codecademy

Biodiversity for the National Parks

Introduction to Data Analysis Capstone
Adam Zavala
June 13, 2019

Table of Contents

- Focusing Conservation Efforts for Protected Species
  - Understanding the source data
  - Species conservation status
  - Investigating Protected Species
  - Testing Protected Status Rates
  - Where to focus conservation efforts
- Reducing Foot and Mouth Disease Rates in Sheep
  - Sample size determination
  - Observations
  - Test Duration

Focusing Conservation Efforts for Protected Species

Testing for significance in the difference of protection status rates between species

Understanding the Source Data

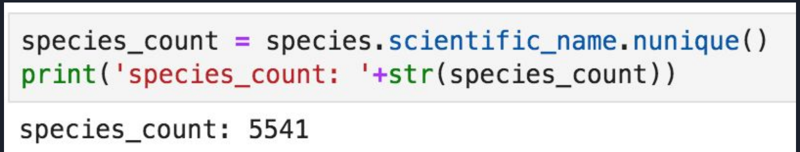
All data was provided on a file **species.csv** which contained 5,824 unique records containing the following information for each:

- Species Category (category)
- Scientific Name of Species (scientific_name)
- Common Name of Species (common_name)
- Conservation Status of Species (conservation_status)

Understanding the Source Data

Other important items of note on the data:

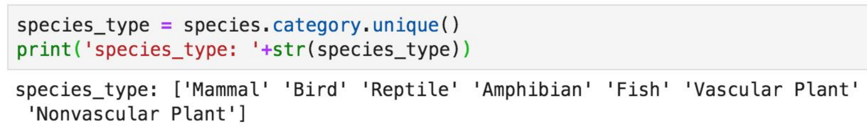
There are **5,541** unique scientific names listed in the data



```
species_count = species.scientific_name.nunique()
print('species_count: '+str(species_count))
```

```
species_count: 5541
```

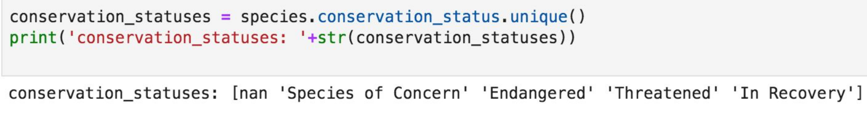
There are **7** unique categories of species.



```
species_type = species.category.unique()
print('species_type: '+str(species_type))
```

```
species_type: ['Mammal' 'Bird' 'Reptile' 'Amphibian' 'Fish' 'Vascular Plant'
 'Nonvascular Plant']
```

There are **4** unique conservation statuses listed across 191 records. The remaining records had a null value listed for their conservation status.



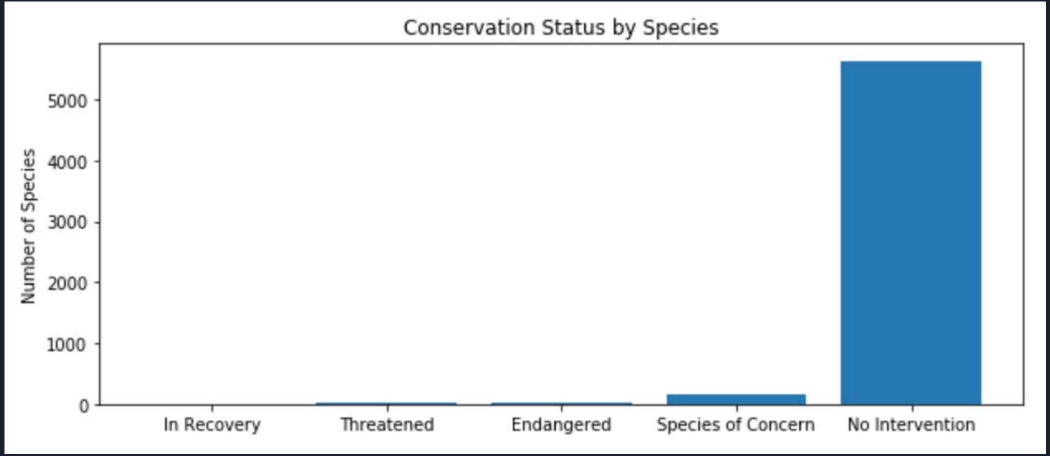
```
conservation_statuses = species.conservation_status.unique()
print('conservation_statuses: '+str(conservation_statuses))
```

```
conservation_statuses: [nan 'Species of Concern' 'Endangered' 'Threatened' 'In Recovery']
```

Species Conservation Status

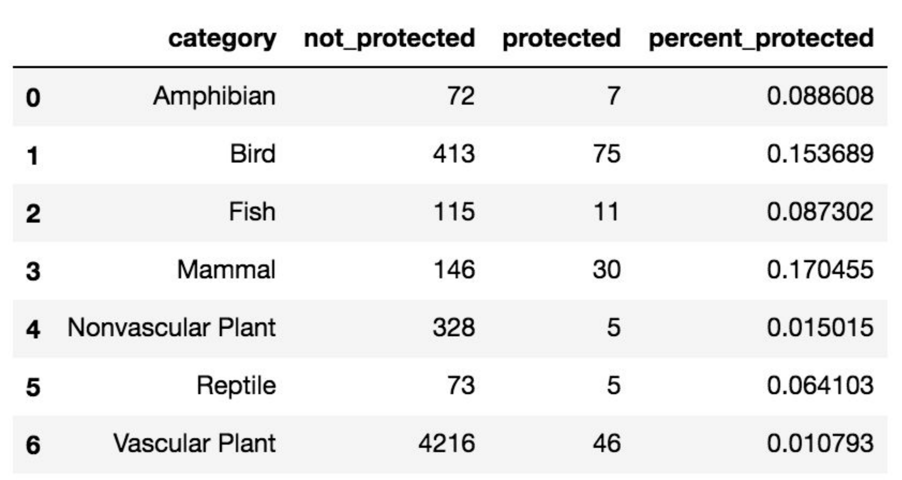
The vast majority of species in the data have no conservation status listed and thus require no intervention. We will label those as “No Intervention” and preserve the conservation status for the others.

Conservation Status	Count
In Recovery	4
Threatened	10
Endangered	15
Species of Concern	151
No Intervention	5,363



Investigating Protected Species

In order to strategically direct conservation efforts, we need to determine if some species are more likely to become endangered than others. First, we must segment the data into two groups: Protected and Not. All species *not* listed as “No Intervention” in the conservation_status field will be marked as **is_protected = True**. All others will be **is_protected=False**. If we further group by **category** and **is_protected** we get the following crosstab.



	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

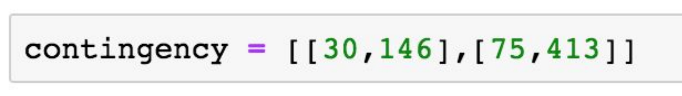
Investigating Protected Species

Now that we have our crosstabs and rates of being designated as a protected species, we need to test to see if the differences between their **percent_protected** values are significant. Because there are multiple pieces of data being evaluated and the data is categorical, we used a **Chi Squared Test**.

Testing Protection Status Rates

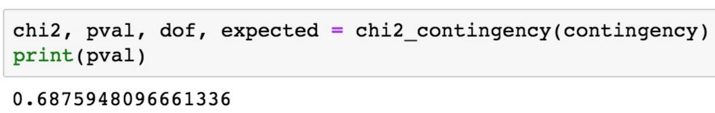
Our first comparison will be to look at birds and mammals which have the highest rates of protection status (15.3% and 17.0% respectively) to see if there is a difference between them.

First, we need to create a contingency table for the values of **not_protected** and **protected** for birds and mammals.



```
contingency = [[30,146],[75,413]]
```

Then, after importing the necessary python libraries, we run a Chi Square test to get the p-value to determine significance of the difference between their protection status



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

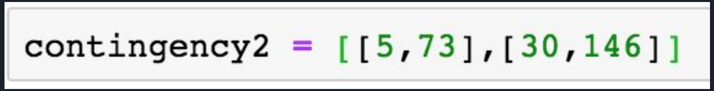
```
0.6875948096661336
```

Because the p-value returned by our test is greater than 0.05 we cannot reject the null hypothesis and thus we conclude that there is **not** a significant difference between the protection status rates of birds and mammals.

Testing Protection Status Rates

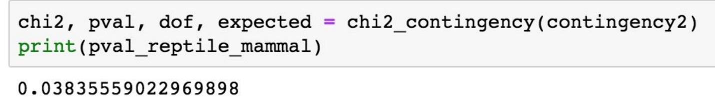
Next, we will compare the protection status rates of mammals (17.0%) and reptiles (6.4%) in order to determine if *this* difference is significant.

First, we need to create a contingency table for the values of **not_protected** and **protected** for birds and mammals.



```
contingency2 = [[5,73],[30,146]]
```

Then we run a Chi Square test to get the p-value to determine significance of the difference between their protection status rates.



```
chi2, pval, dof, expected = chi2_contingency(contingency2)
print(pval_reptile_mammal)
```

```
0.03835559022969898
```

Because the p-value returned by our test is less than 0.05 we reject the null hypothesis and thus we conclude that there **is** a significant difference between the protection status rates of reptiles and mammals.

Where to Focus Conservation Efforts

Given the Chi Square Test results indicated to us that there’s no significant difference between protection status rates of birds and mammals *and* that there **is** a significant difference between mammals and reptiles, we can conclude that birds and mammals are more likely to be designated with a protection status that would require some form of intervention.

As a result of this, it is recommended that conservation efforts are directed primarily to species of birds and mammals are they are more susceptible to endangerment, concern, or other threats to their survival.

Reducing Foot and Mouth Disease Rates in Sheep

Sample Size Determination and Required Observation Duration

Reducing Foot and Mouth Disease at Yellowstone National Park

With current efforts underway to reduce the prevalence of Foot and Mouth disease in species of sheep at Yellowstone National Park by 5%, we want to make sure we're testing the right number of sheep to know if our detected impact of 5% is statistically significant.

What we need to determine are:

- The right sample size to measure this difference in a significant way.
- The amount of time needed to conduct observations.

Sample Size Determination

Sample size determination requires 3 figures be input into a sample size calculator: Baseline, Confidence Level, and Minimum Detectable Effect.

Baseline: We are using **15%** as our baseline, which is the percentage of sheep at Bryce National Park that have Foot and Mouth disease

Confidence Level: We will use a **90%** confidence level for our calculation

Minimum Detectable Effect: We're looking to reduce the prevalence of Foot and Mouth disease by 5% from our baseline. The calculation for this is *0.05 / 0.515% = 0.333 =* **33.3%**

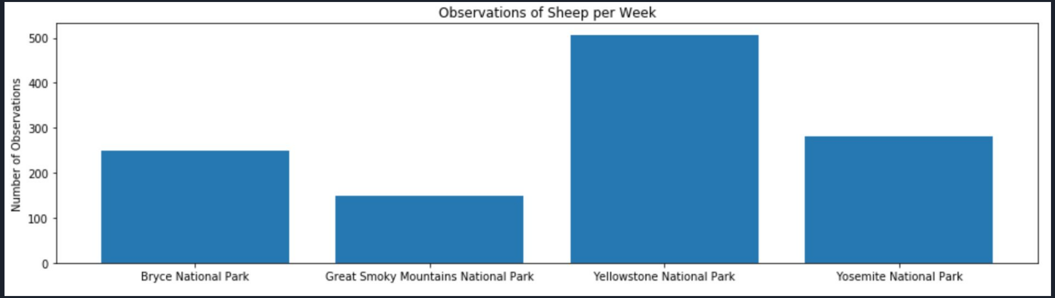
Sample Size: Entering these three figures into the Codecademy Sample Size Calculator gives us a sample size per variant of **870**

Observations

Now that we know we need a sample size of **870** sheep to have confidence in the results of our observations, we need to know how long our observations will need to take place. In order to determine this, we first need to know how many observations of sheep there are per week at the various national parks, including Yellowstone.

We can gather this by merging our **species.csv** file with **observations.csv** and using **groupby** to group the data by park_name and the sum of **observations**

The results are displayed on a chart below.



Bryce:	**250**
Smoky:	**149**
Yellowstone:	**507**
Yosemite:	**282**

Test Duration

We have our required sample size (**870)** for each variant in order to have confidence in whatever results we get in our observations for the prevalence of Foot and Mouth disease. We also know how many sheep are observed each week at each of four national parks. Now we can calculate how long we'd need to conduct observations at each national park to yield results in which we can have confidence*.

- Yellowstone National Park: 870 / 507 = 1.71 → **2 Weeks**
- Bryce National Park: 870 / 250 = 3.48 → **4 Weeks**
- Great Smoky Mountains National Park: 870 / 149 = 5.83 → **6 Weeks**
- Yellowstone National Park: 870 / 282 = 3.08 → **4 Weeks**

Required Sample Size: **870**

Weekly Observations:
Yellowstone: **507**
Bryce: **250**
Smoky: **149**
Yosemite: **282**

* Calculations are rounded up to the nearest whole integer to ensure the required sample size is reached

Thank you!