Biodiversity in U.S. National Parks

A Codecademy Pro Data Analysis Capstone Project

Abstract: Data for 5,541 different species found in American National Parks was provided. Of those, 96.7% require no intervention to preserve the species population; 2.7% are species of concern, 0.25% are endangered, 0.16% are threatened, and the remainder are in recovery. Mammal species are more likely to fall into a protected category than are reptile species, fish species, or plant species; birds are marginally more likely to fall into a protected category than those same species types. Vascular and non-vascular plants have the lowest percentage of species in a protected category.

In terms of an ongoing study of Foot and Mouth Disease in sheep species found in American National Parks, observations from 4 different National Parks were compiled and analyzed to determine the sample size needed to conclude if the intervention by rangers at Yellowstone National Park is effective at reducing disease rates. I conclude that 870 observations will be required (at the 90% confidence interval), and report the predicted number of weeks it will take the 4 national parks to compile the needed number of observations.

Note: It was suggested by Codecademy that the results of this Capstone Project be summarized via a series of Microsoft PowerPoint slides; however, this seems to be the default suggestion of almost every Capstone Project they devise, and since I know that writing reports will be an integral part of the work of any data analyst/data scientist, I decided to summarize my results in written form instead.

Introduction and Context

The U.S. National Parks Service, established in 1916, is charged with preserving access to the natural and cultural resources of the National Park System[1]. While the National Park Service oversees a wide variety of areas, ranging from well-known parks like Yellowstone National Park and Yosemite National Park to National Battlefields such as the Fort Necessity National Battlefield and National Historic Sites such as the Brown V. Board of Education National Historic Site, this project focuses specifically on National Parks and the wildlife found therein. Many bird, mammal, reptile, and amphibian species call America's National Parks home, as do numerous plant species; this great biodiversity makes National Parks an active area for scientific investigation[2]. However, it also means that some species of flora and/or fauna in America's National Parks fall under the protection of the Endangered Species Act of 1973. This act provides for the conservation of threatened and endangered species, as well as the habitats in which they are found[3]. Clearly, rangers in Americas National Parks must comply with the provisions of the Endangered Species Act to ensure the conservation of protected species, which may involve resources in the form of ranger labor and structures meant to limit visitors' access to certain wildlife habitats, among other things.

The aims of this project are to answer two broad questions:

1. Are certain types of species more likely to need protection than others?
2. How many observations are required to demonstrate the effectiveness of a Yellowstone National Park program to reduce Foot and Mouth Disease rates in native sheep populations by 33%?

This project used data developed by Codecademy for use in one of their data analysis Capstone Projects. As they explain, the data they provided is mostly fictional, but was inspired by real data. Since the two provided data tables were generated by Codecademy, I believe them to be Codecaedmy's intellectual property, so I cannot provide them as a supplement to this report. However, I can provide the code I used to analyze the data as a supplement.

[1] https://www.nps.gov/aboutus/index.htm

[2] https://www.nps.gov/nature/science-in-parks.htm

[3] https://www.epa.gov/laws-regulations/summary-endangered-species-act#:~:text=(1973),in%20which%20they%20are%20found.&text=The%20law%20also%20prohibits%20any,of%20endangered%20fish%20or%20wildlife.

Conservation Status of Different Types of Species

The first question I investigated is whether or not certain types of species are more likely to be in a protected category than others. Knowing which types of species are more likely to be endangered, threatened, or a species of concern can help conservation program organizers decide where time, money, and resources can be best spent to achieve their goal of preserving these species. To this end, data for different species found in various American National Parks, in the form of a .csv data set entitled "species_info," was cleaned and analyzed.

The data set contained 5,824 entries, which corresponded to 5,541 unique species (as listed by scientific name). Clearly, there are duplicate entries for many species, a point to which I will return below. These unique species were categorized into 7 broad groups: mammal, bird, reptile, amphibian, fish, vascular plant, and non-vascular plant. Ultimately, we wish to know if any of these species types is more likely than others to contain endangered or threatened species, or species of concern. But first, I needed some idea of how many species in the data base fell into an active conservation category, and how many required no intervention to maintain the current species population. By counting the number of unique species name assigned to each conservation status (including "No Intervention"), the following results were obtained:

Conservation Status	Number of Species
In Recovery	4
Threatened	9
Endangered	14
Species of Concern	151
No Intervention	5.363

Table 1: The number of National Park species assigned to each conservation status.

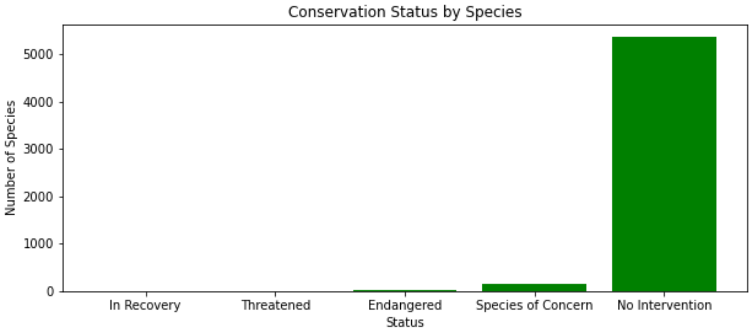

Figure 1: The number of National Park species assigned to each conservation status.

Two things were noted in the compilation of the above data. First, the species *Canis lupus* (i.e., the gray wolf), was listed twice, once under "Endangered" and once under "In Recovery." According to a Google search conducted on 07/23/2020, it should be listed as In Recovery (Least Concern, population stable), and so the "Endangered" entry for this species was removed from the data set. Also, the species *Oncorhynchus mykiss* (i.e., rainbow trout) was listed twice, under both "No Intervention" and "Threatened." I ran a Google search on 07/23/2020 for rainbow trout, and according to the US Fish and

Wildlife service the species is not listed in the Endangered Species Act. I took this to mean "No Intervention," and so the “Threatened” entry for this species was removed from the data set.

I next counted how many species of each type were protected, and how many required no intervention. For the purposes of this analysis, “protected” means having species status Endangered, Threatened, Species of Concern, or In Recovery. The results are included below as Table 2.

Species Category	Number of Protected Species	Number of Species needing No Intervention	Percent of species in a Protected category
Amphibian	7	72	8.86
Reptile	5	73	6.41
Fish	10	115	8.00
Mammal	30	146	17.05
Non-Vascular Plant	5	328	1.50
Bird	75	413	15.37
Vascular Plant	46	4216	1.08

Table 2: Number of protected species, and of species needing no intervention, for each type of species.

We can clearly see that birds and mammals are the two species categories having the highest percentage of species in a protected category. Does this mean that birds and/or mammals are in fact more likely than other species to be in a protected category?

To investigate this, a number of chi-squared tests were run, with the null hypothesis being that there is no difference between species type in terms of likelihood of having species in a protected category. First, I tested to see if there was a significant difference between mammals and birds. There was not, as the test resulted in a p-value of 0.687. Next, a series of chi-squared tests were run on birds and all other non-mammal species categories, pair-wise; birds were significantly more likely to have protected species than vascular plants and non-vascular plants, with each test resulting in a p-value of less than $2 \cdot 10^{-10}$. The chi-squared tests run between birds and reptiles and birds and fish resulted in p-values of 0.053 and 0.047, respectively, and I do not consider either p-value as indicating a significant difference. The chi-squared test run between birds and amphibians resulted in a p-value of 0.176, clearly indicating no significant difference. Finally, I ran a series of chi-squared tests between mammals and all other non-bird species categories, pairwise; mammals were significantly more likely to have protected species than vascular plants (p-value $1.48 \cdot 10^{-10}$), non-vascular plants (p-value $1.44 \cdot 10^{-55}$), reptiles (p-value 0.038) and fish (p-value 0.035). In fact, the only pair-wise chi-squared test that did *not* reveal a significant difference was between mammals and amphibians, as the test yielded a p-value of 0.128. (Of course, the previously run chi-squared test between birds and mammals also did not reveal a significant difference.)

I conclude from this that, when comparing between various species types, mammal species are more likely than all but bird species and amphibian species to be in a protected category. Hence, given limited time and resources devoted to conservation efforts, it makes sense to plan on needing to intervene for mammalian species. Birds are much more likely to be in a protected category than any plant species, and marginally more likely to be protected than reptiles or fish. Finally, while I did not run chi-squared tests for the following statement explicitly, it seems reasonable to assume that vascular and non-vascular plants are less likely than other species types to be in a protected category.

Reducing the Rate of Foot and Mouth Disease in sheep at American National Parks

Foot and Mouth Disease is a serious disease in bovid mammals, resulting in high fever and blisters whose rupturing may cause lameness in infected animals[4]. It is highly infectious, and sometimes outright fatal[5]. Thus, any outbreak of Foot and Mouth disease among bovid wildlife in National Parks should be quickly addressed. In the fictional data set provided, it is known that 15% of sheep (across all sheep species) at Bryce National Park have Foot and Mouth Disease. I assume that this rate is similar, if not identical, for the Great Smoky Mountains National Park, Yellowstone National Park, and Yosemite National Park. I was also told that the fictional park rangers at Yellowstone have undertaken a program aimed at reducing the rate of Foot and Mouth Disease in that park. The task at hand was to see if the program was working, with "working" operationally defined as being a 5 percentage point reduction in the rate of Foot and Mouth Disease at a park (i.e., 10% instead of 15%). Given the nice numbers involved, it is clear that this operational definition of "working" translates to a 33% reduction in cases (that is to say, a reduction of 5/15, or 1/3).

With this information, it is possible to calculate the number of observations of sheep species in a National Park needed to determine if the Foot and Mouth Disease reduction program is working, according to the above definition of "working." Here, I am assuming that "observation" goes beyond just recording whether or not sheep were seen in a park, and also entails some investigation into whether or not the sheep/herd of sheep observed was showing signs of Foot and Mouth infection. With this assumption in mind, relevant parameters were entered into Codecademy's sample size calculator[6]: baseline of 15%, "lift" of 33.33%, and a significance level of 90%. The number of required observations was thus determined to be 870.

It is a fair question to ask how long it would take rangers at a given National Park to compile 870 observations of sheep in their park. To that end, fictional data for 7 days' worth of observations of various wildlife species at four National Parks was provided in the form of a .csv file named "observations.csv." Before merging this data set with the data provided in "species_info.csv," I selected from the latter only species which had the word "sheep" in their common name, with further filtering needed to ensure that species of vascular plant were eliminated from the data. This yielded three specific mammalian sheep

[4] Source: https://en.wikipedia.org/wiki/Foot-and-mouth_disease

[5] Ibid.

[6] https://s3.amazonaws.com/codecademy-content/courses/learn-hypothesis-testing/a_b_sample_size/index.html

species: common sheep (or mouflon), bighorn sheep, and Sierra Nevada bighorn sheep. This sheep-limited data was merged with the “observation.csv” data set to select only those observations which were of mammalian sheep species. From there, the number of observations of sheep in each of the 4 National Parks under consideration was determined. These figures are included below.

Park Name	Number of Sheep Observations
Great Smoky Mountains	149
Bryce	250
Yosemite	282
Yellowstone	507

Table 3: The number observations of sheep in each National Park. The words “National Park” have been omitted from park names for brevity.

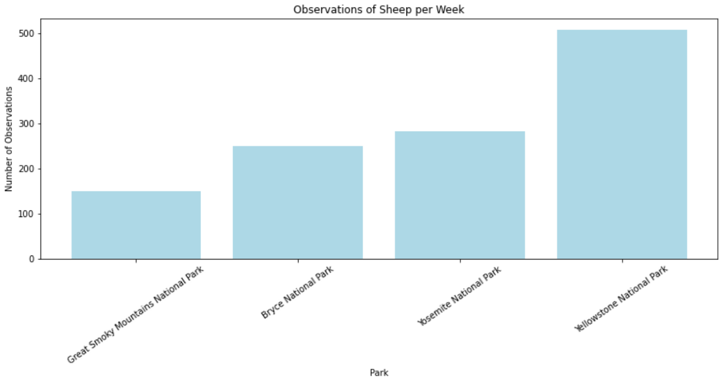

Figure 2: The number of sheep observations in each National Park.

Now that the number of observations per 7 day period of sheep in a park is known, we can determine how many weeks it will take to compile the requisite 870 observations needed to determine if the Foot and Mouth Disease reduction program is working. As is apparent from the above data table, it will take 5.8 weeks in the Great Smoky Mountains National Park, 3.5 weeks in Bryce National Park, 3.1 weeks in Yosemite National Park, and 1.7 weeks in Yellowstone National Park. It is a good thing that the program was started in Yellowstone, as it will not take long for park rangers there to determine if their Foot and Mouth Disease reduction program is working.

Conclusions

I determined that mammal species are more likely than reptiles, fish, vascular plants, and non-vascular plants to fall into a protected category; bird species are marginally more likely fall into a protected category than reptiles or fish, and definitely more likely to be protected than plant species. This has implications for conservation programs, which must determine how to best protect endangered species, threatened species, and species of concern given limited time, money, and resources. It is likely that their work will involve protecting mammal species, so conservation programs should plan on allocating more resources to mammal care and preservation than other species types. Very few resources should be devoted to plant protection, and bird care should be the next highest prioritization of resources.

Given the current (fictional) rate of Foot and Mouth Disease at Bryce National Park, 870 observations of native sheep species are required to see if a disease reduction program has successfully reduced the rate of infection from 15% to 10%. Given the current (fictional) rate of mammalian sheep observations at the 4 parks considered, it will take over a month to compile the necessary observations in the Great Smoky Mountains National Park, but just shy of 2 weeks in Yellowstone National Park.