Biodiversity for the National Parks

Codecademy Introduction to Data Analysis
Capstone Project option 2

Yoann Copreaux

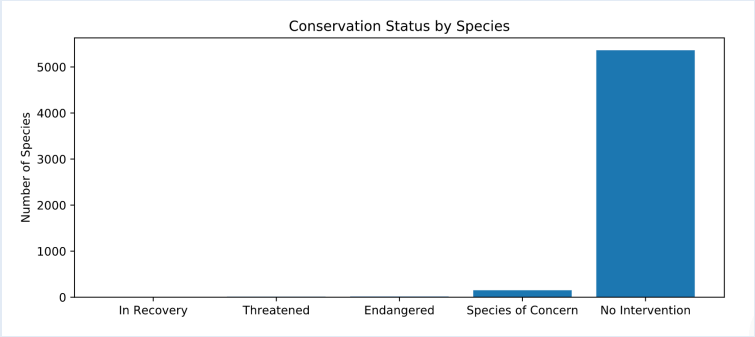
“species_info.csv” is providing detailed data on thousands of species and their endangered status

“species_info.csv” is providing detailed data on thousands of species and their endangered status…

- **7 categories of animals**: 'Mammal’, 'Bird’, 'Reptile’, 'Amphibian’, 'Fish’, 'Vascular Plant’, 'Nonvascular Plant’
- **5 conservation statuses**: ‘N/A’/’No Intervention’, 'Species of Concern’, 'Endangered’, 'Threatened’, 'In Recovery’
- **5824 species entries, of which over 5500 unique species** described by their scientific and common names:
  - 5541 scientific names
  - 5504 common names

… and only 3.2% of them require intervention

- No Intervention: 5363 species (96.8%)
- Species of Concern: 151 species (2.7%)
- Endangered: 15 species (0.3%)
- Threatened: 10 species (0.2%)
- In Recovery: 4 species (0.1%)



Given the type of data, the relative likelihood of categories of species to be endangered was analysed through a chi-squared test

Data was manipulated to get a clear pivot table of the categories of species and the number of species in each categories which were or not protected

Category	Nb of not protected species	Nb of protected species	% of protected species
Amphibian	72	7	8.86%
Bird	413	75	15.37%
Fish	115	11	8.73%
Mammal	146	30	17.05%
Nonvascular Plant	328	5	1.50%
Reptile	73	5	6.41%
Vascular Plant	4216	46	1.08%

In order to test whether certain categories of species were more or less likely to be endangered/ protected, the data was analysed through a chi-squared test, since:

- The data is categorical
- There are more than two pieces of data

Conservationists could split these categories in animals versus plants groups

P-value of the chi-squared test between each pair of categories of species

	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile	Vascular Plant
Amphibian							
Bird	0.176						
Fish	0.825	0.077					
Mammal	0.128	0.688	0.056				
Nonvascular Plant	0.002	0.000	0.000	0.000			
Reptile	0.781	0.053	0.741	0.038	0.034		
Vascular Plant	0.000	0.000	0.000	0.000	0.662	0.000	

- **The first very clear distinction is between animals and plants:**
  - The null hypothesis that the difference in data is due to chance is not rejected (p-value=0.662)
  - The null hypothesis between either category of plants is rejected with every animal categories (p-value <= 0.002, which is extremely low, except Reptile with Nonvascular Plant at p-value = 0.034, which is still statistically significant)

- **Within animal categories, two categories seem to slightly emerge, but the difference is not statistically significant (with the exception of one case):**
  - P-value between pairs from Fish, Amphibian and Reptile categories are fairly high (above 0.74, very high), so the null hypothesis is very clearly not rejected
  - Similarly, the p-value for the test between Bird and Mammal categories is high (0.688), so the difference in the data is most likely due to chance
  - Chi-squared test between two categories of these two groups show much lower p-values (below 0.18), although only one is below the standard threshold of 0.05 (between Reptile and Mammal).
  - **Multiplying tests and p-values calculations always compiles the risk of chance just having led to a p-value inferior to 0.05. Therefore, we can't conclude these categories are different with respect to being endangered. Further data may be required to infirm or confirm the existence of two sub-groups**

1 week of observation at Yellowstone National Park would be needed to observe enough sheep and determine program efficiency

Using the Sample Size Determination technique…

- **Baseline used: 15%** of sheep at Bryce National Park have foot and mouth disease
- **Park rangers want to be able to detect reductions of at least 5 percentage point** (for instance, if 10% of sheep in Yellowstone have foot and mouth disease). Therefore, the minimum detectable effect would be 33.33% (100*5%/15%)
- **90% confidence used to provide comfort to Park rangers**

… Park rangers would need to observe 510 sheeps to get confidence in the potential effect of their program

- Based on the assumptions explained, **the required sample size is 510 sheeps**
- Given observation levels at the various parks, Yellowstone National Park could be suitable, as it had the most observation (507) in the past week. **Based on this, it would take approximately a week of observation at Yellowstone National Park to get to the required sample size.**

Appendix: Graphs produced during analysis