Biodiversity for the National Parks

Introduction to Data Analysis

Describing the Data

Initial data comes from a file named ‘species_info.csv’ which contains 4 columns containing category, scientific name, common names, and conservation status. There are a total of 5541 species, 7 unique categories, and 5 different conservation status types. Initially the conservation status contained NaN values, but were replaced using pandas fillna method.

```
species.head()
```

	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

Count of scientific names grouped by conservation status

```
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

Endangered Status Analysis

Are certain types of species more likely to be endangered?

	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

Chi squared test

This scipy function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table observed. The expected frequencies are computed based on the marginal sums under the assumption of independence.

Significance calculations

Test if species in category Mammal are more likely to be endangered than species in Bird. Using a contingency table setup in the following format:

||protected|not protected| |-|-|-| |Mammal|?|?| |Bird|?|?|

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

We can then compute the p-value: chi2_contingency(contingency)

Which then returns the output:
(0.1617014831654557, 0.6875948096661336, 1, array([[27.8313253, 148.1686747], [77.1686747, 410.8313253]]))

So the difference between Mammal and Bird ISN'T significant

Significance calculations 2

Test if species in category Mammal are more likely to be endangered than species in Reptile. Using a contingency table setup in the following format:

||protected|not protected| |-|-|-| |Mammal|?|?| |Reptile|?|?|

contingency = [[30, 146],
[5, 73]]

We can then compute the p-value: chi2_contingency(contingency)

Which then returns the output:
(4.289183096203645, 0.03835559022969898, 1, array([[24.2519685, 151.7480315], [10.7480315, 67.2519685]]))

So the difference between Mammal and Reptile IS significant

Recommendation

Based on significance calculations it appears that Mammals and Birds are more likely to be endangered than Reptiles. This information should be used to request more resources to be allocated to the preservation of these species.

Foot and Mouth Disease Study

Our scientists know that 15% of sheep at Bryce National Park have foot and mouth disease. Park rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park. The scientists want to test whether or not this program is working. They want to be able to detect reductions of at least 5 percentage points. For instance, if 10% of sheep in Yellowstone have foot and mouth disease, they'd like to be able to know this, with confidence.

	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

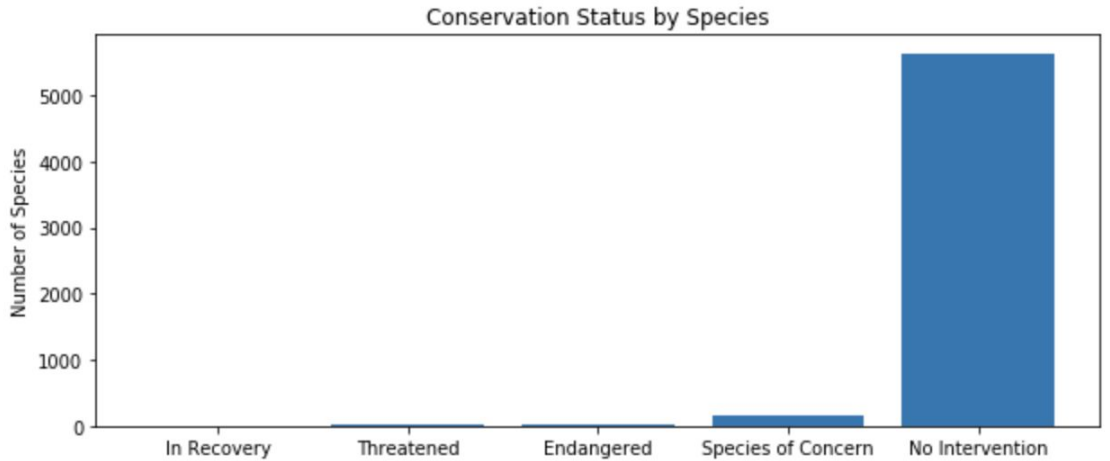
Sheep sightings at different National Parks.

Sample Size Calculations

Using the Codecademy sample size calculator it was determined that the number of sheep needed to observe from each park was 890. This was computed using a baseline of 15%, confidence of 90%, and minimum detectable effect of 33%. This means in order to observe enough sheep it would take 3.56 weeks at Bryce and 1.75 weeks at Yellowstone.

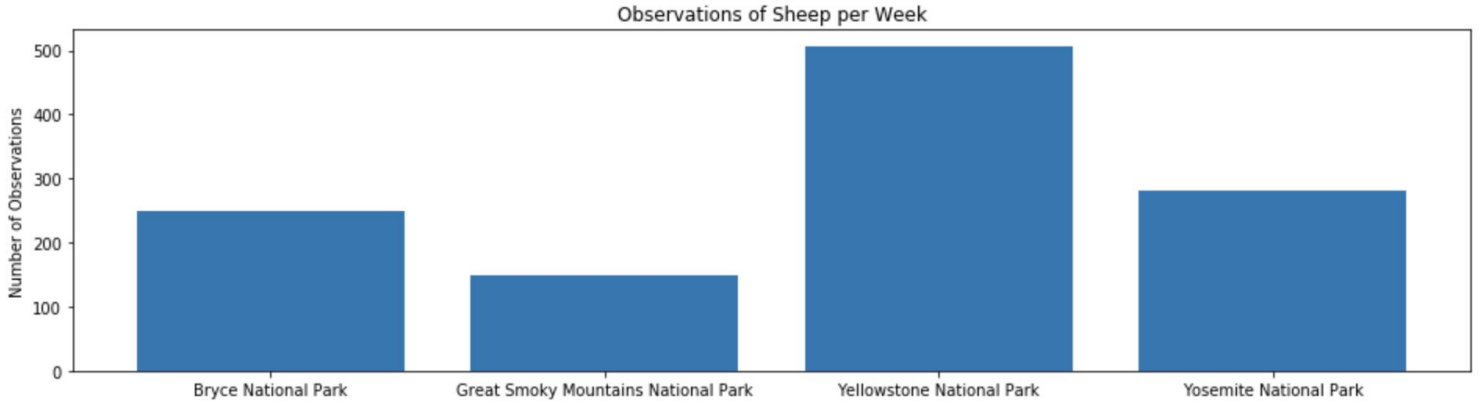
Conservation Status by Species Graph

```
plt.figure(figsize=(10, 4))
ax = plt.subplot()
plt.bar(range(len(protection_counts)),
        protection_counts.scientific_name.values)
ax.set_xticks(range(len(protection_counts)))
ax.set_xticklabels(protection_counts.conservation_status.values)
plt.ylabel('Number of Species')
plt.title('Conservation Status by Species')
plt.show()
```



Observations of Sheep per Week Graph

```
plt.figure(figsize=(16, 4))
ax = plt.subplot()
plt.bar(range(len(obs_by_park)),
        obs_by_park.observations.values)
ax.set_xticks(range(len(obs_by_park)))
ax.set_xticklabels(obs_by_park.park_name.values)
plt.ylabel('Number of Observations')
plt.title('Observations of Sheep per Week')
plt.show()
```



Thank You

Capstone project by Ian De Bie