

CAPSTONE PROJECT: BIODIVERSITY OF THE NATIONAL PARKS

INTRODUCTION TO DATA ANALYSIS WITH PYTHON

Christopher Marasco (lucioangelom)
May 5th, 2018

codecademy / PRO



lucioangelom
Pro Intensive Member



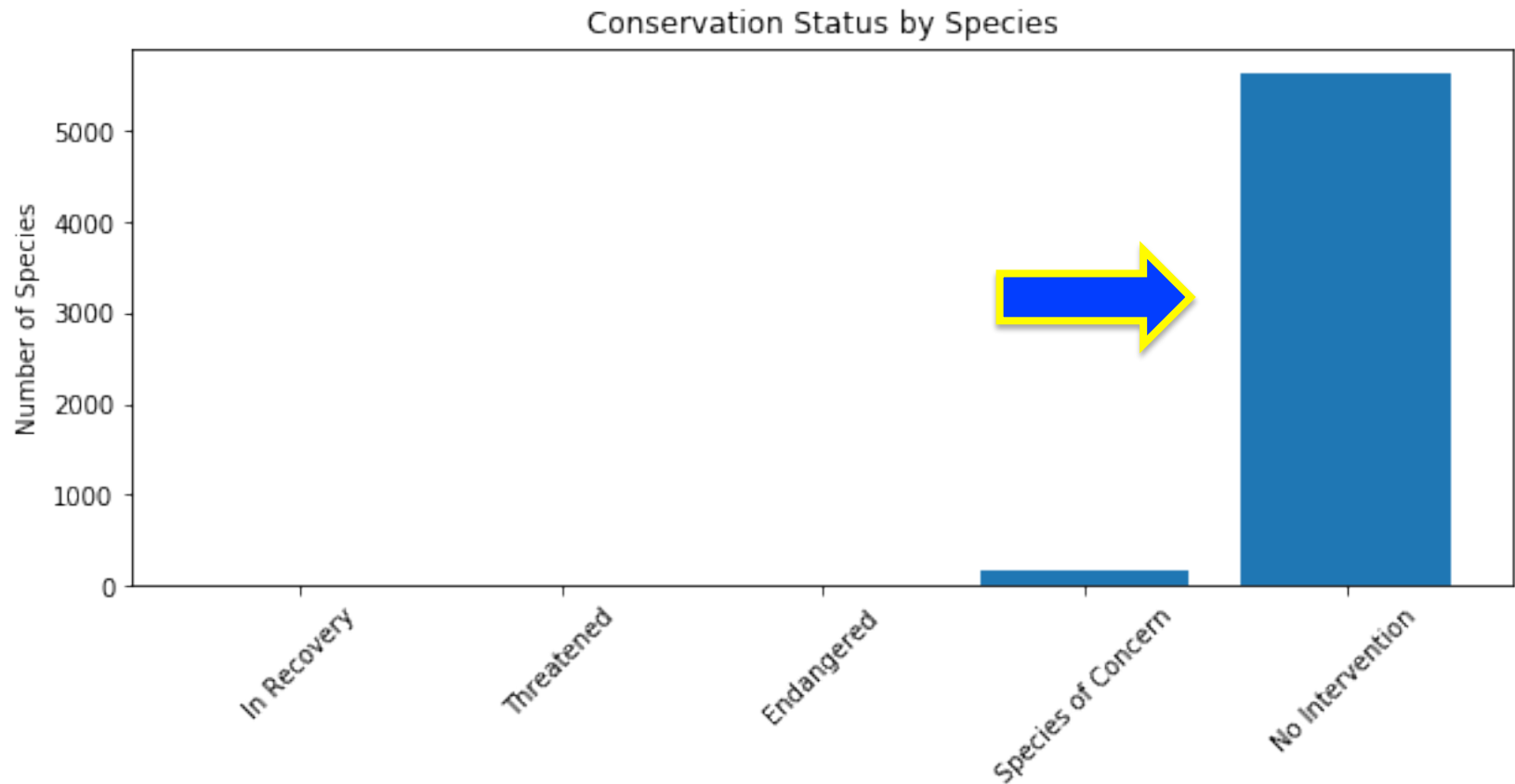
SPECIES_INFO.CSV DATA SET. What does my data look like?

	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
7	Mammal	Canis latrans	Coyote	Species of Concern
8	Mammal	Canis lupus	Gray Wolf	Endangered
9	Mammal	Canis rufus	Red Wolf	Endangered

Our data frame contains 5,824 observations over 4 columns: Category, Scientific Name, Common Names, and Conservations status. A sampling of the data is above.

- **category** - There are 7 unique categories of species: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant.
- **scientific_name** – There are 5,541 unique instances within this field, far too many to list and nearly 1 for every record within the data set.
- **common_names** – This field contains corresponding common names for each of the scientific_names entries. This data is untidy as there are multiple values for each observation. We did not do extensive research into this column.
- **conservation_status** – This field contains 5 unique values: Species of Concern, Endangered, Threatened, In Recovery, and nan.

SPECIES_INFO.CSV DATA SET. What does my data look like?



After grouping our data by conservation status, we can see there are a large number of species on the list without intervention concerns. We plotted our data using matplotlib imported as plt, but we also used the Pandas library imported as pd to use their robust database manipulation tools.

SPECIES_INFO.CSV DATA SET. What does my data look like?

```
In [59]: #Create a lambda function to apply
mylambda = lambda x: 'True' if x != 'No Intervention' else 'False'
#Apply mylambda to the newly created column called 'is_protected'
species['is_protected'] = species.conservation_status.apply(mylambda)
#Check data frame to ensure colum creation and lambda are functioning properly
species.head(10)
```



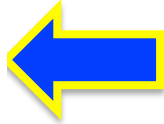
Out[59]:

	category	scientific_name	common_names	conservation_status	is_protected
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	No Intervention	False
1	Mammal	Bos bison	American Bison, Bison	No Intervention	False
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Dom...	No Intervention	False
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False
4	Mammal	Cervus elaphus	Wapiti Or Elk	No Intervention	False
5	Mammal	Odocoileus virginianus	White-Tailed Deer	No Intervention	False
6	Mammal	Sus scrofa	Feral Hog, Wild Pig	No Intervention	False
7	Mammal	Canis latrans	Coyote	Species of Concern	True
8	Mammal	Canis lupus	Gray Wolf	Endangered	True
9	Mammal	Canis rufus	Red Wolf	Endangered	True

As such, we considered only those species who are were not labeled as “No Intervention”. We did so by creating a lambda function that checked the conservations status, when “No Intervention” was found, a flag of “False” was applied to a new column labeled “is_protected”. In this way we were not able to quickly subset by the remaining animals.

SPECIES_INFO.CSV DATA SET. What does my data look like?

	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



After grouping the applicable data, then pivoting to make our table more legible, we created a “percent_protected” column to help us further visualize the protection status of each category of species.

SPECIES_INFO.CSV DATA SET. How confident am I with my data?

	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

	protected	not protected
Mammal	146	30
Bird	413	75
	protected	not protected
Mammal	146	30
Reptiles	73	5



After importing our SciPy library, we want to test our data. In order to test the significance of our assumption, that the Mammal category is more likely to be endangered than the species in Bird, we're going to run a chi-squared test. In order to do so, we need to create a contingency table. But while we're at it, let's also compare the species in Mammal and Reptiles as well.

SPECIES_INFO.CSV DATA SET. How confident am I with my data?

	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

	protected	not protected
Mammal	146	30
Bird	413	75
	protected	not protected
Mammal	146	30
Reptiles	73	5



```
chi2, pval, dof, expected = chi2_contingency(contingency)
print "The p-value of {} shows there is NOT a significant difference between Birds and Mammals".format(pval)
```

The p-value of 0.687594809666 shows there is NOT a significant difference between Birds and Mammals

```
contingency2 = [[146,30],[73,5]]
chi2, pval, dof, expected = chi2_contingency(contingency2)
print "The p-value of {} shows there IS a significant difference between Reptiles and Mammals".format(pval)
```

The p-value of 0.0383555902297 shows there IS a significant difference between Reptiles and Mammals

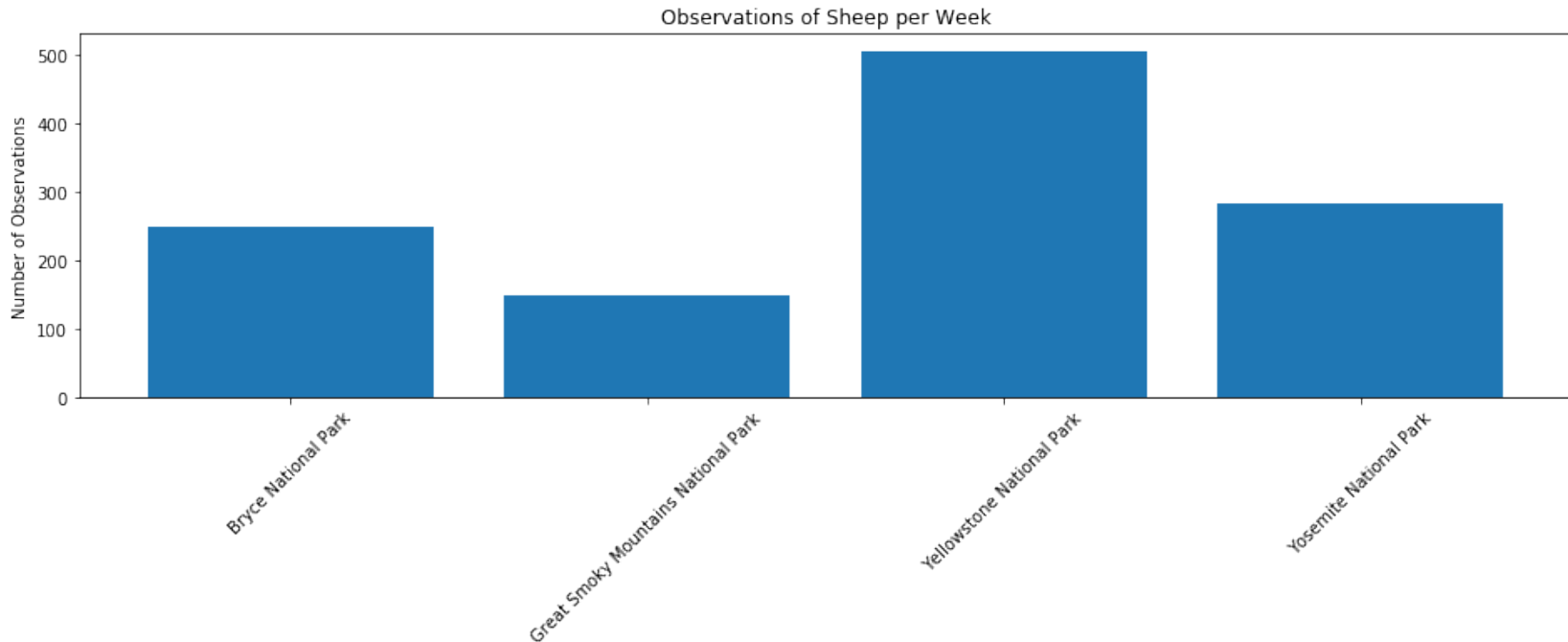
After running our two tests on our two contingency tables it's become clear, there are NOT significant differences between the Birds and Mammals species and there ARE significant differences between the Birds and Reptiles species.

SPECIES_INFO.CSV + OBSERVATIONS. Insights from the data?

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep	park_name	observations
0	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yosemite National Park	126
1	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Great Smoky Mountains National Park	76
2	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Bryce National Park	119
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True	Yellowstone National Park	221

We merged our database with another titled Observations, by amending the average number of sheep observations per week and the park they were observed.

SPECIES_INFO.CSV + OBSERVATIONS. Insights from the data?

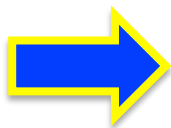


After grouping our data into observation parks, we plotted the weekly sheep sightings per location. While we didn't subset the data by sheep who were labeled as "no intervention" from those "of concern" or "endangered", we can clearly see a healthier population of sheep in the Yellowstone National Park. This could be attributed to simple volume of population, volume of observations commensurate with observers, or some other unforeseen variable. But given the large shifts in observations, I'd recommend that conservationists dig deeper into the data to ensure there aren't best practices being applied in Yellowstone that may be applicable to the other national parks.

SPECIES_INFO.CSV + OBSERVATIONS. Disease Population Sampling

A/B Test Sample Size Calculator

Powered by Optimizely's Stats Engine



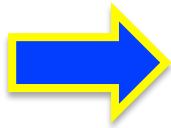
Baseline Conversion Rate

15



%

Your control group's expected conversion rate. [\[?\]](#)



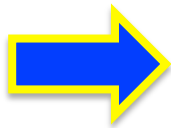
Minimum Detectable Effect

33.3



%

The minimum relative change in conversion rate you would like to be able to detect. [\[?\]](#)



Statistical Significance

90%

[EDIT](#)

95% is an accepted standard for statistical significance, although Optimizely allows you to set your own threshold for significance based on your risk tolerance. [\[?\]](#)

Finally, we would like to acknowledge the amount of Foot and Mouth disease our sheep population at Bryce National Park have. Let's use these data to calculate whether or not our goal of reducing instances of Foot and Mouth by 5% was successful, with confidence.

First I calculate my minimum relative change in conversion. In this case it's simply $100 \times 5\% / 15\% = 33.3\%$

Then I was provided a baseline conversions (15% of the population) and our required statistical significance (90%)

SPECIES_INFO.CSV + OBSERVATIONS. Disease Population Sampling

A/B Test Sample Size Calculator

Powered by Optimizely's Stats Engine

Baseline Conversion Rate

15

%

Your control group's expected conversion rate. [\[?\]](#)

Minimum Detectable Effect

33.3

%

The minimum relative change in conversion rate you would like to be able to detect. [\[?\]](#)

Statistical Significance

90%

[EDIT](#)

95% is an accepted standard for statistical significance, although Optimizely allows you to set your own threshold for significance based on your risk tolerance. [\[?\]](#)

Sample Size per Variation

510



I use my online A/B Test Sample Size Calculator from Optimizely to determine my minimum population size required. And that number is 510!

SPECIES_INFO.CSV + OBSERVATIONS. Disease Population Sampling

A/B Test Sample Size Calculator

Powered by Optimizely's Stats Engine

Sample Size per Variation

510

```
print "In Bryce National Park I see an average of 250 sheeps per week, as such, I would need ~ {} weeks to complete my r
```

In Bryce National Park I see an average of 250 sheeps per week, as such, I would need ~ 2 weeks to complete my required samples size

```
print "In Yellowstone National Park I see an average of 507 sheeps per week, as such, I would need ~ {} week to complete
```

In Yellowstone National Park I see an average of 507 sheeps per week, as such, I would need ~ 1 week to complete my required samples size

If I were to test that my 5% reduction was accurate at Bryce National Park, it would take me ~ 2 weeks, given I have an average of 250 observations per week.

If I were to test that my 5% reduction was accurate at Yellowstone National Park, it would take me ~ 1 weeks, given I have an average of 507 observations per week.