Biodiversity for the national parks

By Boris Kharitontsev

Species Information

Species Information

I was given a file *species_info.csv*, which contained some information on the variety of species.

Each piece of information provided the following:

- Category
- Common name
- Scientific name
- Conservation status

	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	NaN

Species Information - Factual Summary

- The database contains information on **5541** unique species.
- All of the species are divided into **six** categories - *Mammals*, *Birds*, *Reptiles*, *Amphibians*, *Fishes*, *Vascular and Nonvascular plants.*
- The species conservation status could be represented by 5 different statuses:
    - No Intervention Required
    - Species of Concern
    - Endangered
    - Threatened
    - In Recovery

Conservation Status Statistics

Conservation Status Statistics

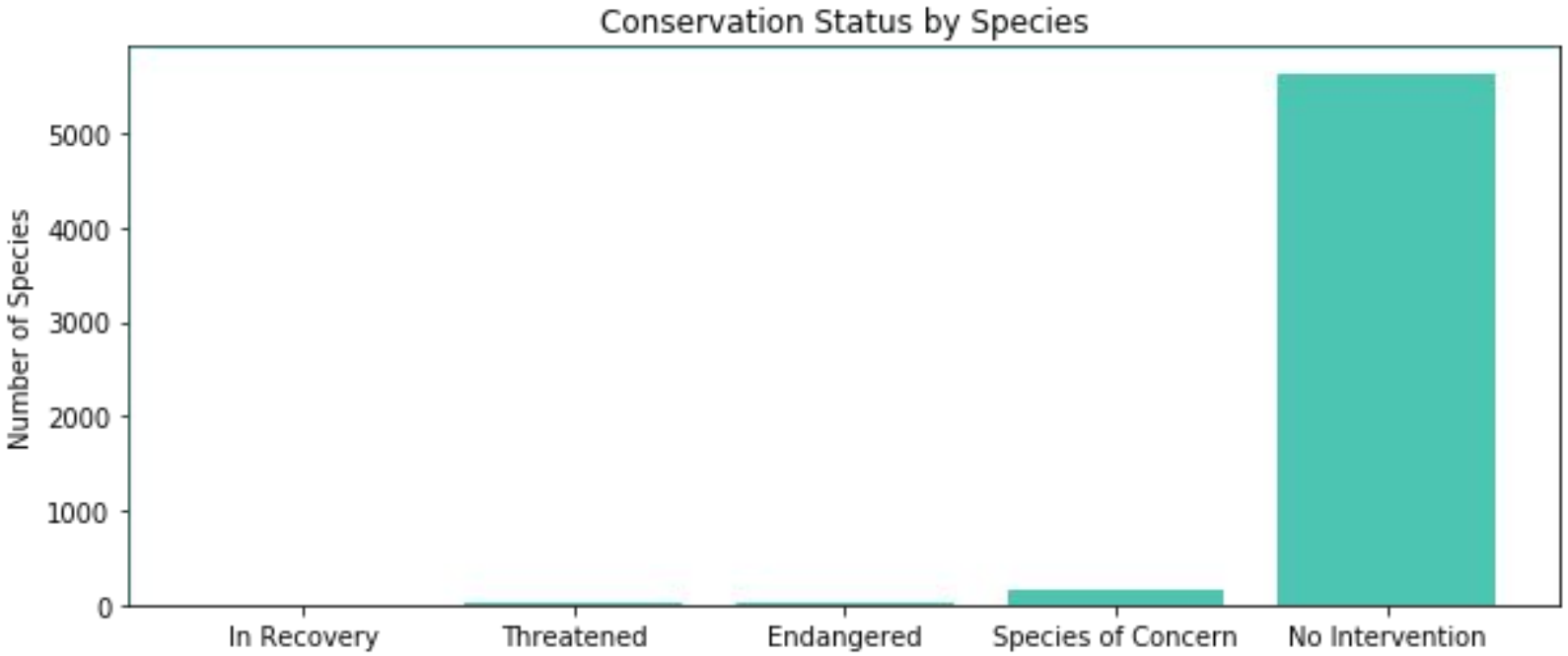
Table on the side provides the number of species, from the least to the largest, in each Conservation Status.

- Most of the species require *No Intervention* - **~96.7%.**
- *Species of Concern* Comprise just **~2.7%** of all.
- Only **2%** of all species that require intervention are *in recovery*.

conservation_status	
In Recovery	4
Threatened	10
Endangered	16
Species of Concern	161
No Intervention	5633

Just 2%

- *Only 2 % of all species that require intervention are in recovery.*
- *Just 4 species out of 191 that require intervention*



Previous empirical data represented as a bar chart

Significance Caluclations

Significance Calculations - Birds and Mammals

Having processed the given data, I was able to extract the numbers of *protected*[1] and *non-protected* species, grouped by *category*.

I have found that the class, with the highest percent of protected species, is ***mammals (~17%).*** They are then followed by ***birds (~15%).***

I furthered my analysis then by completing a significance test, which was about to show whether this difference of 2 percentage points was due to a random error or an actual tendency.

P-value = 0.688 ⟹ Difference is not significant, between *mammals* and *birds*, so it is probably due to a random error.

[1] - *species that require some kind of intervention*

Significance Calculations - Reptiles and Mammals

- On the other hand, only **6%** of ***reptiles*** were protected.
- Conducting a significance test between mammals and reptiles has produced a P-value = 0.038.
- The difference is significant! (even at a 95% significance level).
- Therefore the difference in percentages of the protected species in *mammals* and *reptiles* is probably a real tendency.

Note : these significance tests were conducted using a chi squared method and contingency tables.

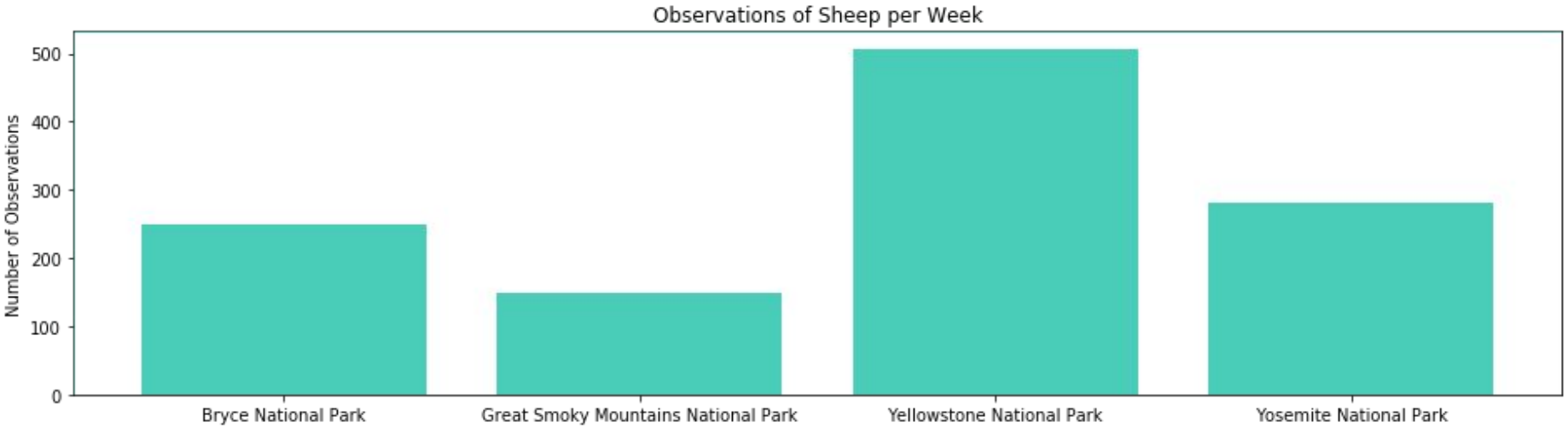
Sample Size Determination

Sheep Observations

Diverting to another file provided, *observations.csv*, I could access all registered observation of various species. Then I was able to extract all observations of the sheep. After appropriate formating, I further developed the data provided to represent the total number of observations of the sheep in each of the four parks, namely *Bryce National Park*, *Great Smoky Mountains National Park*, *Yellowstone National Park and Yosemite National Park*.

	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 Observations Graphed



Sample Size Determination

“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.”

I used Codecademy’s Sample Size Calculator, which required some initial data input, namely *Significance Level, Minimum Detectable Effect and Baseline Conversion Ratio.*

Those had to be calculated.

Sample Size Determination

- Baseline Conversion Ratio = 15%
- Minimum detectable effect = 100* 5/15 = ~33.3%
- Significance Level = 90%

Sample Size Determination - Conclusion

To complete a test, we need to get a large enough sample of sheep in each of the parks. As we have just found, the required sample size is **870.**

We have information on the approximate number of observations per week in each park:

- 250 per week at Bryce National Park
- 507 per week at Yellowstone National Park

So, we could evaluate the number of weeks required to get enough sheep observations:

- For Bryce national Park, it is 870/250 = ~4 weeks
- For Yellowstone National park, it happens to be 870/507 = ~ 2 weeks

The End

Unit 6 - Capstone Project Option 2: “Biodiversity for the National Parks”

By Boris Kharitontsev

For Codecademy.com

2018