Investigating Protected Species in US National Parks

Codecademy Capstone Project: Introduction to Data Analysis

March 26, 2018

Outline of Presentation

- **Endangered Species Analysis**
  - Description of National Parks species data
  - Analysis of Endangered Species
  - Observations of Analysis
  - Recommendations
- **Foot and Mouth Disease Study**
  - Overview
  - Sample Size Determination
  - Results
- **Annex: Chi-square Testing**

Description of National Parks Data

- The database has information on the conservation status of thousands of species across seven broad categories of species
- An excerpt of the database is shown below:

Category	Scientific Name	Common Name(s)	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), Domesticated Cattle	nan
Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan

Source: National Parks Service

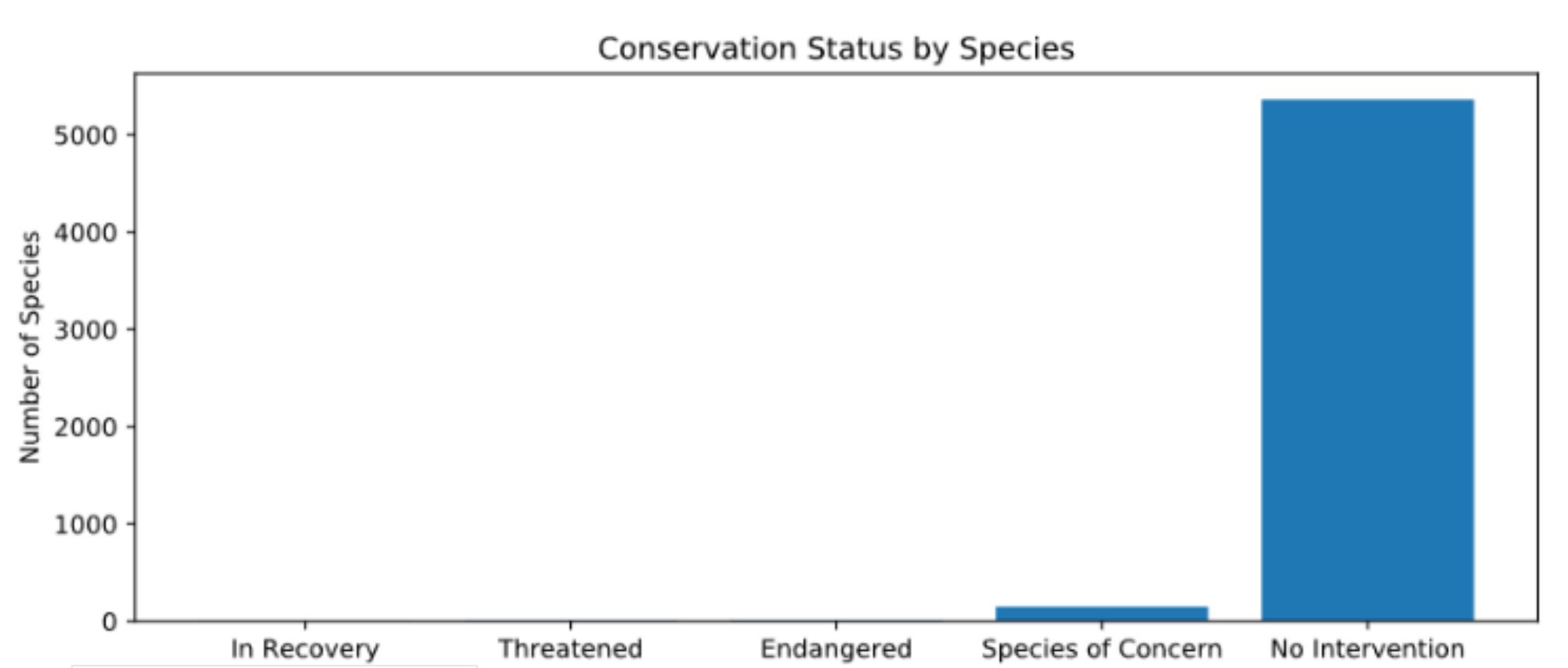
Description of National Parks Data

- There are a total of 5,541 unique species in the database
  - The species types covered in the database are categorized as:
    - Mammals
    - Birds
    - Reptiles
    - Amphibians
    - Fish
    - Vascular Plants
    - Nonvascular Plants

Description of National Parks Data

- Species are given one of five conservation statuses:
  - ***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.
  - ***No Intervention (nan)***: No intervention necessary at this time.
- Analysis of the data follows

Analysis of Endangered Species



Source: National Parks Service

As shown above, the majority of species tracked in the database do not require intervention at this stage. Nonetheless, there are a sizeable number of species (179) classified as endangered to some degree.

Analysis of Endangered Species

- Of those 179 species, it is helpful to know which categories they are and test whether certain species are more likely to be endangered.
  - The table below breaks down each species by category and conservation status:

Category	Not Protected	Protected	% Protected
Amphibian	72	7	8.9%
Bird	413	75	15.4%
Fish	115	11	8.7%
Mammal	146	30	17.0%
Nonvascular Plant	328	5	1.5%
Reptile	73	5	6.4%
Vascular Plant	4,216	46	1.1%

Source: National Parks Service

Analysis of Endangered Species

- Using a chi-squared test, we can determine if a certain type of species is more likely to be endangered than others.
  - For example, the percentage of birds protected (15.4%) and mammals protected (17.0%) appears similar. But are mammals more likely to be endangered than birds?
  - Under a chi-squared test, our Null Hypothesis is that this difference is due to chance.
  - Our Alternative Hypothesis is that the difference is significant and that mammals are more likely to be endangered.
  - The results of a chi-squared test (~0.688) is greater than 0.05, our preferred level of significance. We can conclude that the different percentages (15.4% vs. 17.0%) is not significant and a result of chance.
  - Mammals are statistically not more likely to be endangered than birds.

Analysis of Endangered Species

(% is endangered rate for each species)	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile	Vascular Plant
Amphibian (8.9%)		0.176	0.825	0.128	0.002	0.781	0.000
Bird (15.4%)			0.077	0.688	0.000	0.053	0.000
Fish (8.7%)				0.038	0.000	0.741	0.000
Mammal (17.0%)					0.000	0.038	0.000
Nonvascular Plant (1.5%)						0.034	0.662
Reptile (6.4%)							0.000
Vascular Plant (1.1%)							

Source: Derived from analysis of National Parks Service data.
Green highlights indicate significance at the 95% level. Yellow highlights indicate significance at the 90% level.

- The table above presents results for all chi-squared tests across all species-category pairs. Let us compare reptiles (6.4%) against mammals (17.0%). Is this difference due to chance?
- The p-value from the chi-squared test is 0.038, lower than our level of significance of 0.05. Hence, we can conclude this is significant, and that mammals are statistically more likely to be endangered than reptiles (the difference is not due to chance).
- All results in green indicate that differences in endangered rates are not due to chance.

Observations of Endangered Species Analysis

- Mammals are statistically more likely to be endangered than fish.
- Mammals are also statistically more likely to be endangered than reptiles.
- Although not statistically significant, birds are more likely to be endangered than both fish and reptiles
  - A p-value of 5% (indicating 95% confidence in our results) is the statistically significant cut-off point. However, birds compared to fish and reptiles have a p-value of less than 10% (90% significance), indicating that we should take note of this observation.
- Plants, both vascular and nonvascular, are statistically less likely to be endangered than all other species categories

Recommendations

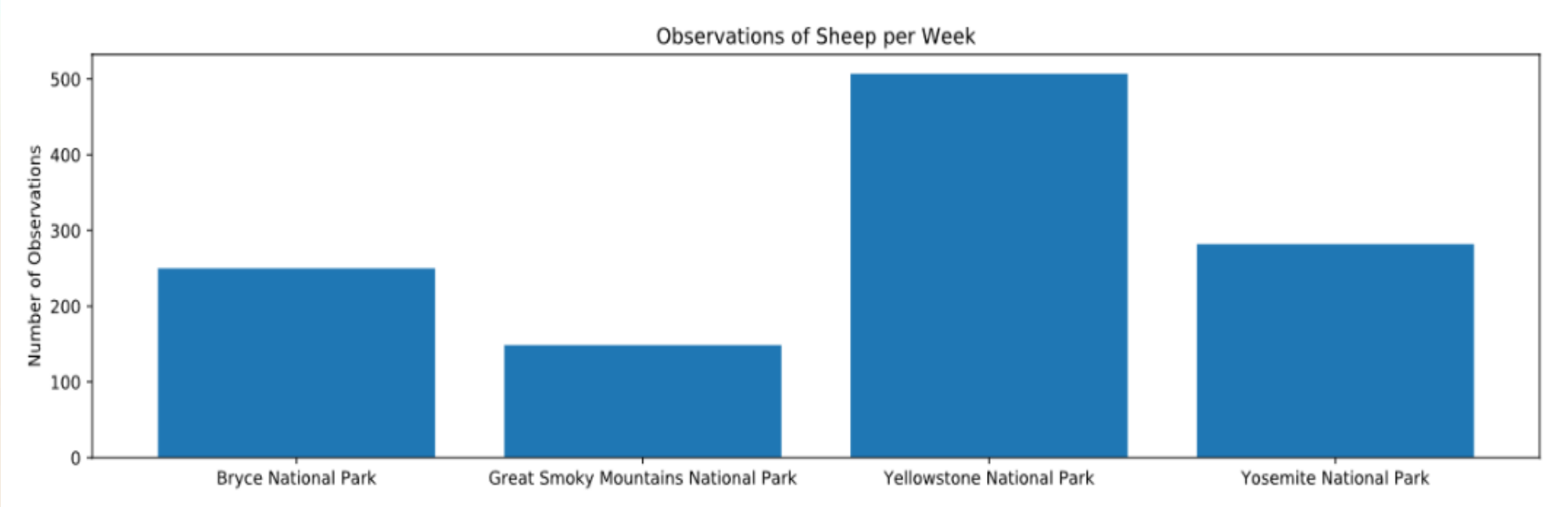
- Mammals are far more likely to be endangered than other species.
  - Resources should be increased to ensure conservation of this species.
- Birds are also more likely to be endangered than other species, albeit it not to the same degree as mammals.
  - Efforts should be maintained to ensure conservation of endangered breeds within this species.
- At this moment, amphibians, fish and reptiles do not appear to have a high likelihood of being endangered.
  - Current conservation efforts can be maintained.
- Plants, both vascular and nonvascular have a low likelihood of being endangered.
  - Current conservation efforts can be maintained.

Foot and Mouth Disease Study: Overview

- At Yellowstone National Park, Park Rangers have initiated a program to reduce the rate of foot and mouth disease amongst sheep at that park.
- The stated aim of this program is to detect reductions of at least 5 percentage points in foot and mouth disease incidence in sheep.
  - Current information is that 15% of sheep at Bryce National Park have foot and mouth disease.
- To test if the program is working, we need to take samples from parks where sheep are known to reside and test if there is a significant difference in the incidence of the disease between parks.

Sample Size Determination

- Over the last week, conservationists have recorded sightings of different species at several national parks.
- From this database, we have counts of the number of sheep observed at four different parks in the last week.



Source: National Parks Service

Sample Size Determination

- To determine if the Yellowstone National Park program is working, we need to take samples from each of the four parks where sheep have been observed.
- As we are comparing between subpopulations (of sheep) in our sample, the best way to determine the required sample size from each park is to use a A/B test.
- This type of test requires three variables:
  - Baseline conversion rate
  - Minimum detectable effect
  - Statistical significance

Sample Size Determination

- ***Baseline conversion rate***: this is a percentage of subjects who exhibit a measurable effect. For this program, it is the % of sheep with foot and mouth disease.
  - Based on historical data, we know that 15% of sheep at Bryce National Park have foot and mouth disease. Our baseline is 15%
- ***Minimum detectable effect*** (MDE): this is the smallest difference we need to measure to determine if the program is working. It is expressed as a percent of the baseline conversion rate.
  - For this study, the scientists wish to determine if a 5% drop in observed cases of foot and mouth disease in sheep was significant. Our MDE is hence (5%/15%), or 33.33%.
- ***Statistical significance***: this represents the degree of confidence we have in our results.
  - For this study, we are using the default significance level of 90%. This means that, for multiple repetitions of the survey, we can establish that any observed reduction of foot and mouth disease will be significant (not due to chance) 90% of the time.

Results

- Based on these variables, the minimum sample size needed to ensure a 5% or more drop in observed cases of foot and mouth disease in sheep at Yellowstone National Park is significant is **at least 510 sheep**.
  - Based on data of the number of observations of sheep at various parks, we can determine it would take approximately 1 week to obtain this sample at Yellowstone National Park. Full results are below.

Park Name	# Observations (last 7 days)	Estimated time to complete survey
Bryce National Park	250	2 weeks
Great Smoky Mountains National Park	149	3.5 weeks
Yellowstone National Park	507	1 week
Yosemite National Park	282	1.8 weeks

Source: National Parks Service

Annex: Chi-Square Testing

- For datasets where we have more than two discrete categories of data, binomial testing is not appropriate; chi-square becomes the preferred test to determine if differences are significant.
  - For the National Parks dataset analyzed here, our discrete categories were species, protected status and not protected status.
  - A chi-square test was hence used to determine if one species was more likely to be endangered than another.
- A chi-square test requires our data to be arranged as a contingency table, where columns represent different outcomes (endangered vs. not endangered) and rows represent different conditions (species 1 vs. species 2).

Annex: Chi-Square Testing

- For example, to test if mammals are more likely to be endangered than reptiles, the contingency table would be:

	Protected	Not Protected
Mammals	30	146
Reptiles	5	73

- Our Null Hypothesis is that differences in rates of endangerment are due to chance
  - The alternative is that the difference is significant and one species (mammals) is more likely to be endangered than another (reptiles)
- A chi-squared test on this data (all calculations in this report utilized Scipy) yields a p-value of 0.038
  - A p-value of 0.05 (95% significance) is chosen as our cut-off point for significance. Our calculated p-value is lower than 0.05 – as such we can reject the Null and conclude that mammals are more likely to be endangered than reptiles.