Biodiversity Capstone Project

Bolun ZHANG

1. Biodiversity Project

	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), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

- In our dataset ‘species_info.csv’, there are 4 columns, which are category, scientific_name, common_names and conservation_status.

- What’s more, in Column category, we have a list of unique value: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant and Nonvascular Plant.

- We can easily find that in Column conservation_status, there are Species of Concern, Endangered, Threatened and In Recovery, 4 unique values. But there are also lots of Nan values in this column. We should make an analysis to replace them with other comment for eviting problems .

2. Inspected the DataFrame

- How many different species are in the species DataFrame?

  species_count = 5824

- What are the different values of category in the DataFrame species?

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

- What are the different values of conservation_status?

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

3. Analyze Species Conservation Status

Conversation_counts:

	Conservation_status	Scientific_name
0	Endangered	16
1	In Recovery	4
2	Species of Concern	161
3	Threatened	10

We know that there are 5824 species in our data set according to our previous exercise.

But from the table above, we find that there are still many species without a conservation status.

Therefore, we add a now conservation status <No Intervention> to those species who do not have a status before.

Conversation_counts_fixed:

	Conservation_status	Scientific_name
0	Endangered	16
1	In Recovery	4
2	No Intervention	5633
3	Species of Concern	161
4	Threatened	10

- Is the data numerical or categorical?

	Category	Not_protected	Protected	Percent_protected
0	Amphibian	73	7	0.08
1	Bird	442	79	0.15
2	Fish	116	11	0.08
3	Mammal	176	38	0.17
4	Nonvascular Plant	328	5	0.01
5	Reptile	74	5	0.06
6	Vascular Plant	4424	46	0.01

From the chart above, we can easily find that the dataset is categorical (protected or not protected)

- How many pieces of data are you comparing?

Mammal VS Bird:

p value = 0.446 > 0.05, there is no significant difference between Mammal and Bird in terms of being endangered more likely.

Mammal VS Reptile:

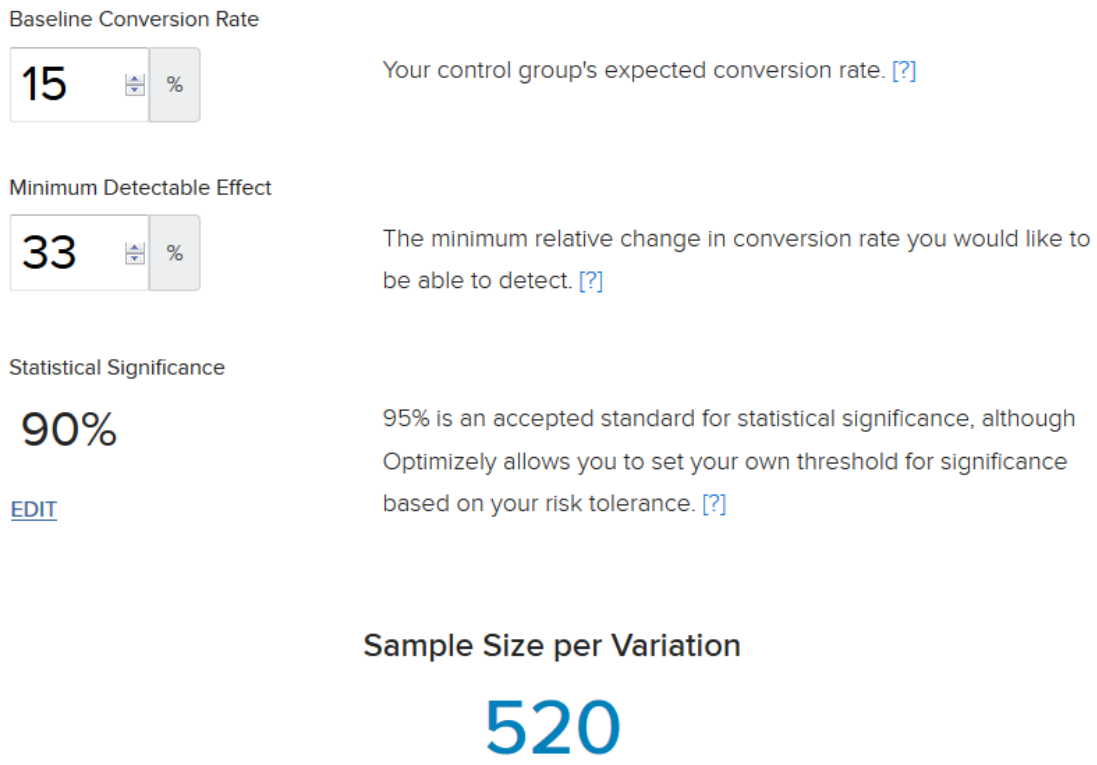
p value = 0.023 < 0.05, there is a significant difference between Mammal and Reptile.

4. Sample size determination

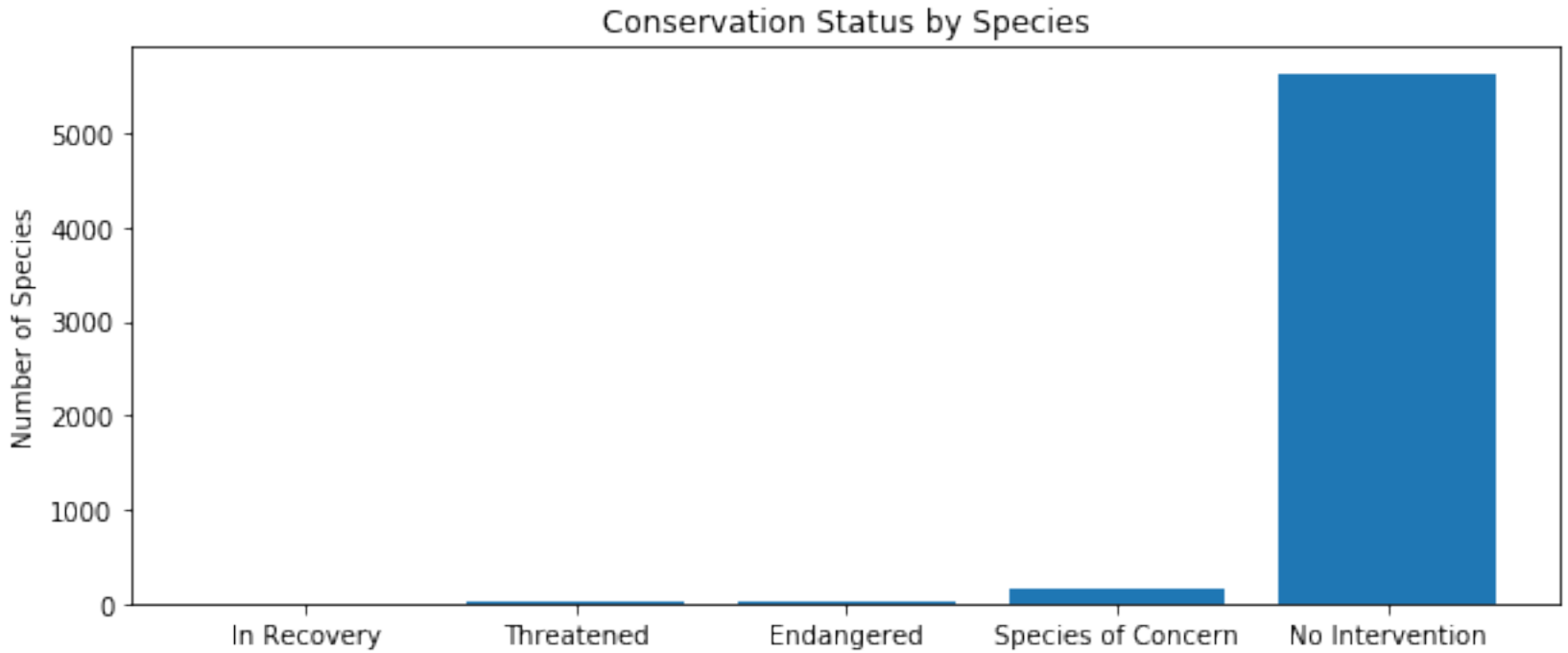
- According to the project, we find that:

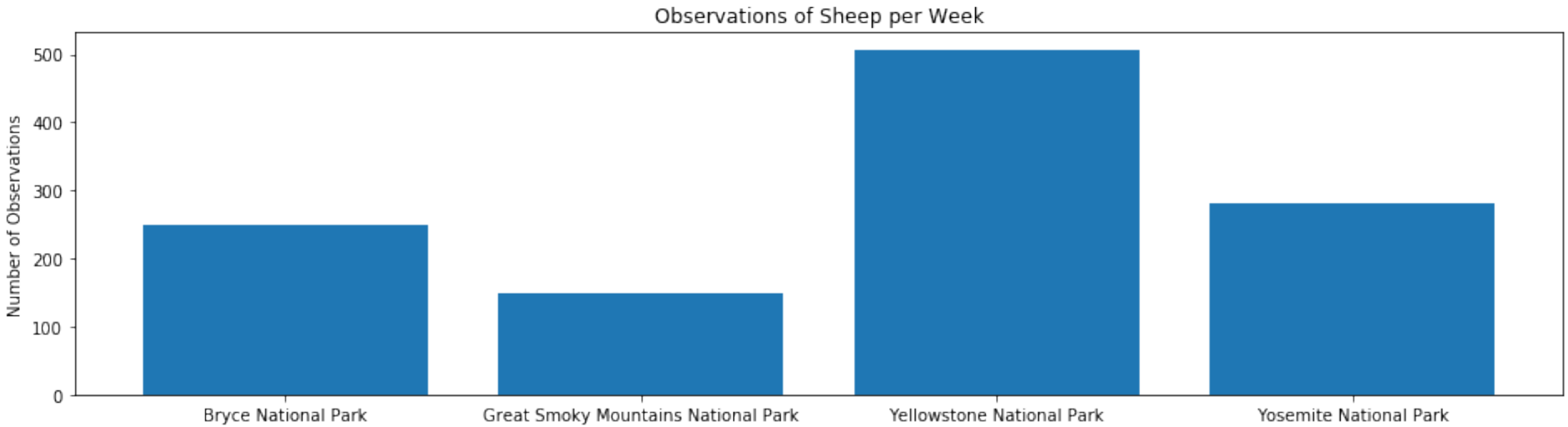
Minimum Detectable Effect = 100 * 0.05 / 0.15 = 33%

With the sample size calculator,
we can calculate the sample size is 520



5. Graphs


Conservation Status by Species
Number of Species
5000
4000
3000
2000
1000
0
In Recovery
Threatened
Endangered
Species of Concern
No Intervention


Observations of Sheep per Week
Number of Observations
0
100
200
300
400
500
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park