Biodiversity for the National Parks

Andreea Chidu

- The Data

The Data

Species Data Frame

- Fields
  - Category
  - Scientific_Name
  - Common_Names
  - Conservation_Status

	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
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

Figure 1. Head of Species Data Frame

The Data

Species Data Frame

- 7 species types
  - Mammal
  - Bird
  - Reptile
  - Amphibian
  - Fish
  - Vascular Plant
  - Nonvascular Plant

The Data

Species Data Frame

- 5541 Unique Species
  - Note: NaN values in the Conservation_Status field were replaced with string ‘No Intervention.’ Had I tried to count the total number of unique species using the Group By function to create a new data frame with the unique count of Scientific_Name grouped by Conservation_Status without replacing the NaN values, we would have only counted 180 unique species because NaN values are not included in the count (see Figure 2.)
- The majority of the species (97%) in this data frame are not under conservation

The Data

Species Data Frame

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

Figure 2. Unique Count of Species Grouped By Conservation_Status *before* replacing NaN With a String; 'conservation_counts' Data Frame

The Data

Species Data Frame

	conservation_status	scientific_name
1	In Recovery	4
4	Threatened	10
0	Endangered	15
3	Species of Concern	151
2	No Intervention	5363

Figure 3. Unique Count of Species Grouped By Conservation_Status *after* replacing NaN With a String; ‘conservation_counts_fixed’ Data Frame

- Extracting Value

Extracting Value

Are certain species more susceptible to endangerment?

- To visualize the data, I created a bar graph (Figure 4.) to compare the total number of species in each conservation group
- Used 'conservation_counts_fixed' data frame to include species that have no intervention status (Figure. 3)
- Before I created the graph, I sorted the data frame by count in ascending order to provide more value (instead of the default alphabetical order)

Extracting Value

Are certain species more susceptible to endangerment?

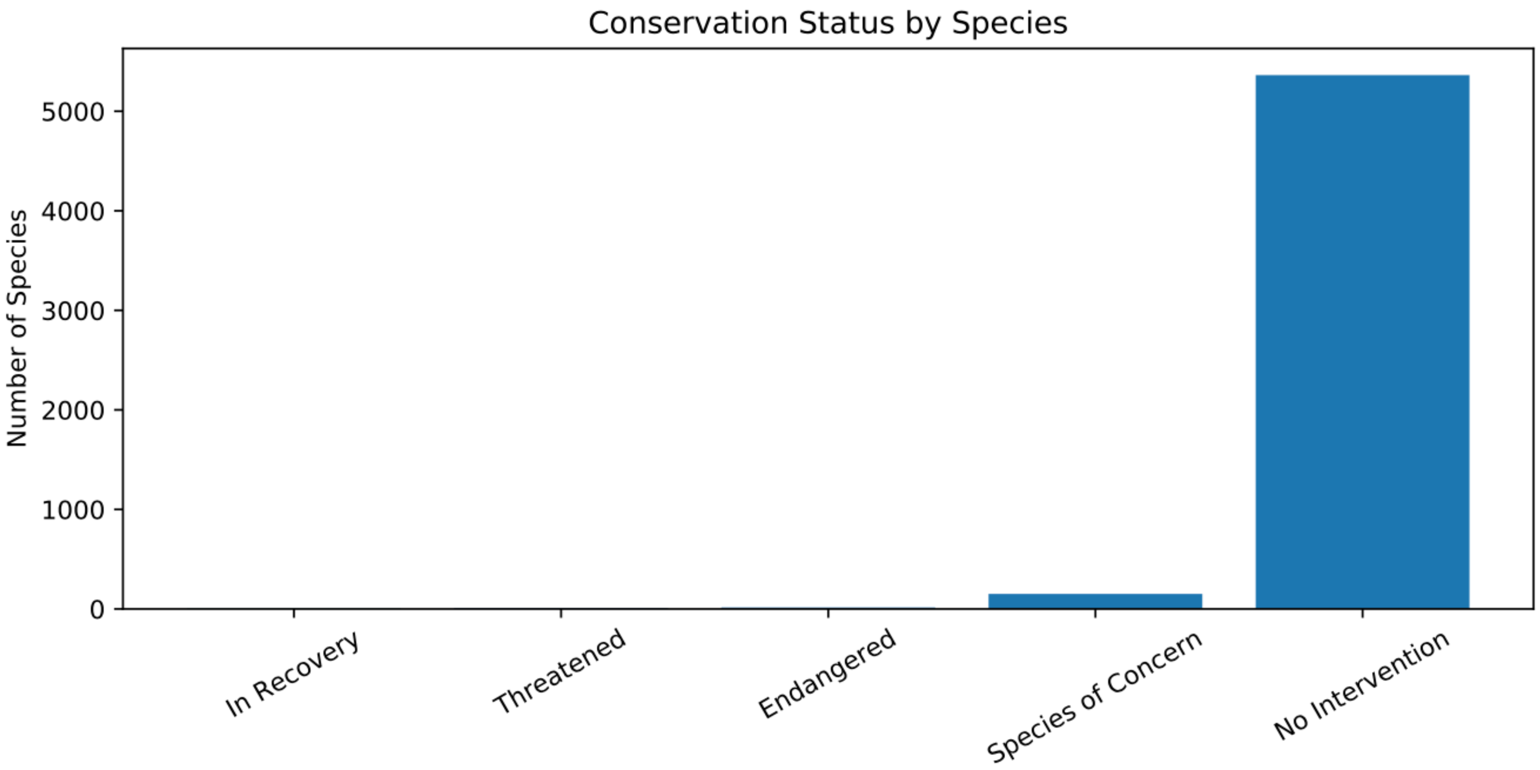

Figure 4. Conservation Status by Species

Extracting Value

Are certain species more susceptible to endangerment?

- To gain more insight into specific species' likelihood to be endangered, I created a pivot table (Figure 5.)
- The pivot table is easier to read and provides data that is more straight-forward

	category	not_protected	protected
0	Amphibian	72	7
1	Bird	413	75
2	Fish	115	11
3	Mammal	146	30
4	Nonvascular Plant	328	5
5	Reptile	73	5
6	Vascular Plant	4216	46

Figure 5. Category Pivot Table

Extracting Value

Are certain species more susceptible to endangerment?

- To make the data even more useful, I added a new column to the pivot table that contained percentage values (% of each species that is protected) (Figure 6.)

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

Figure 6. Category Pivot Table With Percent Column

Extracting Value

Are certain species more susceptible to endangerment?

- At first glance, it looks as though mammals and birds are much more likely to be protected (17 and 15 percent, respectively)
- To test this hypothesis, I needed to check if the differences between the number protected versus not protected for each species were due to chance or statistically significant using the pivot table
- Because the data I am using for this test is categorical and I am comparing four pieces of data (the number protected and not protected for both categories that I am comparing at a time), I need to use a Chi-Squared test

Extracting Value

Are certain species more susceptible to endangerment?

- Chi-Squared Results:
  - Mammal / Bird [p-value = .69]
  - Mammal / Reptile [p-value = .04]
  - Mammal / Fish [p-value = .06]
  - Mammal / Vascular Plant [p-value = 1.4e-55]
  - Bird / Fish [p-value = .08]
  - Bird / Reptile [p-value = .05]
  - Bird / Amphibian [p-value = .18]
  - Bird / Vascular Plant [p-value = 4.6e-79]
  - All animals / all plants [p-value = 3.2e-85]

Extracting Value

Are certain species more susceptible to endangerment?

- P-values less than .05 reject the null hypothesis and highlight a statically significant difference between the two species' likelihood to be endangered
  - Mammal / Bird [p-value = .69] not significantly different
  - Mammal / Reptile [p-value = .04] **significantly different**
  - Mammal / Fish [p-value = .06] **significantly different**
  - Mammal / Vascular Plant [p-value = 1.4e-55] **significantly different**
  - Bird / Fish [p-value = .08] not significantly different
  - Bird / Reptile [p-value = .05] not significantly different
  - Bird / Amphibian [p-value = .18] not significantly different
  - Bird / Vascular Plant [p-value = 4.6e-79] **significantly different**
  - All animals / all plants [p-value = 3.2e-85] **significantly different**

- Recommendation

Recommendation

To conservationists concerned about endangered species

- Based on my significance calculations, it appears as though Mammals are more endangered than the other species.
- Animals are much more likely to be endangered than both plant species.
- I would recommend that the conservationists direct most of their efforts towards Mammals, primarily, followed by Birds.

- The Data Pt.2

The Data

Observations Data Frame

- Fields
  - Scientific_Name
  - Park_Name
  - Observations

	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85
5	Elymus virginicus var. virginicus	Yosemite National Park	112

Figure 7. Head of Observations Data Frame

The Data

Observations Data Frame

- Because the scientists wanted data from both the Species and Observations data frames, I merged the two together.
- However, the Observations data frame only contains the scientific name of each species
- Before I merged the data frames I created a new column in Species to tell me (True or False) if the species name contains "Sheep."
- I created a new data frame that only contained data from the Species data frame that 'Mammal' as the category and 'True' for the new Is_Sheep column.
- I then merged this new data frame with the Observations data frame.
- The new data frame only contains data for sheep from both the Observations and Species data frames

- Extracting Value

Extracting Value

How many total sheep sightings across all three species?

- Before creating a meaningful visual to portray the sheep sightings across all three species, I had to organize the data in a way that would be best to work with.
- I created a new data frame called Obs_by_Park that returned the sum of Observations grouped by Park (Figure 8.)

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

Figure 8. Number of Observations by Park

Extracting Value

How many total sheep sightings across all three species?

- Using the data frame in Figure 8., I created a bar graph to show the number of sightings per week at each of the four national parks (Figure 9.)

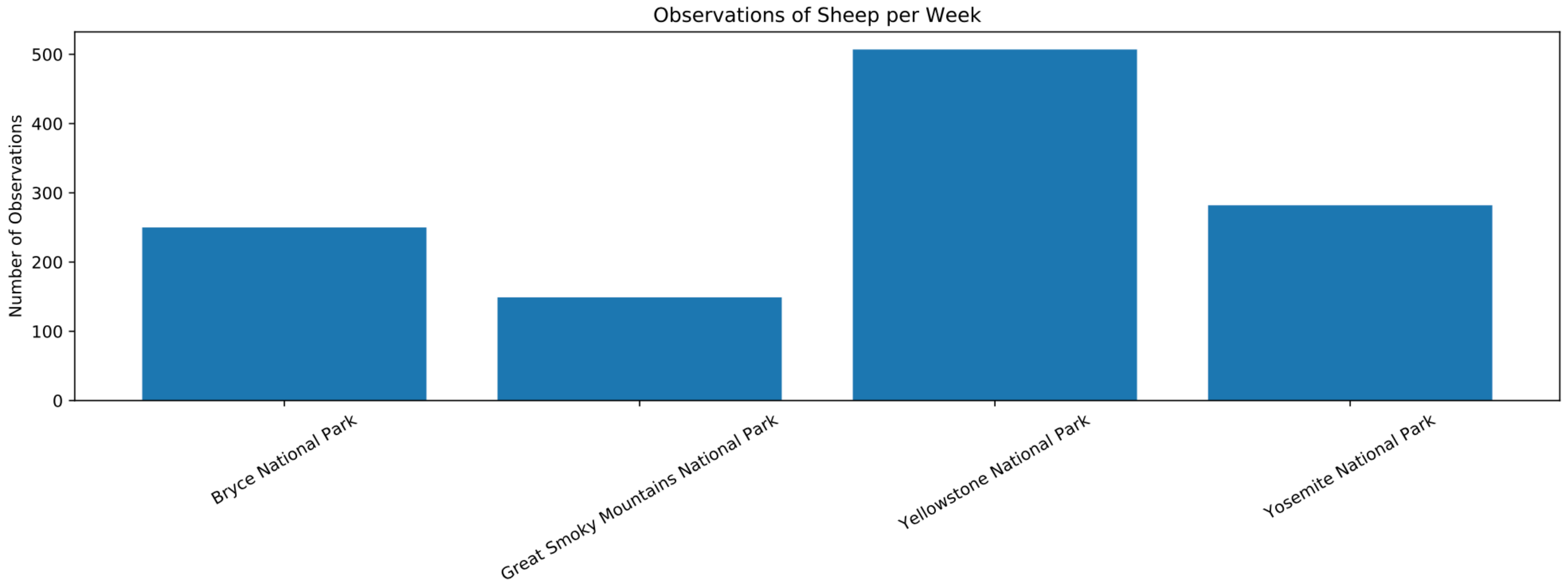

Figure 9. Observations of Sheep per Week

Extracting Value

Foot and Mouth Reduction Effort

- Overview
  - The scientists wanted to know if the program designed to reduce foot and mouth disease at Yellowstone National Park is working.
  - To do this, they need to observe a specific amount of sheep for a specific amount of time.

Extracting Value
Foot and Mouth Reduction Effort

- Using the sample size calculator with the below values (Figure 10.), I was able to determine that the scientists need to observe at least 870 sheep.

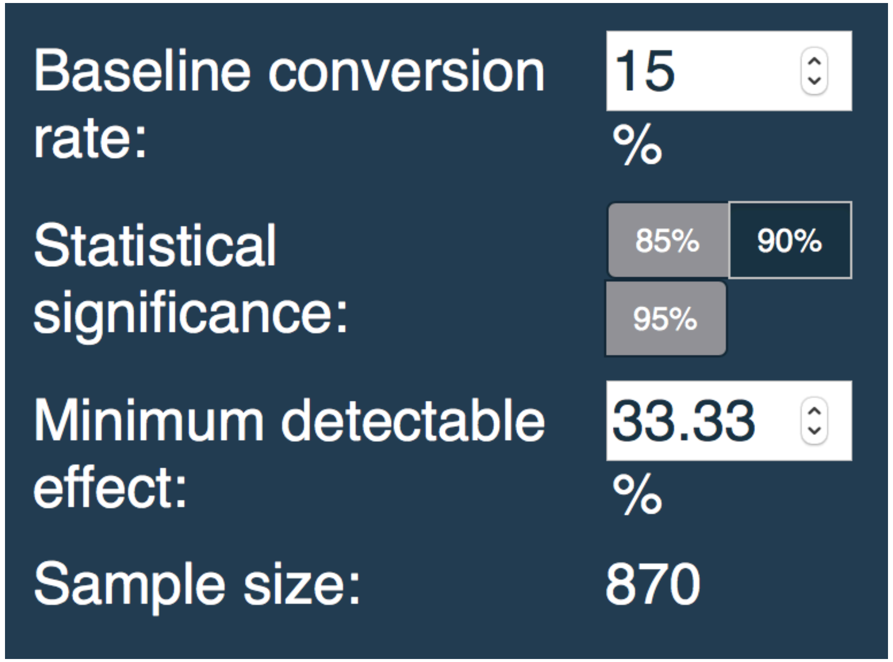

Figure 10. Observations of Sheep per Week

- 15% of sheep at Bryce Canyon National Park have foot and mouth disease
- Want to use the default level of 90%
- Looking for a 5% change – 5% change from 15% is 1/3 or 33.33%
- The output sample size with the above inputs

Extracting Value

Foot and Mouth Reduction Effort

- Using the number of observations of sheep per park from the Obs_by_Park data frame, I calculated how many weeks it would take to observe 870 sheep:
  - Yellowstone National Park ~ 2 weeks
  - Bryce Canyon National Park ~ 3.5 weeks
- These calculations were made by dividing the number of sheep the scientists need to observe by the number of sheep observed per week at the respective parks.

Thank you for an awesome course! I hope to take another intensive and the future, and will be thrilled if it’s as good as this one was.

On behalf of all my new skills & knowledge, thank you for all your hard work!

Andreea