# Biodiversity for the National Parks

Capstone Option 2
10/04/18

Image: https://commons.wikimedia.org/wiki/Yosemite_National_Park#/media/File:El_Capitan,_Yosemite_NP.jpg

# Objective

- Two data sets from the National Parks Service have been provided:
  - ***species.csv*** details the conservations status of different animal species
  - ***observations.csv*** which records data on sightings of different types of sheep from various National Parks

- By examining the species data, seek to determine whether certain species are more or less likely to be endangered

- The observations data can be used to determine the amount of time needed to detect the result of efforts to reduce foot and mouth disease

# Examining the data

Initial observations

# Observations from data -summary

| | category | scientific_name | common_names | conservation_status |
|---|---|---|---|---|
| 0 | Mammal | Clethrionomys gapperi gapperi | Gapper's Red-Backed Vole | nan |
| 1 | Mammal | Bos bison | American Bison, Bison | nan |

- Data about different species is included in the file **species_info.csv**
- This file includes scientific names, commons names, categories, and conservation status for a range of species.

- There are some duplicate values within this data.
- The figure shows the first two rows within the data for two different types of mammal
- If there is no conservation status, then a “nan” result is listed

# Observations from data - categories

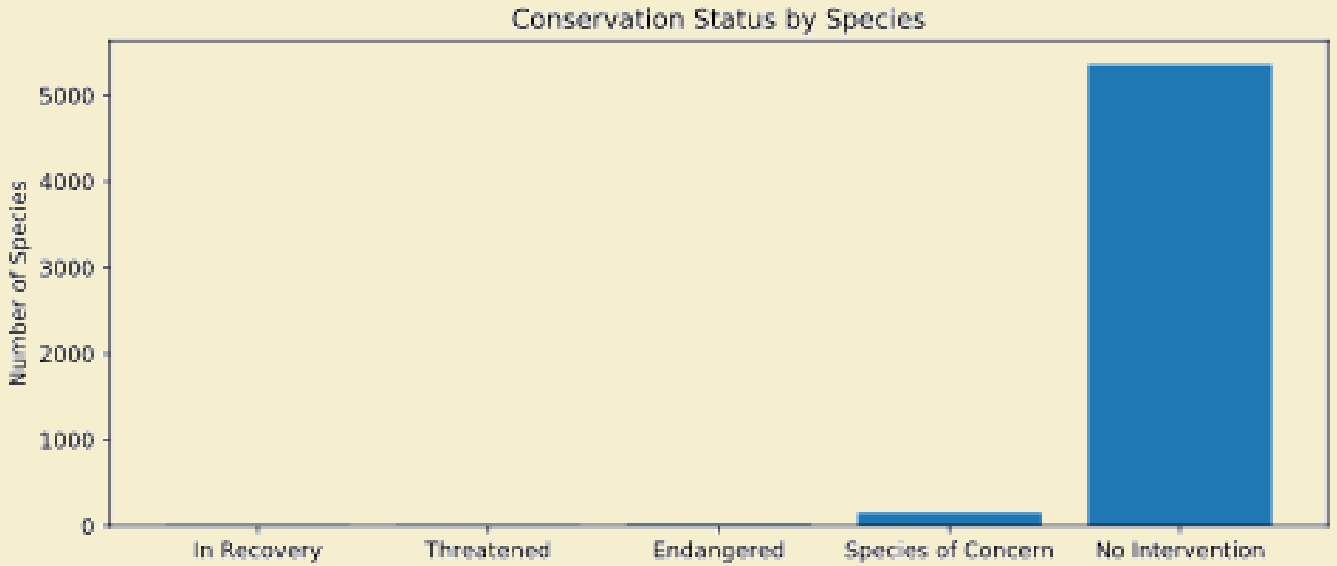
| | category | scientific_name | common_names | conservation_status |
|---|---|---|---|---|
| 0 | Mammal | Clethrionomys gapperi gapperi | Gapper's Red-Backed Vole | nan |
| 1 | Mammal | Bos bison | American Bison, Bison | nan |

- Data about different species is included in the file **species_info.csv**
- This file includes scientific names, commons names, categories, and conservation status for a range of species.

- Overall, there are 5541 ***unique*** specifies recorded.
- Categories are:
  - Mammal
  - Bird
  - Reptile
  - Amphibian
  - Fish
  - Vascular Plant
  - Nonvascular Plant

# Observations from data - conservation

| | category | scientific_name | common_names | conservation_status |
|---|---|---|---|---|
| 0 | Mammal | Clethrionomys gapperi gapperi | Gapper's Red-Backed Vole | nan |
| 1 | Mammal | Bos bison | American Bison, Bison | nan |




- Conservation statuses are ‘Species of Concern’, ‘Endangered’, ‘Threatened’, ‘In Recovery’ and ‘Null’
- There are:
  - 15 endangered
  - 4 in recovery
  - 151 of concern
  - 10 threatened
- The rest have no conservation status – these are updated to have a status of “No Intervention”. There are 5363 of these.

# Different types of species

Are certain types of species more likely to be endangered than others?

# Different types of species - data

| Category | Not Protected | Protected | Percent Protected |
|---|---|---|---|
| Amphibian | 72 | 7 | 8.75% |
| Bird | 413 | 75 | 15.16% |
| Fish | 115 | 11 | 8.66% |
| Mammal | 146 | 30 | 17.76% |
| Nonvascular Plant | 328 | 5 | 1.50% |
| Reptile | 73 | 5 | 6.33% |
| Vascular Plant | 4216 | 46 | 1.03% |

- All animals with a conservation status of "No Intervention" are "Not protected"
- This table shows the percentage of species protected by each type.
- It appears that birds and mammals are more protected than amphibians, fish, and reptiles, and even more so than plants.
- How significant is this?

# Different types of species - tests

| Category | Not Protected | Protected | Percent Protected |
|---|---|---|---|
| Amphibian | 72 | 7 | 8.75% |
| Bird | 413 | 75 | 15.16% |
| Fish | 115 | 11 | 8.66% |
| Mammal | 146 | 30 | 17.76% |
| Nonvascular Plant | 328 | 5 | 1.50% |
| Reptile | 73 | 5 | 6.33% |
| Vascular Plant | 4216 | 46 | 1.03% |

- A Chi-Squared test can be used to assess the significance of this effect.
- This test can be used to assess multiple samples of categorial data
- The *null hypothesis* is that the difference between samples is due to chance. A higher p-value means that the null hypothesis is more likely to be true
- Typically, a p-value of less than 5% is considered significant

# Different types of species – test 1

| Category | Not Protected | Protected | Percent Protected |
|---|---|---|---|
| Amphibian | 72 | 7 | 8.75% |
| Bird | 413 | 75 | 15.16% |
| Fish | 115 | 11 | 8.66% |
| Mammal | 146 | 30 | 17.76% |
| Nonvascular Plant | 328 | 5 | 1.50% |
| Reptile | 73 | 5 | 6.33% |
| Vascular Plant | 4216 | 46 | 1.03% |

- First, mammals and birds were compared by putting the data into a contingency table.
- A chi-squared test was run, and the p-value was determined to be 68.8%.
- This is much higher than 5% - the difference in the data for birds and mammals is not significant.
- This is not surprising – the percentage protected is similar for both categories.

# Different types of species – test 2

| Category | Not Protected | Protected | Percent Protected |
|---|---|---|---|
| Amphibian | 72 | 7 | 8.75% |
| Bird | 413 | 75 | 15.16% |
| Fish | 115 | 11 | 8.66% |
| Mammal | 146 | 30 | 17.76% |
| Nonvascular Plant | 328 | 5 | 1.50% |
| Reptile | 73 | 5 | 6.33% |
| Vascular Plant | 4216 | 46 | 1.03% |

- A second test was completed for Reptiles and Mammals.
- In this case, the P-value is 3.8%. This is less than 5% and therefore the Null Hypothesis – that the difference is due to chance – can be rejected.
- The alternative hypothesis is that this difference is not due to chance.

# Recommendations for conservationists

- The data suggests that different types of species are more/less likely to be endangered than other types of species.
- This is verified by the data for reptiles and mammals – mammals have a percentage protected of 17.76% and reptiles have a percentage protected of 6.33%. The difference between these two populations is significant, based on the chi-squared test (which returned a p-value of 3.8%).
- The obvious recommendation is that conservationists should focus their efforts on certain types of species (e.g. mammals and, subject to further analysis, birds). It would also be worth trying to determine *why* certain types of species are more likely to be endangered.
- However, conservationists should also reflect on the quality of the data. This data does not actually tell us which types of species are more likely to be endangered, but rather, which types of species are more likely to be protected. Other factors may therefore explain these differences.
- For example, it is possible that people have more of an affinity for mammals compared to reptiles and are therefore more likely to protect them.

# Sample size determination

Observing the effect of efforts to reduce foot and mouth disease

# Sample size determination - observations

| scientific_name | park_name | observations |
|---|---|---|
| Vicia benghalensis | Great Smoky Mountains National Park | 68 |
| Neovison vison | Great Smoky Mountains National Park | 77 |
| Prunus subcordata | Yosemite National Park | 138 |
| Abutilon theophrasti | Bryce National Park | 84 |
| Githopsis specularioides | Great Smoky Mountains National Park | 85 |
| … | … | … |

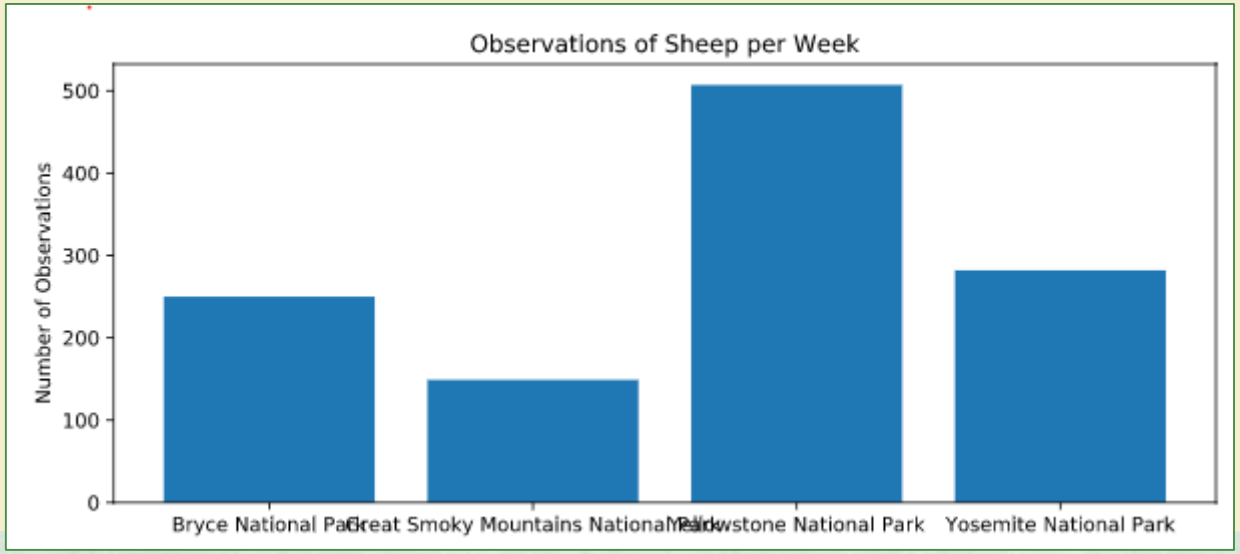
- A second data set on observations of different breeds of sheep at various national parks was provided.
- An extract is shown here

- Within the original species data set, mammals with the string ‘sheep’ within their common name are identified. It is necessary to limit this to mammals, as some plant species also have this in their name.

# Sample size determination - observations

| park name | observations |
|---|---|
| Bryce National Park | 250 |
| Great Smoky Mountains National Park | 149 |
| Yellowstone National Park | 507 |
| Yosemite National Park | 282 |

- By combining the two data sets, the total number of sheep observations in each national park can be determined.
- This is plotted below



# Sample size determination - calculation

- A calculation was undertaken to determine the expected time required to make observations of sheep in different national parks, in order to check for reductions in foot and mouth disease.
- The baseline rate, from previous years, is 15% and the national parks service want to be able to observe a reduction to this rate of at least 5%, with a 90% significance.
- A 5% change from a baseline of 15% is 33%. To detect an effect of at least this size, a sample of at least 870 sheep is required.
- With the weekly observations previous calculated, this would take 1.72 weeks at Yellowstone National Park and 3.48 weeks at Bryce National Park.

# Thank You