Code Academy Capstone project- Option2

Sara Bonatti

July 2018

BLACKROCK®

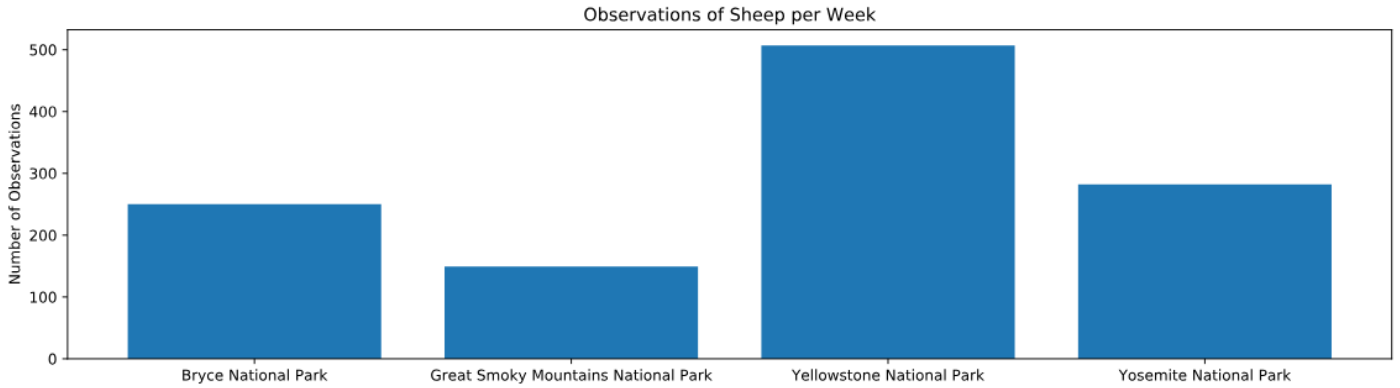
The Data

The data available in the species_info.csv provided the following information on a sample of 5824 animals and plants found in the park:

- The 7 categories of each animal divided into:
  - Mammal
  - Bird
  - Reptile
  - Amphibian
  - Fish
  - Vascular Plant
  - Non-Vascular plant
- The scientific names of each animal for a total of 5541 unique names. This implies that there are multiple animals of a same species in the database
- Common names for a total of 5504 unique name. This implies that multiple species are commonly named in the same way
- Conservation Status divided in 5 categories:
  - Species of concern
  - Endangered
  - Threatened
  - In Recovery
  - Null values: meaning that they are not in any danger

Sheep Everywhere!

Some Sheep enthusiast have kindly recorded the movements of sheeps in our park. As it turns out Yellowstone National Park is full of beautiful peaceful sheeps:

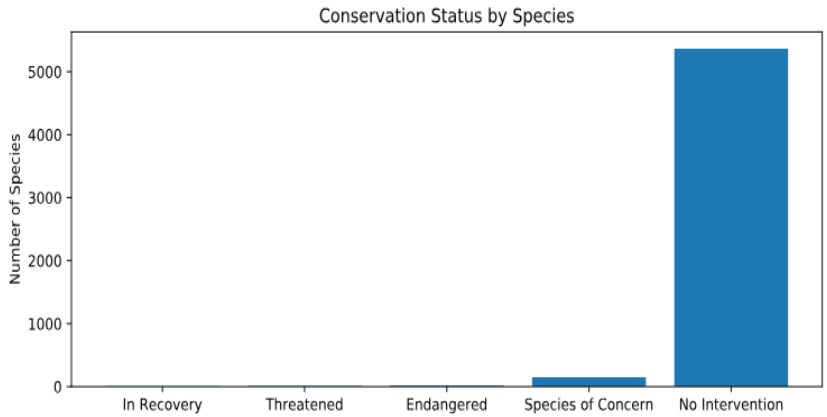


As a side note: There are 6 foxes across all parks so I guess that's why sheeps are free to grow and leave unencumbered in our park

Significance calculations

As part of the project, I had to test whether there were species that were more in danger than others.

As shown in the graph and table below, while most of the species seem to need no intervention, the data in general show that Mammals and Birds seem to be more protected than other species:



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

In order to test whether there was indeed a statistical difference, I run a chi-square test to test whether there was a significant difference between:

- Birds and Mammals
- Mammals and Reptile

Significance calculations cont.

Mammals and Birds:

To check the significance I used the following code:

```
contingency = [[30, 146],
               [75, 413]]
print contingency
from scipy.stats import chi2_contingency

chi2,pval,dof,expected=chi2_contingency(contingency)
print pval
```

The resulting P-Value is 68.76%. **Therefore there is no statistical difference between Mammals and Birds**

Mammals and Reptiles:

To check the significance I used the following code:

```
contingency_2 = [[30, 146],
               [5, 73]]
print contingency_2
chi2,pval_reptile_mammal,dof,expected=chi2_contingency(contingency_2)
print pval_reptile_mammal
```

The resulting P-Value is 3.84%. **Therefore there is a statistical difference between Mammals and Reptiles.** From our initial data we know that this difference means that Mammals are statistically more protected than Reptiles.

Recommendation: The species that are most protected are Mammals and Birds. I would advise conservationist to focus their effort also on other endangered species, particularly plants

Sample Size determination for Foot and Mouth Diseases

I’ve been asked to estimate the minimum sample size need to assert statistically whether the program for the reduction of foot and mouth diseases at Yellowstone National park is indeed working.

What we know is:

- Baseline rate: 15% given past research at Bryce National Park
- They want to be able to detect changes of at least 5%
- Therefore their minimum detectable effect is 5%/15%=33.33%
- They want results with a statistical significance of 90%

Given the information above **the necessary sample size is 870**

if we take into account the weekly observations that the park rangers are able to collect each week it turns out that:

- In Yellow stone park they it will take **2 weeks** to collect the necessary data
- In Bryce National Park it will take **3.5 weeks** to collect the necessary data

park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282