gather the data ($870/507 = 1.71$)