Introduction to Data Analysis

Biodiversity Capstone Project – Investigating Protected Species

Prepared by:
Lim Chong Han
11 January 2019

CodeCademy

Data collected on species

- The file species_info.csv contains information about:
  - Category of species
  - Scientific name
  - Common name
  - Conservation status

- Information could be grouped by conservation status for each unique species (scientific name). This will allow user to zoom into a specific conservation status (e.g. endangered). As there are empty/NaN cells for species which are not categorized, they should also be captured into the data frame (.fillna).

- It is also possible to examine the category of species (e.g. bird, fish, mammal) which are more likely to be endangered through a pivot table and statistical test (chi-square test for significance).

Significance calculations for endangered status between different categories of species

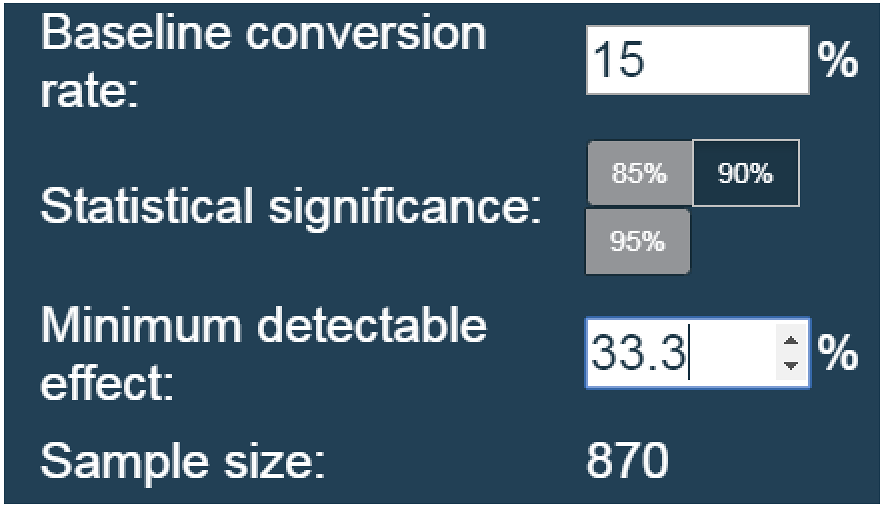
- The main question to be addressed is: Are certain types of species more likely to be endangered?
- There are 7 types of categories of species: amphibian, bird, fish, mammal, nonvascular plant, reptile, vascular plant. With a 7 choose 2 combination, there are 21 sets of comparisons between any of the two categories of species.
- As the comparison is based on categorical variables, a chi-square test is the most appropriate.
- Using the null hypothesis that there is no difference between the endangered status between different categories of species, we measure the p-value and compare it to the significance level that we can accept.
- Using the above approach and a significance level of 5%:
  - Birds and mammals: p-value of 0.688 -> Do not reject null hypothesis (above 5%), hence there is NO significant difference between their endangered status
  - Reptiles and mammals: p-value of 0.038 -> Reject null hypothesis (below 5%), hence there IS significant difference between their endangered status
- Therefore, we can conclude that certain types of species are more likely to be endangered than others.

Recommendation for conservationists

- We have concluded from previous section that significant differences exist between the endangered status of the different categories of species.
- Conservationists should then conduct the chi-square test for the different combinations of species, and identify the species which are more likely to be endangered than others.
- Once these species are identified, conservationists can zoom into that particular category of species and perform additional analyses.
- Such analyses could include identifying specific species within the category, and locating them by the parks or areas where they are observed.
- When information about sighting is available. conservationists can then devise a plan to conserve them, such as setting up protection area (from poaching etc.), monitoring their behavior to create conducive environment or remove obstacles.
- Based on data, mammals should be the category of species receiving the most attention.

Sample size determination for foot and mouth disease study

- Baseline conversion rate: The only information that the scientists currently have is that last year it was recorded that 15% of sheep at Bryce. This will form the baseline conversion rate.
- Statistical significance: Default level of 90% assumed.
- Minimum detectable effect: This is as a percentage of baseline conversion rate, and as scientists want to detect reductions of at least 5 percentage points, the minimum detectable effect is (5%) / (15%) * 100
- Based on the above input, the sample size is computed as 870.



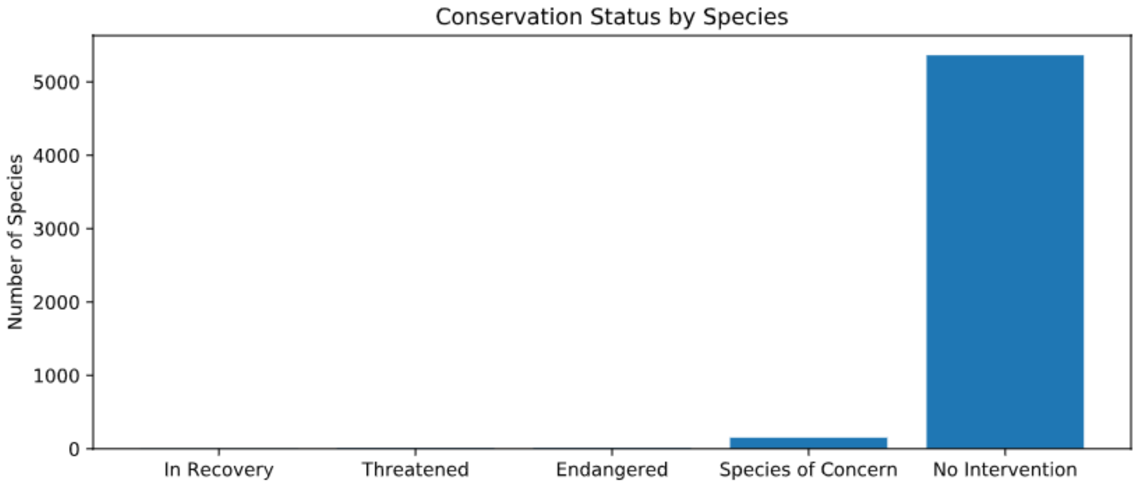
Baseline conversion rate: 15 %

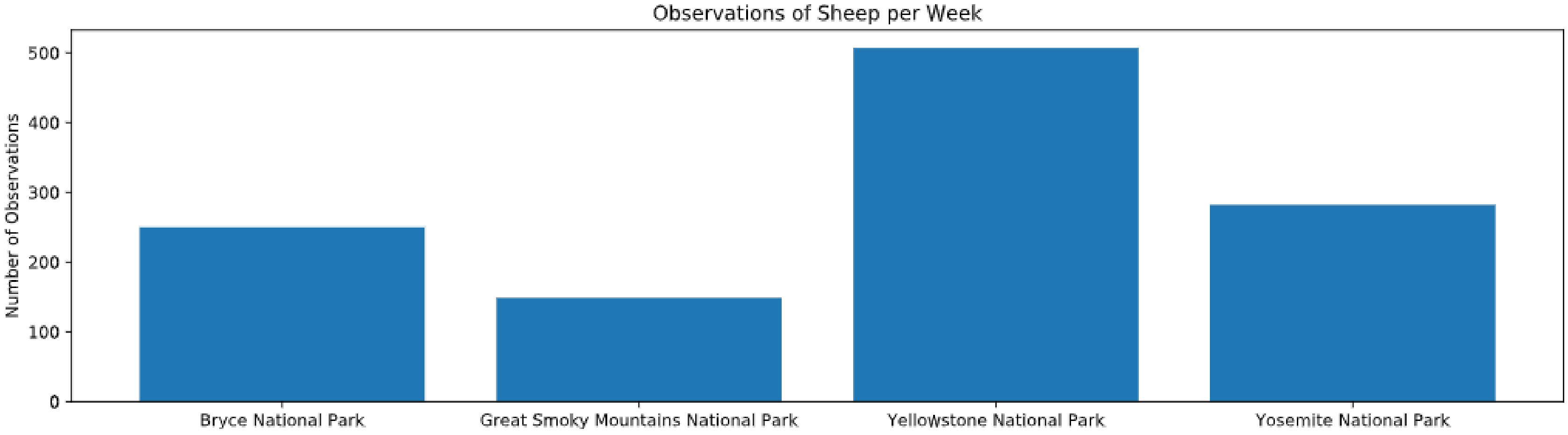
Statistical significance: 85% [✓ 90%] 95%

Minimum detectable effect: 33.3 %

Sample size: 870

Graphs created


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


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