Introduction to Data Analysis

Capstone project:

Biodiversity for the National Parks

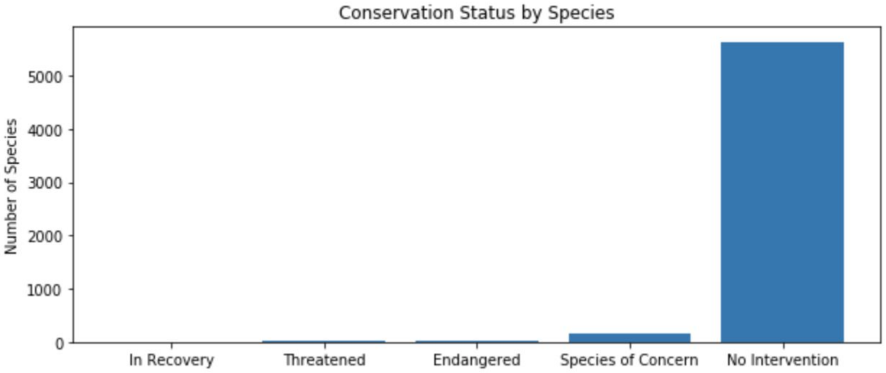
Yu-Ting Shen

Description of the data

The data in the species_info.csv contains 5824 observations (rows) and 4 columns.

- There are 5541 different species in 7 different categories.
  - Catetories are mammal, bird, reptile, amphibian, fish, vascular plant, and non-vascular plant.
- There are 5 conservation status with different number of species.

Conservation status	Number of species
No intervention	5363
Species of concern	511
Endangered	15
Threatened	10
In recovery	4



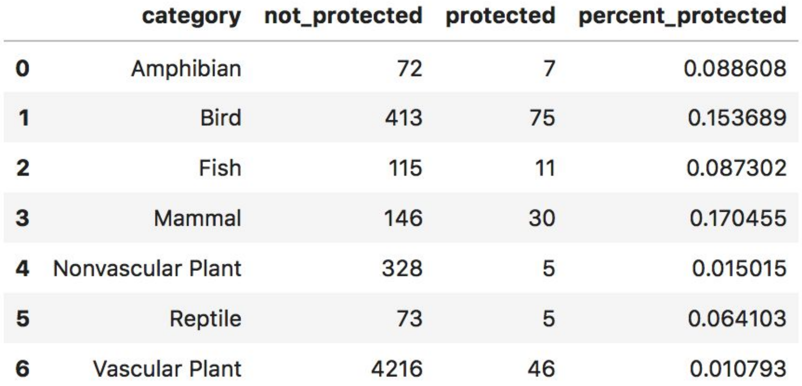
Endangered status

The hypothesis assumes that the difference causes by chance.

We use chi2 test to calculate the p-values between and catetories.

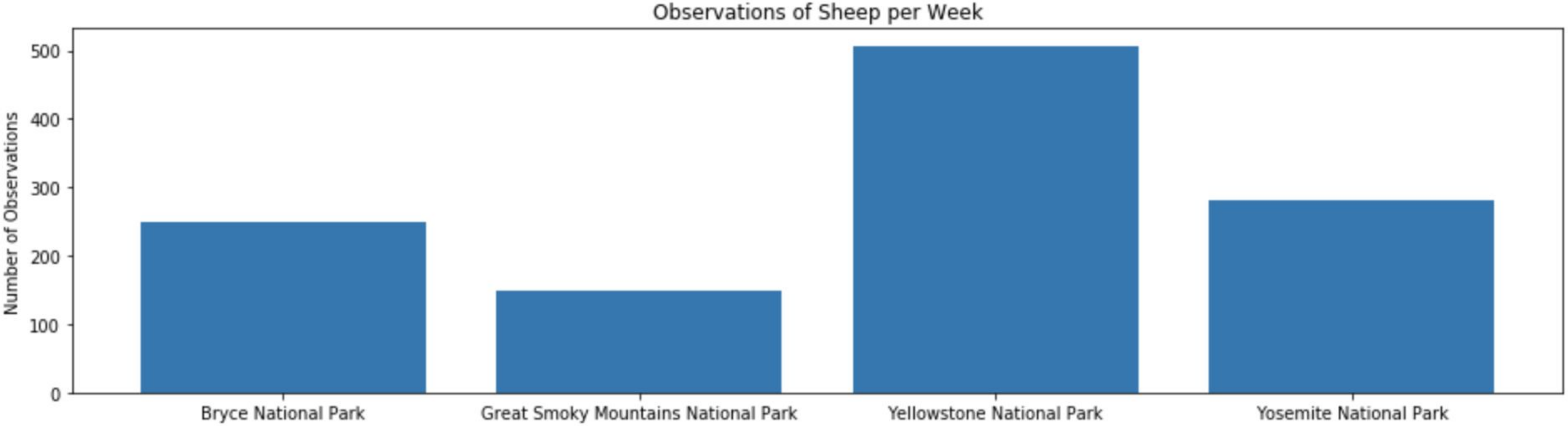
p-value(mammal, bird) = 0.688 > 0.05 → Not significant, accept hypothesis

p-value(reptile, mammal) = 0.038 < 0.05 → Significant, reject hypothesis



	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

Foot & mouth disease study



Foot & mouth disease study

Use baseline = 15%, minimum detectable effect = 33%, and statistical significance = 90%, we get the sample size per variation = 520.

Bryce and Yellowstone National Park have 250 and 507 observations per week , respectively.

We need 520/250 ~ 2.08 weeks for Byrce and 520/507 ~ 1.03 weeks for Yellowstone to observe enough sheep.