BIODIVERSITY FOR THE NATIONAL PARKS

FINAL CAPSTONE PROJECT

CODE ACADEMY – INTRODUCTION TO DATA ANALYSIS

APRIL 10, 2018 – JULY 3, 2018

RAMON ALSINA MUÑOZ

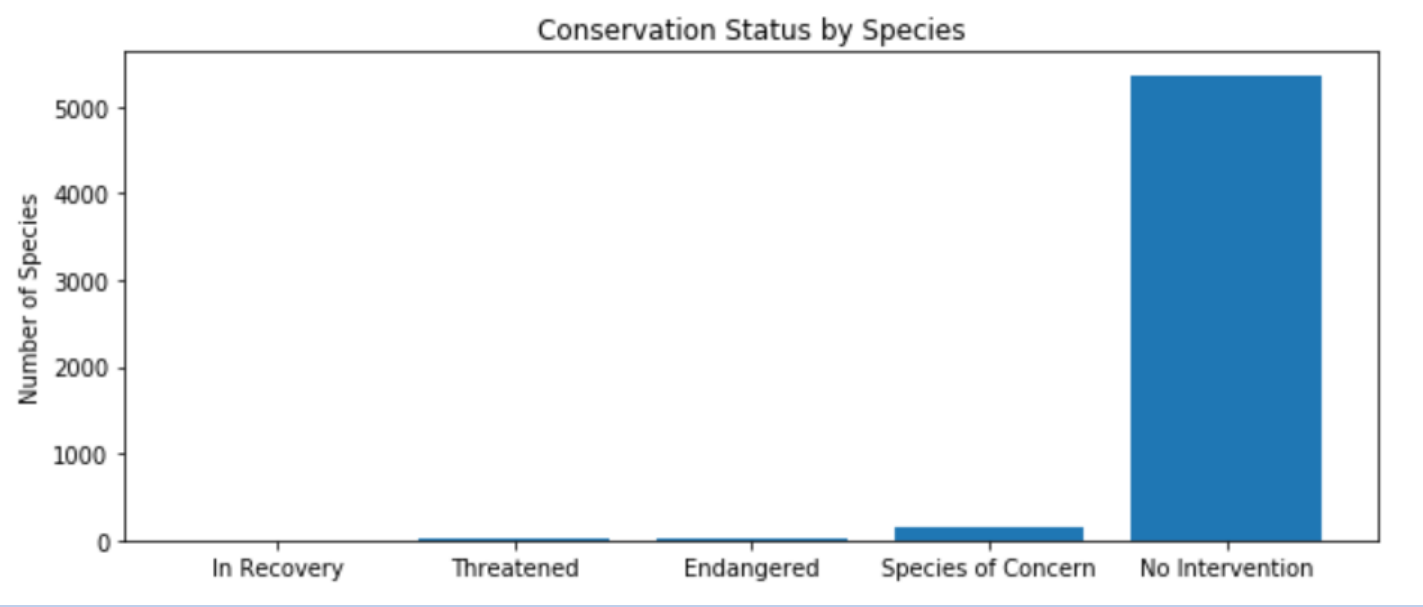
DATA DESCRIPTION

- The data provided describes the situation in terms of conservational status of different species in different US National Parks.
- The data gives details about the scientific name, common name, category and conservational status of each species.
- There are seven types of **Categories**: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant and Non-Vascular Plant.
- There are five types of **Conservational Status**: Species of Concern, Endangered, Threatened, In Recovery and No Intervention.
- There are 5,541 different **Scientific Names.**
- There are 5,504 different **Common Names**.
- There are a total of 5,541 records.

- Looking at the number of Scientific Names by each Conservational Status we see that for most of them there is no need for intervention:

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

- The same conclusion can be also seen in the below bar chart:



- And the same can be also seen in the below summary where the percentage of protected Scientific Names is shown for each Category (protected means a Conservational Status different than No Intervention):

category	not_protected	protected	percent_protected
Amphibian	73	7	8.750000
Bird	442	79	15.163148
Fish	116	11	8.661417
Mammal	176	38	17.757009
Nonvascular Plant	328	5	1.501502
Reptile	74	5	6.329114
Vascular Plant	4424	46	1.029083

CALCULATIONS REGARDING THE SIGNIFICANCE OF ENDANGERED STATUS BETWEEN DIFFERENT CATEGORIES OF SPECIES

- On this section we want to analyze if there are certain types of species that are more likely to be endangered than others.
- We use the same summary as in the previous slide and we see that we are using categorical data (protected vs not protected), and not numerical data:

category	not_protected	protected
Amphibian	73	7
Bird	442	79
Fish	116	11
Mammal	176	38
Nonvascular Plant	328	5
Reptile	74	5
Vascular Plant	4424	46

- Since we are using categorical data we have to use an appropriate significance test.
- We cannot use the Binomial Test for this analysis because more than two pieces of data are used. There is no other choice than using a Chi Squared Test because this test allows to have two or more categorical datasets. In the Binomial Test only one dataset is compared to a specific expectation.

- Based on the provided data, to use the Chi Squared Test we have to build the following contingency table:

CONTINGENCY TABLE		
	protected	not protected
Mammal	38	176
Bird	79	442

- Using the above data we get a **p-value** of **0.16** which is **greater than 0.05**. Since the Null Hypothesis is that there is no significant difference between the datasets and the p-value is greater than 0.15 we can state that there is no significant difference between the two analyzed categories. None of these two species are more likely to be endangered compared one to the other.
- If we now compare Reptiles to Mammals the contingency table will be:

CONTINGENCY TABLE		
	protected	not protected
Reptile	5	74
Mammal	38	176

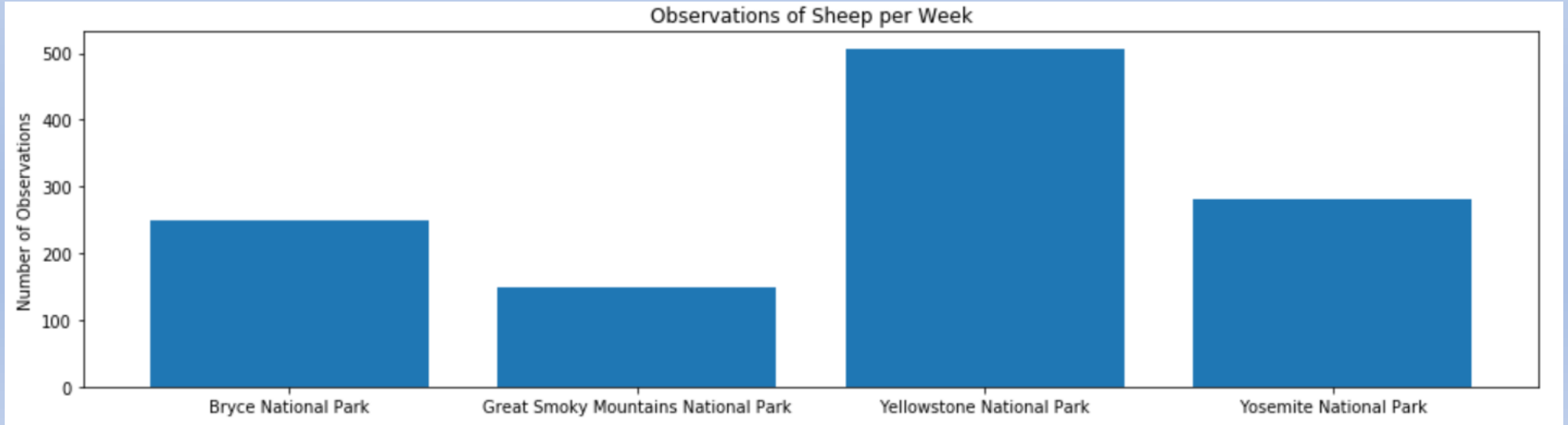
- In this case the result of the **p-value** is **0.02** which is **lower than 0.05**. This means that we can reject the Null Hypothesis and state that that there is a significant difference between the two datasets. In this case the Mammals are prone to be more endangered than the Reptiles (indeed 18% of protected Mammals vs 6% of protected Reptiles).
- The conclusion is that more focus in terms of protection should be put on Mammals, Fish, Birds and Amphibians than in Plants and Reptiles.

SAMPLE SIZE DETERMINATION FOR THE FOOT AND MOUTH DISEASE STUDY

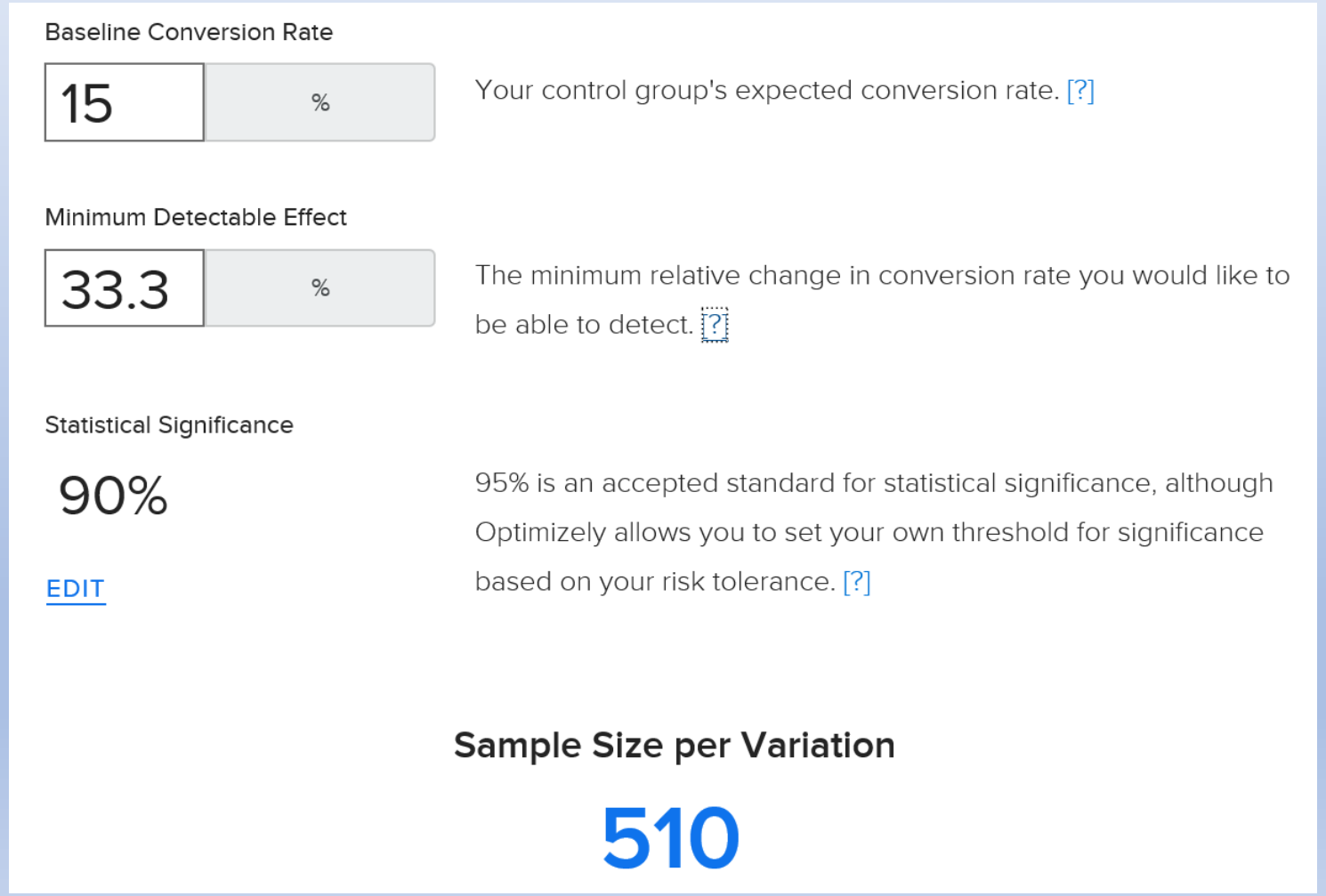
- Conservationists have sent data regarding the number of observations at several national parks for one week.
- Several species of sheep are observed at different parks and combining the initially given data with the number of sheep observations we get the following sheep observation summary:

park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

- The above data can be graphically depicted as:



- Conservationists know that 15% of sheep at Bryce National Park have foot and mouth disease. This makes the **Baseline Conversion Rate = 15%.**
- Since conservationists want to detect reductions of this disease of at least 5%, the **Minimum Detectable Effect** is calculated as follows: (0.05 / 0.15) * 100 = **33.3%**.
- With the above data the **sample size** can be calculated, which in this case is **510 observations**.



- The final step will be to calculate the number of weeks needed to reach the calculated sample size.
- For instance, for Bryce National Park two weeks will be needed to reach the 510 observations: 510 sample size / 250 given observations = **2.04 weeks**.
- For Yellowstone National Park only one week will be needed to reach the 510 observations: 510 sample size / 507 given observations = **1.01 weeks**.