Codecademy Capstone Project

Capstone Option 2: Biodiversity for the National Parks

Matt Jane

Species Data Overview

- Data Type
  - All columns have a String datatype
  - Scientific Names consist of a single string, but may be many words
  - Common Names are a concatenation of strings comma separated
- Data Lengths
  - The following table shows the max and min values for various columns;

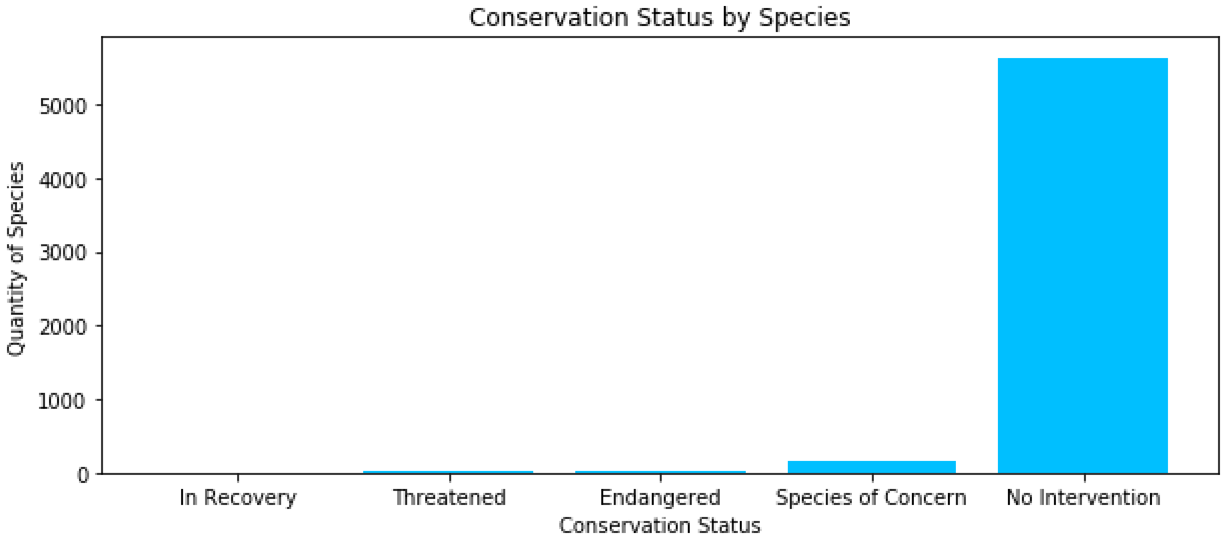
	Min	Max
Scientific Name Length	5	59
Common Name(s) Length	3	218
Qty. of Common Names	1	11

- Unique Values
  - Species: 'Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant', 'Nonvascular Plant'
  - Conservation Status: NaN, 'Species of Concern', 'Endangered', 'Threatened', 'In Recovery‘
  - There are 5541 unique species

Species Data Overview

- The following table and figure show the amount of unique species by conservation status;

Category	# of Scientific Names
In Recovery	4
Species of Concern	151
Threatened	10
Endangered	15
No Intervention	5363



Significance Calculations

- **Chi-squared Test:** A chi-square test looks for relationships (i.e. dependence or independence) between variables. In our example, these variables are category of animal and whether or not they are protected.
- **P-value:** The primary output of a Chi-squared test is the p-value.
  - P ≤ 0.05, statistically you can reject the null hypothesis
  - P > 0.05, statistically you can accept the null hypothesis
- **Contingency Table:** A table showing numerical frequencies relating to various categories. In our example, we have a count of animal species across the various categories that are either protected or not.
- **Significance Testing:**
  - Hypothesis: One species is more likely to be protected than another
  - Null Hypothesis: Not statistically likely that one species is more protected than another

Recommendations to Conservationists

- Once the categories of animals were sorted by the percentage of species that are protected, additional chi-squared tests were ran. Comparing neighboring categories of animals showed us that all but one comparisons have statistical significance.
- The table below shows all of the tested combinations from chi-squared testing;

P-value Table	Vascular Plant	Non-Vascular Plant	Reptile	Fish	Amphibian	Bird	Mammal	% Protected
Vascular Plant	1.000	0.6623	1.45E-4	1.49E-12	1.04E-08	4.61E-79	1.44E-55	1.08
Non-Vascular Plant	-	1.000	0.0336	4.96E-.04	0.002	1.05E-10	1.48E-10	1.50
Reptile	-	-	1.000	0.7406	0.781	0.053	0.03835	6.41
Fish	-	-	-	1.000	0.8247	0.077	0.056	8.73
Amphibian	-	-	-	-	1.000	0.1759	0.128	8.86
Bird	-	-	-	-	-	1.000	0.6875	15.37
Mammal	-	-	-	-	-	-	1.000	17.05

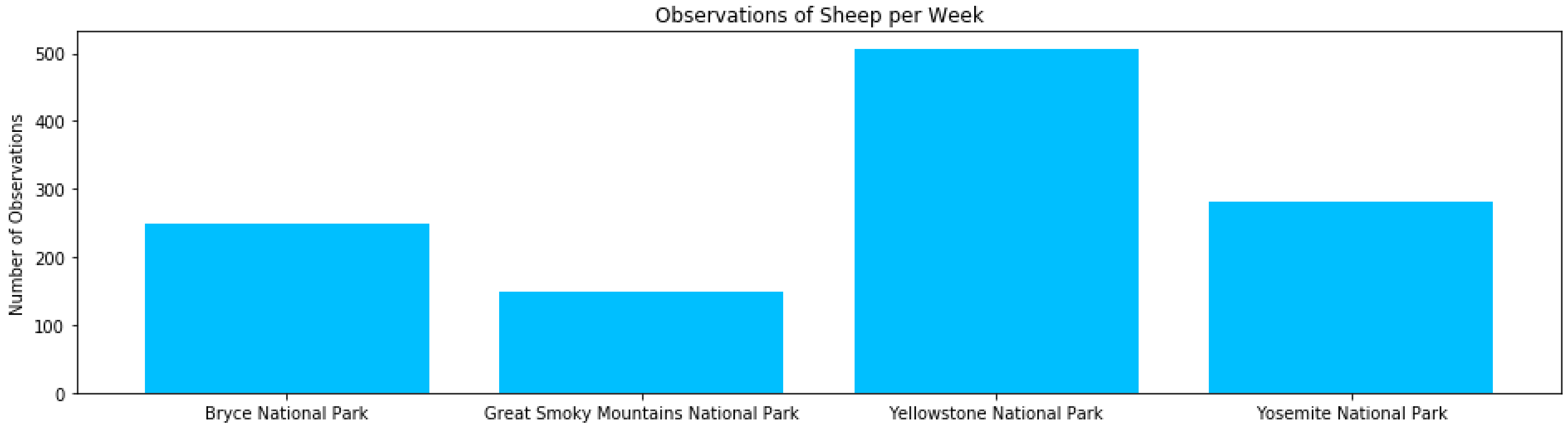
- Comparing the various species categories;
  - There is an statistically likely chance to pick a species at random from; Mammals, Birds, Amphibians and Fish that would be in the Protected classification. Conservation efforts will be similarly effective across these categories.
  - Mammals are statistically more protected than Reptiles. Efforts should be made to protect Mammals over Reptiles.
  - All Animal categories are statistically more protected than Vascular and Non-Vascular Plants.
  - There is no statistical significance between Vascular and Non-Vascular protection rates.

Sample Size Determination

- Sample Size Determination is extremely important in the world of statistics as it allows for the data to be statistically significant, therefore it should hold true for a wide population, while not being so large that it is infeasible to test.
- Formulas can be used to manipulate variables and determine the final sample population size for the study or experiment.
- In the Foot and Mouth (F&M) study we have the following variables;
  - Minimum Detectible Effect = (F&M Reduction) / (F&M Baseline) * 100 = 33.33%
  - Baseline = 15%
  - Significance = 90%
  - Sample Size Determination = 870 sheep to be observed

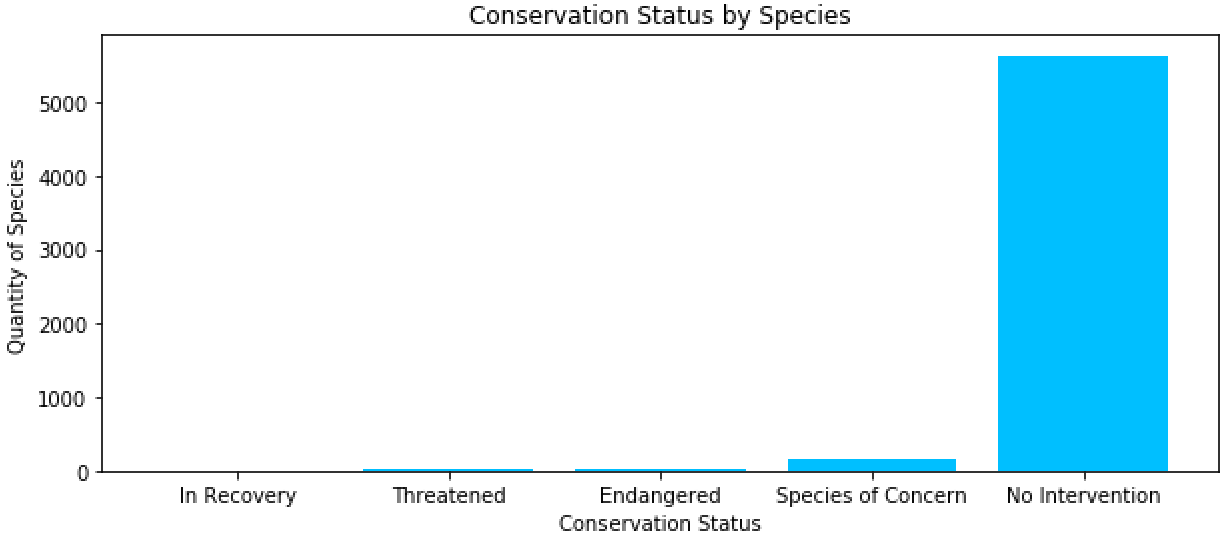
National Park Observations

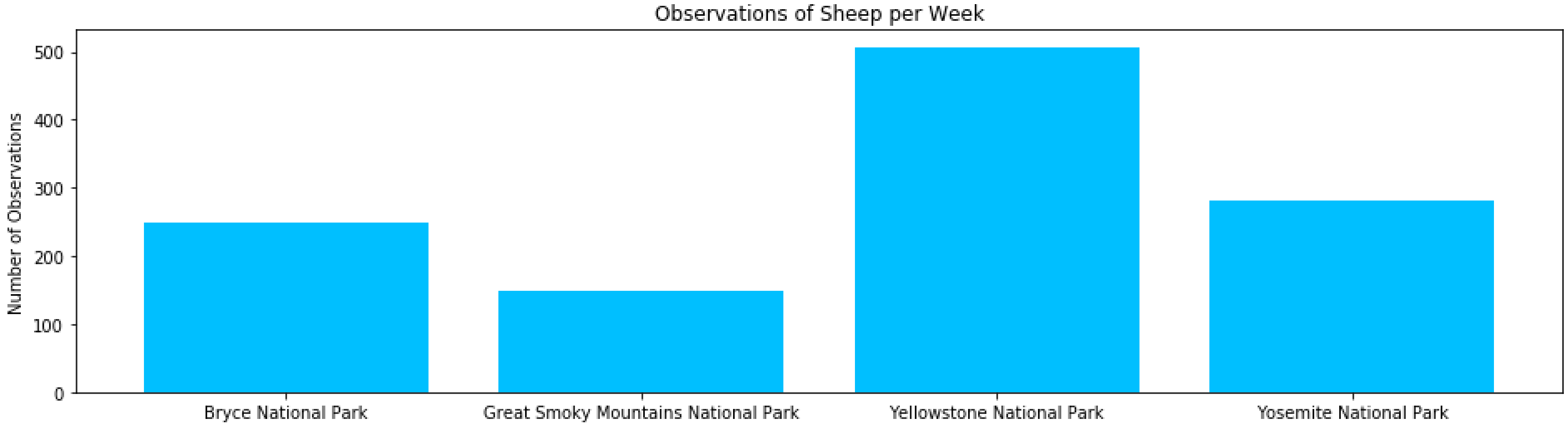
- Using this sample size and known observations, we can calculate the required length of time to observe enough sheep to notice a change in foot and mouth disease.



Park Name	# of Weeks to Observe
Bryce	3.48
Great Smoky	5.84
Yellowstone	1.72
Yosemite	3.09

Graph Generation


Conservation Status by Species
Quantity of Species
In Recovery
Threatened
Endangered
Species of Concern
No Intervention
Conservation Status


Observations of Sheep per Week
Number of Observations
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park