Intro to Data Science Capstone: Biodiversity

Name: Mike Orlando

Username: orla0010

Email: m.orla0010@gmail.com

Data in species_info.csv

The species_info dataframe contains 5,541 species and the 4 columns shown below – all of which are strings.

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	nan
1	Mammal	Bos bison	American Bison, Bison	nan
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

Species categories include

- Mammal
- Bird
- Reptile
- Amphibian
- Fish
- Vascular Plat
- Nonvascular Plant

Each species is assigned a conservation_status:

- Species of concern
- Endangered
- Threatened
- In Recovery
- Null (if not endangered, which is about 96.8% of animals in the dataframe)

Differences in Endangered Status

is_protected	category	False	True
0	Amphibian	72	7
1	Bird	413	75
2	Fish	115	11
3	Mammal	146	30
4	Nonvascular Plant	328	5
5	Reptile	73	5
6	Vascular Plant	4216	46

Two Chi Sqaure tests were conducted to evaluate whether or not the differences in % endangered between species categories are statically significant

- Question: Is there a statistical difference in the proportion of endangered birds vs mammals?
  - Null hypothesis: there is no statistically significant difference in rates of endangerment between birds and mammals
  - Answer: With a p-value of .6876, we cannot reject the null
- Question: Is there a statistical difference in the proportion of endangered mammals vs reptiles?
  - Null hypothesis: there is no statistically significant difference in rates of endangerment between mammals and reptiles
  - Answer: With a p-value of .0384, we can reject the null and conclude that there is a statistically significant difference in the proportion of endangered mammals compared to reptiles

Recommendation to conservationists based on hypothesis tests

- Although we can calculate the percent endangered of each species category, we cannot use these data points alone to make strategic decisions.
  - We first need to understand if the differences we are seeing could be the result of sample size
- For example, we can conclude that mammals are statistically significantly more endangered than reptiles – and thus may want to prioritize our efforts accordingly
  - But, we cannot conclude that birds are more likely than mammals to be endangered

Next Steps:

- Continue to use hypothesis testing to disprove or validate the significance of what the conservationists believe are the important points the data is indicating
- I would suggest evaluating by species category and then any notable category differences within specific parks

Foot and Mouth Disease Study

Sample Size Determination

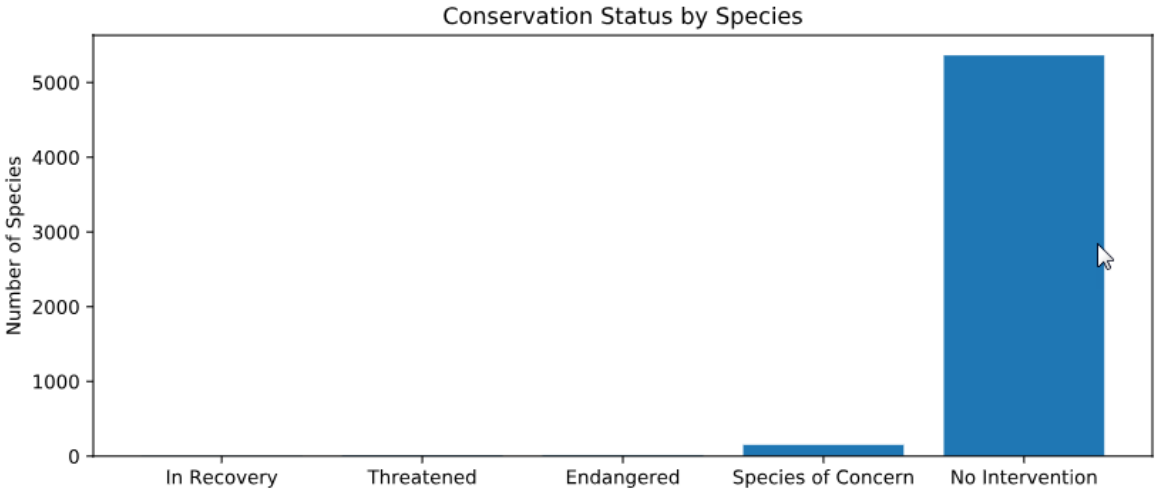
To ensure the foot and mouth sheep study is representative of the population, we calculated a required sample size of **870** based on:

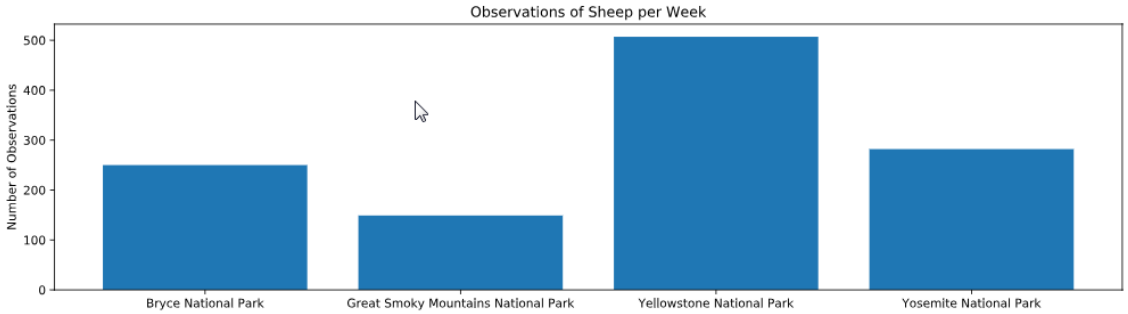
- Last year's baseline rate of 15%
- A desire to detect reductions of at least 5 percentage points
- 90% level of significance

Based on our observation data, this study is likely to take:

- 1.72 weeks in Yellowstone
- 3.48 weeks in Bryce

Graphs


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


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