Endangered Species Analysis by National Park

By Christopher J Routley
Data Analysis Contractor for National Parks Service

Data Provided: species_info.csv

After loading and displaying the information I was able to see that the file contained 4 distinct columns

1. Category
2. Scientific Name
3. Common Name
4. Conservation Status

	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), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN

Data Provided: species_info.csv Cont'd

First it's important to see what these columns contain; focusing on relevant data Categories, Scientific Name and Conservation Status :

```
species.scientific_name.nunique()
```

```
5541
```

Tells me the total number of unique Scientific Names

Tells me the Categories used to Classify each Species

```
species.category.unique()
```

```
array(['Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant',
       'Nonvascular Plant'], dtype=object)
```

```
species.conservation_status.unique()
```

```
array([nan, 'Species of Concern', 'Endangered', 'Threatened',
       'In Recovery'], dtype=object)
```

Tells me the Conservation Status of Each Species

Questions we'd like to Answer through Analysis:

How many species are in each Conservation Status, separated by Category of Species?

a. I separated by Category of Species because to look at each Species individually would be too cumbersome and nearly impossible to determine any information of significance.

2. **Are certain Categories of Species more likely to be Endangered?**

Let’s look at Question # 1:

How many species are in each Conservation Status, separated by Category of Species?

First lets create a new table to display this data:

```
species.groupby('conservation_status').scientific_name.nunique().reset_index()
```

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

As you can see there are only 180 Species represented here. We know that there are over 5000 Species in total. So what happened to those other Species?

If we look at the first table we can see that there are some Species with Conservation Status of NaN. What does this mean? It means that there are some species that are not in any kind of protective status.

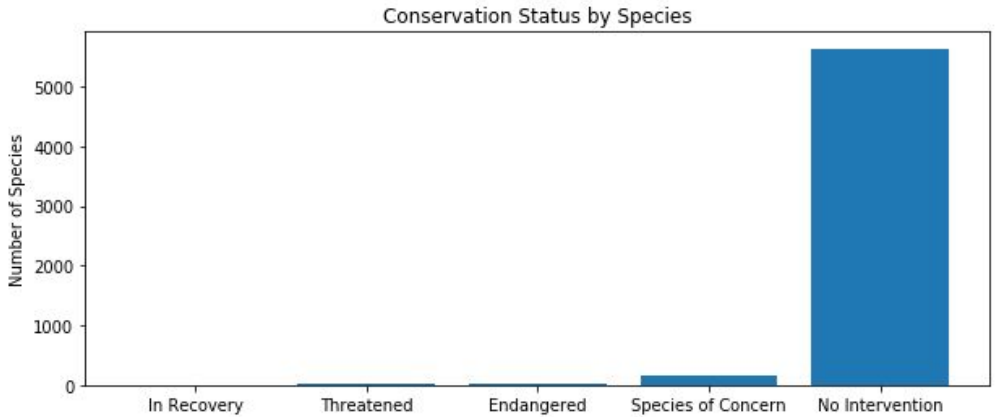
	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	NaN
1	Mammal	Bos bison	American Bison, Bison	NaN
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep,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

Let’s look at Question # 1: Cont’d

First we need to label those Species with a Conservation Status as NaN with **“No Intervention”**. Then we will display this newly updated table to see our Conservation Status by Category of Species.

```
species.groupby('conservation_status').scientific_name.nunique().reset_index()
```

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10



Let’s look at Question # 2:

Are certain Categories of Species more likely to be Endangered?

First it’s important to limit the number of Conservation Status we look at. We can limit this by giving all Species with “No Intervention” status a label of “Not Protected” and all others can be labeled as “Protected”. Also we need to determine what Data we need to look at. We need to count the total number of Species by Category that are either “Protected” or “Not Protected”. It is also helpful to see the percent in a “Protected” status compared to the total population of each Category.

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

Let’s look at Question # 2: Cont’d

By looking at the chart we can see that Mammals are more likely to be endangered than the next most likely Birds.

Can we confidently say that Mammals at 17% are more likely than Birds at 15% to be endangered? Not yet as there are still possibilities for errors.

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

To solve this dilemma I ran a *chi squared contingency test* to compare two categorical data-sets for significance. The results of such test provided a *probability value* (*pval* for short) of 0.687. This value means that the difference between Birds and Mammals are not significant enough to call Birds and Mammals most likely to be endangered. Now it’s important to test Mammals against a category with a much lower percent protected like Reptiles at 6%. Using the *chi squared contingency test* a *pval* was calculated of 0.038. Using the standard of 0.05 as our cutoff we can clearly state that the difference is significant enough to take into account. Which was the original Hypothesis generated by looking at the percent protected of both categories.

Therefore there is only 1 category of Species most likely to be endangered Mammals.

Recommendations Based on Findings:

Through my analysis there were two areas of concern that Conservationists should take notice of.

1. First, there are an alarming number of Species in “Species of Concern” status and should be continued to be monitored and looked at closer for areas of improvement in terms of protection to ensure the longevity of these Species.

2. Second, Mammals are significantly more likely to be endangered than all other categories of Species, with the exception of Birds. What this means is that although Mammals have the highest likeliness of being endangered Birds are also very likely to endangered as well. With this observations it is my recommendation that Conservationists look at ways to improve protection of Mammals and Birds. Also, it would be recommended that new programs be instituted to make the public aware of these findings to allow for more possibilities of funding and support.

Foot and Mouth Disease Testing Effectiveness:

Task 1: What is the ideal sample size to be tested to determine the effectiveness of the foot and mouth disease prevention program for sheep?

Task 2: How long will it take to observe enough sheep to determine effectiveness of the foot and mouth disease prevention program?

Let’s look at Task 1:

What is the ideal sample size to be tested to determine the effectiveness of the foot and mouth disease prevention program for sheep?

First let's look at the Observation file provided. It shows 3 columns Scientific Name, Park Name, and Observations. We can use Scientific Name to combine with our previous Species File to filter out only sheep.

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

First I look for all common names with the word sheep and create a new column that shows True if common name has “sheep” in it.

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention	False	False
1	Mammal	Bos bison	American Bison, Bison	No Intervention	False	False
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	No Intervention	False	False
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
4	Mammal	Cervus elaphus	Wapiti Or Elk	No Intervention	False	False

Let’s look at Task 1: Cont’d

It looks like there were plants also included in our table so I applied one more filter to search for both “sheep” in the Common Name and “Mammal” in the Category of Species

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

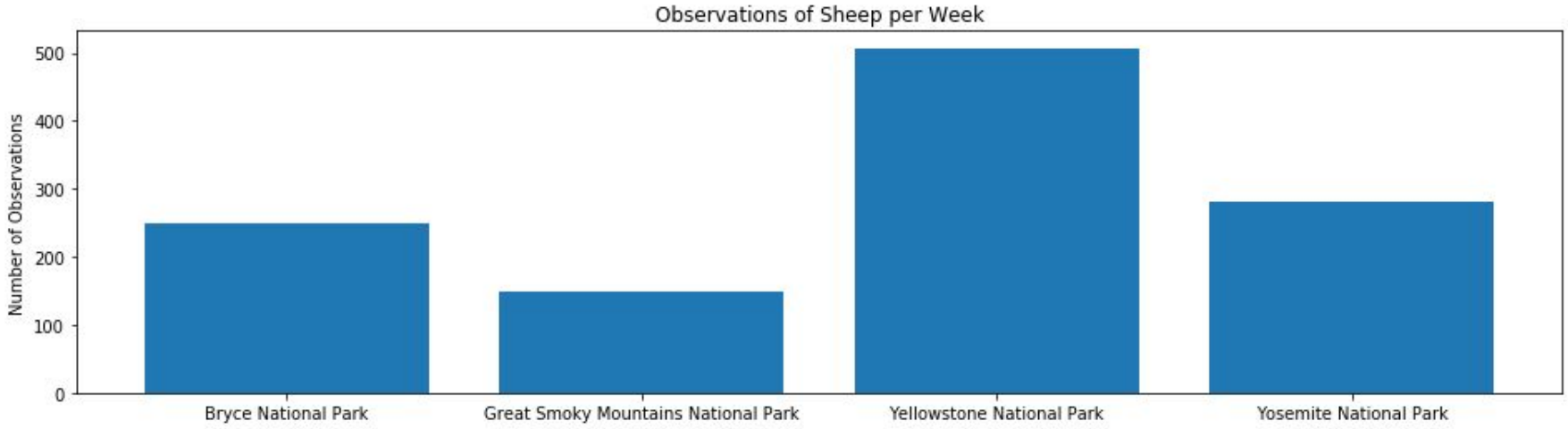
Next, we need to combine the Observations File and Species File to see the number of observations by Park Name and also by Scientific Name of the sheep. Only sheep are shown in this graph.

	scientific_name	park_name	observations	category	common_names	conservation_status	is_protected	is_sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
2	Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
3	Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4	Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
5	Ovis canadensis sierrae	Yosemite National Park	39	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
6	Ovis canadensis sierrae	Bryce National Park	22	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
7	Ovis canadensis sierrae	Great Smoky Mountains National Park	25	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True	True
8	Ovis aries	Yosemite National Park	126	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
9	Ovis aries	Great Smoky Mountains National Park	76	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
10	Ovis aries	Bryce National Park	119	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
11	Ovis aries	Yellowstone National Park	221	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True

Let’s look at Task 1: Cont’d

In order to fully understand what the total number of observations of sheep are in each Park we needed to simply group our data by Park to display our totals. With these numbers found we can no move on to looking for the ideal sample size for determining effectiveness.

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282



Let’s look at Task 1 and 2:

We know that 15% of sheep at Bryce National Park currently have foot and mouth disease and we want to see a reduction of 5 percentage points or in other words we’d like to see that number drop to 10%.

First we must use a sample size calculator to find out our ideal sample size. I choose Optimizely for it’s ease of use and reliability.

Next we need to know our current baseline. 15% as laid out at Bryce National Park

We are going to use a significance of 90% as this is the general used standard.

Lastly we must determine the minimum detectable effect. The formula for determining that is as follows:

100 * (old percentage - new percentage) / old percentage

In this case 100 * (0.15 - 0.10) / 0.15 or 33.3333%

- **Simply input this data into the calculator and you’ll see a value of 510 as the sample size needed.**
- **Next to determine the length of time in weeks to see a change we take 510 / number of observations at Park.**
- **At Bryce National Park it will take approximately 2 weeks to see a change as there are 250 observations per week currently at the Park.**
- **At Yellowstone National Park it will take approximately 1 week to see a change as there are 507 observations per week currently at the Park.**