Biodiversity for the National Parks

maria.morales@minted.com

Introduction

About the Data

The data used for this analysis can be found in **species_info.csv**. It lists the observed species, categorized by their different classes (mammal, reptile, vascular plant, etc.) that are found in national parks across America. Both the scientific name and the common name are included for each species.

Most importantly, the data specifies whether or not that species is protected, under “conservation_status”.

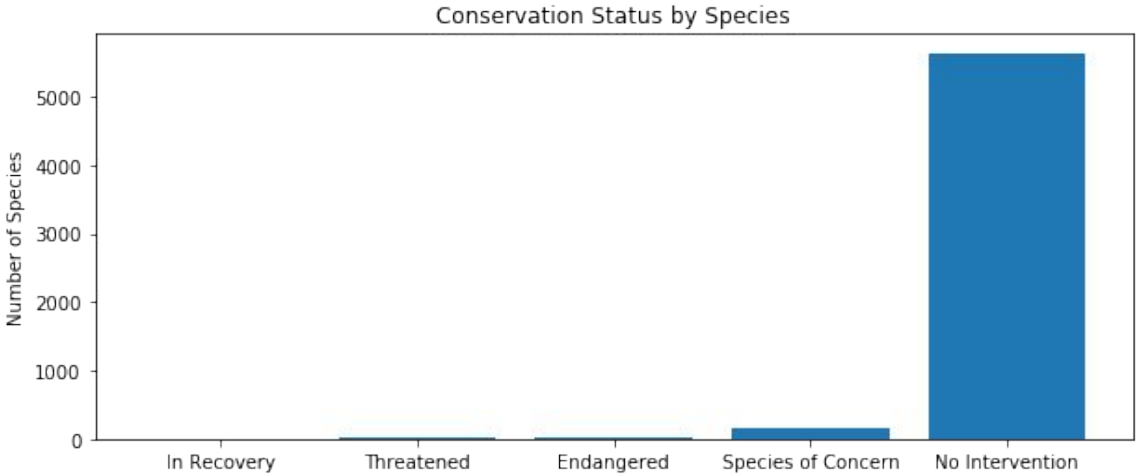
About the Data: Conservation Status

There are four possible values for **conservation_status:**

- No Intervention
- Species of Concern
- Threatened
- Endangered
- In Recovery

About the Data: Conservation Status

Thankfully, the vast majority of the species analyzed had a conservation status of “No Intervention.”



Significance Calculations

Manipulating the Data

We added a boolean variable, **is_protected**, and pivoted the data on that column. From there, it was easy to calculate the percentage of each species that was protected.

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.097222
1	Bird	413	75	0.181598
2	Fish	115	11	0.095652
3	Mammal	146	30	0.205479
4	Nonvascular Plant	328	5	0.015244
5	Reptile	73	5	0.068493
6	Vascular Plant	4216	46	0.010911

Contingencies and the Chi-squared Test

Based on the pivot table, one could easily assume that mammals are more likely to be endangered than birds, but is this true?

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.097222
1	Bird	413	75	0.181598
2	Fish	115	11	0.095652
3	Mammal	146	30	0.205479
4	Nonvascular Plant	328	5	0.015244
5	Reptile	73	5	0.068493
6	Vascular Plant	4216	46	0.010911

Contingencies and the Chi-squared Test

No. The p-value that was returned after running the chi-squared test was 0.69.

The difference is not significant.

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.097222
1	Bird	413	75	0.181598
2	Fish	115	11	0.095652
3	Mammal	146	30	0.205479
4	Nonvascular Plant	328	5	0.015244
5	Reptile	73	5	0.068493
6	Vascular Plant	4216	46	0.010911

Contingencies and the Chi-squared Test

However, running the same test on mammals versus reptiles shows returns a p-value of 0.038, which shows there **is** a significant difference; **mammals are more likely to be endangered than reptiles.**

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.097222
1	Bird	413	75	0.181598
2	Fish	115	11	0.095652
3	Mammal	146	30	0.205479
4	Nonvascular Plant	328	5	0.015244
5	Reptile	73	5	0.068493
6	Vascular Plant	4216	46	0.010911

Recommendation

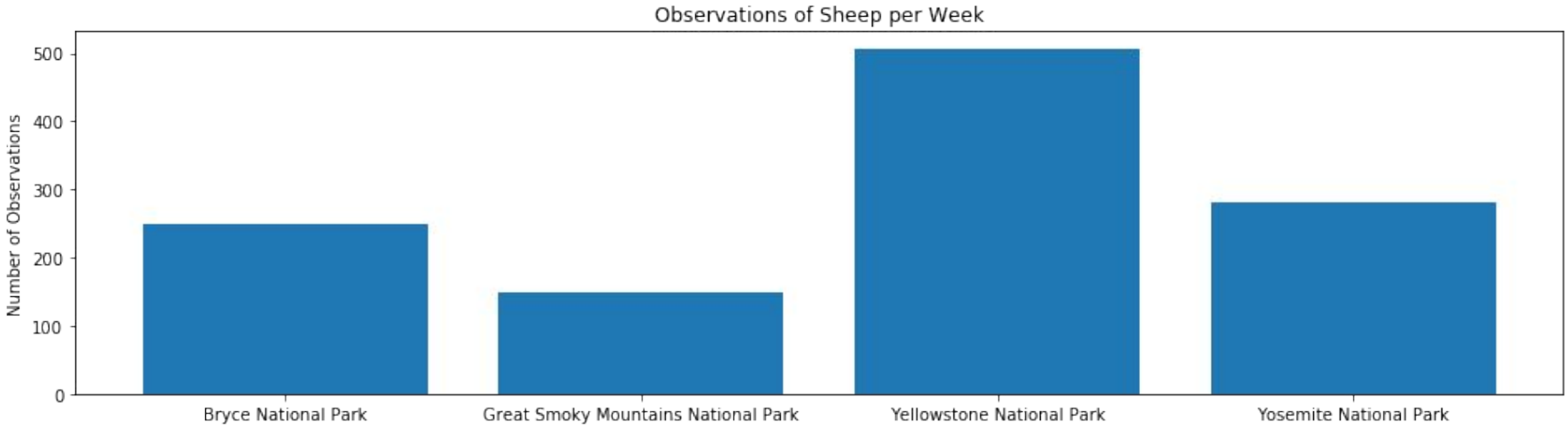
Some Advice for Conservationists

Comparing broad categorical data can trap analysts into making false hypotheses. Use the correct hypothesis test to based on the particular scenario to ensure any findings are statistically significant.

Sample Size Determination

Calculating the Sample Size

Conservations have been recording sheep sightings across several national parks across America for a week.



Calculating the Sample Size: Minimum Detectable Effect

Using **Optimizely**'s sample size calculator, we calculated the number of sheep necessary to observe from each park.

First, we need to determine the **Minimum Detectable Effect**, assuming the baseline of 15% of sheep having foot and mouth disease.

```
baseline = 0.15
minimum_detectable_effect = (100 * 0.05) / baseline
```

33.3%

Minimum Detectable Effect

Calculating the Sample Size: Length of Observation

We also calculated the number of weeks necessary to observe enough sheep at both Bryce and Yellowstone National Parks.

```
sample_size_per_variation = 520

yellowstone_observation_weeks = sample_size_per_variation / 507
bryce_observation_weeks = sample_size_per_variation / 250
```

1 ~ 2 weeks

Conservationists would need to observe sheep at Yellowstone for approximately one (1) week, and approximately two (2) weeks at Bryce.