Biodiversity for the National Parks

To protect or not to protect: time to answer the question

Project Contributors:
Mitch McKinnon, Codecademy, Unknown Park Observers

The Circle of Life:

Or the CSV of life

Data Structure

An Overview of the Dataset Features

*category**
A description of the animal class of the species. This includes: mammals, birds, reptiles, amphibians, fish, vascular plants, and nonvascular plants

scientific_name
The scientific name of the species listed.

common_names
The commonly used name(s) of the species listed.

*conservation_status***
The current conservation status of the species. This includes: species of concern, endangered, threatened, in recovery, and no intervention.

is_protected
Indicates if a species currently has a conservation status aside from the 'no intervention' label.

*Please note: categories do not appear in the exact format as written here for presentation readability purposes (i.e. 'mammals' is 'mammal' in the csv file
**We manually added the 'no intervention' label to the data for further analysis.

Numbers of Note

Number of Species

5541 Unique Species

Unprotected Species

5363 are not currently protected

Species Protection

Mammals are the most likely to be protected at a 10.27% protection rate

Species in Recovery

Only 4 species are currently in recovery

The Meaning of Life:

Diving into the numbers of species_info.csv

Analysis Methodology

Chi-squared Test

What is a Chi-squared Test?
A Chi-squared Test allows the data scientist to analyze multiple sets of similar data to see if there's a statistically significant difference between the sets .

Why are you using this test?
We chose to use this test because it allows us to:

- Look 2 or more categorical datasets
- Calculate a statistically significant difference in the form of a p-value

But wait, can't you use a binomial test?
Good try, but although binomial tests can test a categorical dataset, they cannot test multiple categorical datasets.

Analyzing the Data

Mammals V. Birds

p-value = 0.6876

Conclusion:
Given the p-value is > 0.05, we accept the null hypothesis that there is no statistically significant difference between the two datasets.

Mammals V. Reptiles

p-value = 0.0384

Conclusion:
Given the p-value is < 0.05, we reject the null hypothesis that there is no statistically significant difference between the two datasets.

*Please note: categories do not appear in the exact format as written here for presentation readability purposes (i.e. 'mammals' is 'mammal' in the csv file
**We manually added the 'no intervention' label to the data for further analysis.

Interpreting the Results

Looking at Significance

What can we conclude from the Chi-Squared Tests?
The results indicate that there's not a significant difference between the mammal and bird populations, but a significant difference exists between the mammal and reptile populations.

In our dataset, 10.27% of the mammals were protected while only 3.43% of the reptiles were afforded the same protections. We can assume that something other than chance is causing the reptile rate to be lower or causing the mammal rate to be higher.

10.27%

Mammal
Protection Rate

3.43%

Reptile
Protection Rate

Evolving:

Recommendation for the Future

Data-driven Recommendation

Three Approaches

Investigate biases in the process of classifying protection statuses
Given the significant difference between the protective status being applied to mammals more than reptiles, it's worthwhile to look at the classification process. There is some research that has been done in this area already*.

Evaluate the current approach to reptiles
Since the protection rate of reptiles is significantly lower than that of mammals, we can recommend looking at current approaches to conserving reptilian species and see where efforts can be applied to mammals.

Re-evaluate the current approach to mammals
As a result of a higher protection rate, it's worthwhile to look at current mammalian conservation efforts to see where they can be bolstered with resources or re-worked.

* https://www.tandfonline.com/doi/abs/10.1080/14888386.2011.642663

Sampling is Afoot:

A detour to calculate the sample size relating to foot and mouth disease among sheep

Determining Sample Size

Components

15%

Baseline Conversion Rate

The baseline conversion rate is provided as the current percent of sheep with foot and mouth disease: 15%.

33.33%

Minimum Detectable Effect

The Minimum Detectable Effect, also known as lift, is the percent of the conversion rate that we would like to see impacted:

$$\frac{0.05}{0.15} \times 100 = 33.33\%$$

90%

Statistical Significance

This number can vary depending on how confident we'd like to be in the outcome, 90% is a standard statistical significance. If this is increased for more certainty, sample size will increase.

Result

510

Sample Size

Using the three components to the left, we can use a tool such as the one at Optimizely to calculate this sample size.

The End

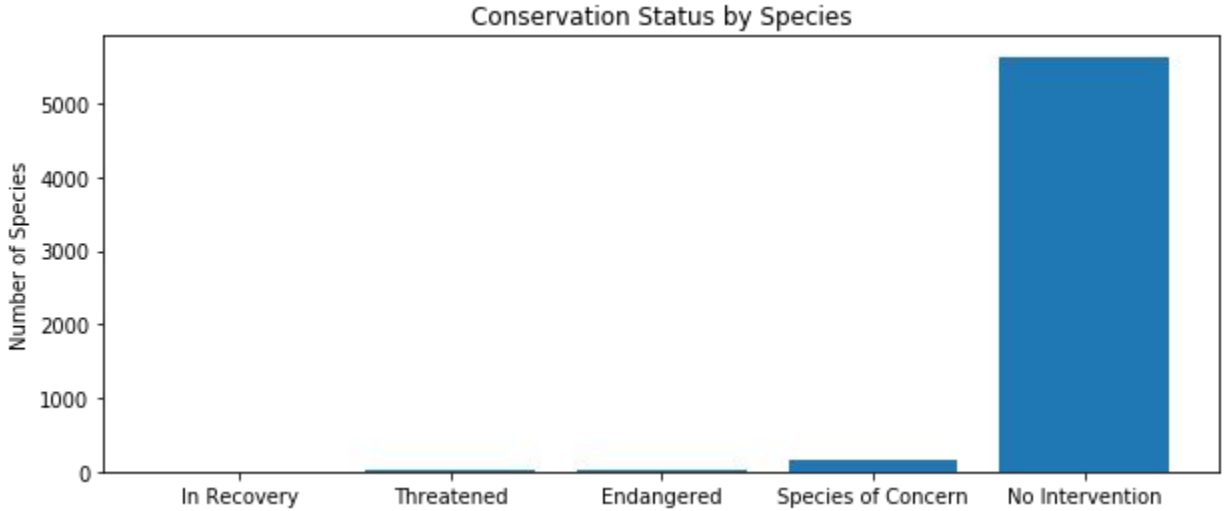
[‘Hip’, ‘Hip]

(read: “Hip, Hip, Array!)

Appendix

Charted Territories Ahead

Conservation Status by Species



Observations of Sheep per Week