BIODIVERSITY CAPSTONE PROJECT

MARY EATON

16 MARCH 2018

OBJECTIVES – PART I

- Part I: Biodiversity in National Parks – Investigating Protected Species
  - Using species data collected by the National Parks Service, analyze conservation statuses across species categories
  - Investigate if there are any patterns or themes to the types of species that become endangered
  - Recommend focus for conservationists concerned about endangered species

OBJECTIVES – PARTS II & III

- Part II: Discovering locations and count of specific species
  - Using a week of observation data, establish location and count of sheep in National Parks
  - Provide bar chart of results
- Part III: Assisting program evaluation of foot and mouth disease reduction in National Parks
  - Assist Program evaluation – is reduction effort working
  - Determine sample size of sheep observations and time needed to investigate efficacy of foot and mouth treatment

PART I: BIODIVERSITY IN NATIONAL PARKS – SPECIES DATA COLLECTED

- National Parks Service provided file *species_info.csv*, which provided category and conservation status data on 5541 species tracked in the National Parks.
- The following information was in the file for each species:
- (1) Category of species (category), which is one of the following:
  - Mammal
  - Bird
  - Reptile
  - Amphibian
  - Fish
  - Vascular Plant
  - Nonvascular Plant

PART I: BIODIVERSITY IN NATIONAL PARKS – SPECIES DATA COLLECTED II

- (2) Scientific name of species (scientific_name), which is a unique name
- (3) Common names of species (common_names), with multiple common names possible
- (4) Conservation status (conservation_status), which is one of following:
  - No Intervention (listed as null or NaN on input file)
  - Species of Concern - declining population or appears to be in need of conservation
  - Endangered - seriously at risk of extinction
  - Threatened - vulnerable to endangerment in the near future
  - In Recovery - formerly Endangered, but currently not in danger of extinction throughout all or a significant portion of its inhabitable range

Example of some rows of data in this file are on the next slide.

PART I: BIODIVERSITY IN NATIONAL PARKS – EXAMPLE OF DATA FROM SPECIES_INFO.CSV

	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), Dom...	NaN
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	NaN
4	Mammal	Cervus elaphus	Wapiti Or Elk	NaN

PART I: BIODIVERSITY IN NATIONAL PARKS – DATA INVENTORY

- How many of each species fall into the designated categories?
- The distribution of tracked species by category is:

Category	Count of Species
Amphibian	79
Bird	488
Fish	125
Mammal	176
Nonvascular Plant	333
Reptile	78
Vascular Plant	4262
	5541

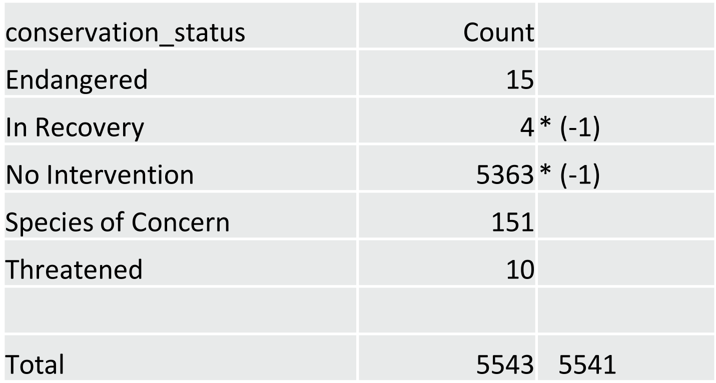
- The majority of species tracked are plants, especially vascular plants.

PART I: BIODIVERSITY IN NATIONAL PARKS – RESEARCH QUESTION

ARE SOME CATEGORIES OF SPECIES MORE THREATENED THAN OTHERS?

PART I: BIODIVERSITY IN NATIONAL PARKS – COUNT OF CONSERVATION STATUS BY SPECIES

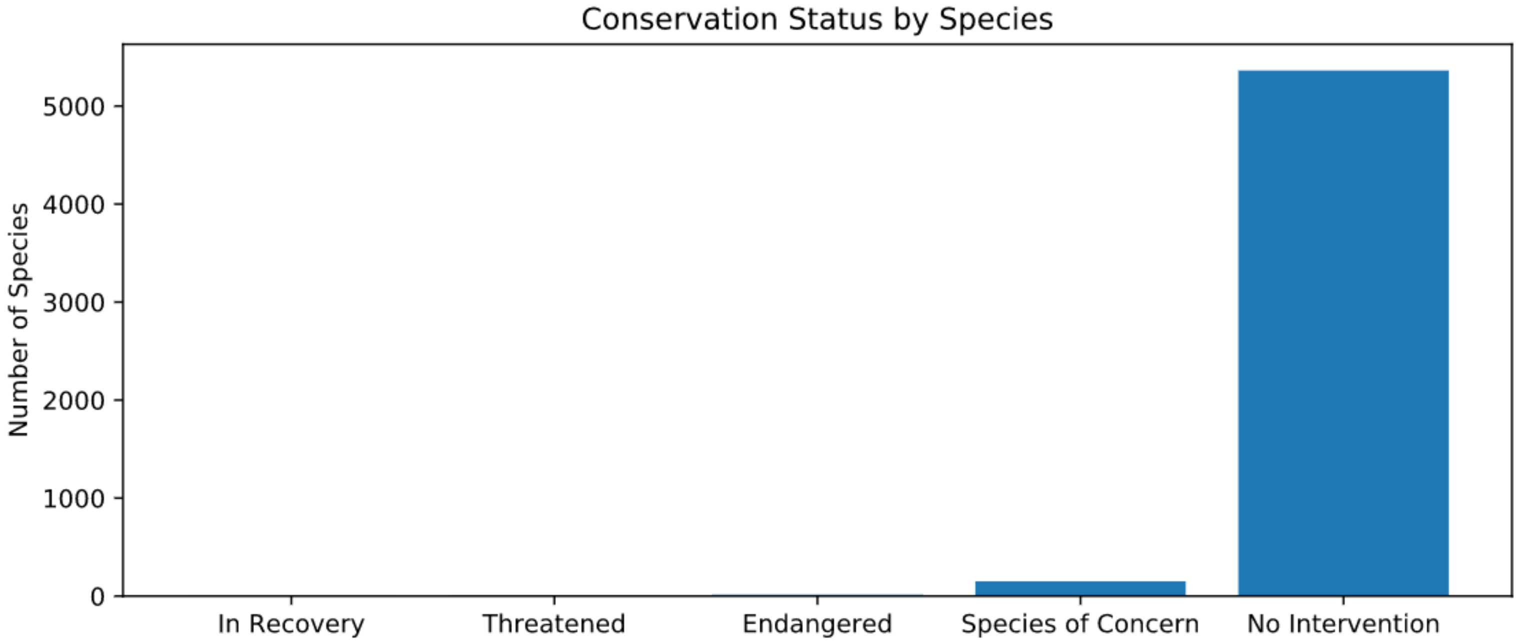
- How many of each species fall into these conservation statuses?
- The 5541 species have the following distribution of conservation statuses:



conservation_status	Count	
Endangered	15	
In Recovery	4	* (-1)
No Intervention	5363	* (-1)
Species of Concern	151	
Threatened	10	
Total	5543	5541

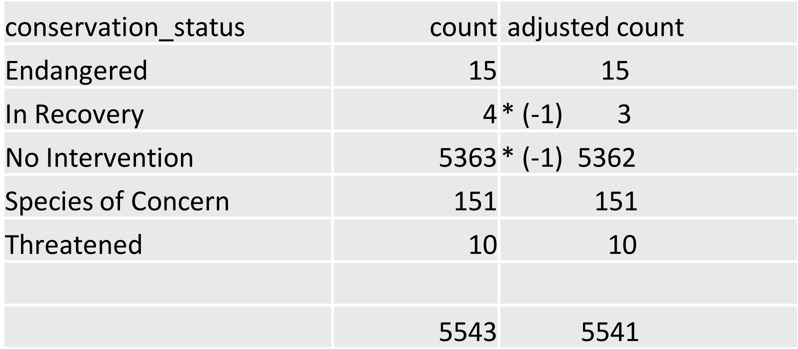
- This information is presented graphically on the next slide.
- All null values in conservation status were assigned “No Intervention.”
- NOTE: The asterisks represent adjustments due to duplicate conservation statuses on same species. See explanation on slide after graph.

PART I: BIODIVERSITY IN NATIONAL PARKS – DISTRIBUTION OF CONSERVATION STATUSES BY NUMBER OF SPECIES



PART I: BIODIVERSITY IN NATIONAL PARKS – *EXPLANATION OF COUNT DISCREPANCIES

- There are 5541 species, yet the total on the initial distribution yields a count of 5543.

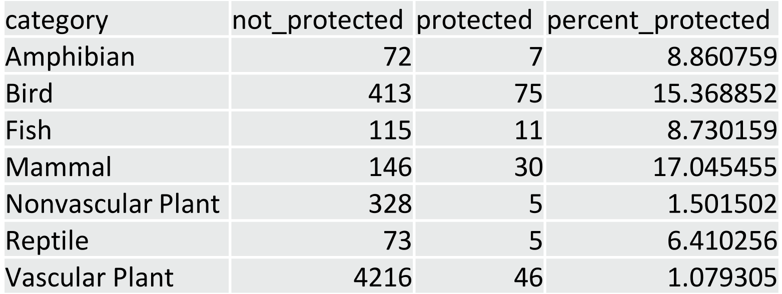


conservation_status	count	adjusted count
Endangered	15	15
In Recovery	4	* (-1) 3
No Intervention	5363	* (-1) 5362
Species of Concern	151	151
Threatened	10	10
	5543	5541

- Two species had multiple conservation status values:
  - Species *Canus lupi* (gray wolf) was listed three times: twice as being “Endangered,” and once as “In Recovery.” The “In Recovery” count was removed for that species in this chart (not in dataset).
  - Similarly, the species *Oncorhynchus mykiss* (rainbow trout) was listed twice, once as “Threatened” and once as null. The “Threatened” status was preserved, and the null (“No Intervention”) was removed from the count in the chart (data set was not modified).

PART I: BIODIVERSITY IN NATIONAL PARKS –
ARE CERTAIN TYPES OF SPECIES MORE LIKELY TO BE ENDANGERED?

- To answer this question, combined all variations of protection status into one status of "Protected," and those with status of "No Intervention" were considered "Not Protected."
- For each category, calculated the count and percentage of protected versus not protected species:



category	not_protected	protected	percent_protected
Amphibian	72	7	8.860759
Bird	413	75	15.368852
Fish	115	11	8.730159
Mammal	146	30	17.045455
Nonvascular Plant	328	5	1.501502
Reptile	73	5	6.410256
Vascular Plant	4216	46	1.079305

- The highest percentage of species in protected status are in the categories of Mammals (17.0%) and Birds (15.4%).
- Are Mammals more likely to be endangered than Birds? Performed a chi-squared test, which indicated the difference between these two categories is not significant (p-value of 0.6876).
- Are Mammals more likely to be endangered than Reptiles? A separate chi-squared test between Mammals and Reptiles indicates that Mammals are significantly more likely to have a protected status than Reptiles (p-value of 0.0384).

PART I: BIODIVERSITY IN NATIONAL PARKS – RECOMMENDATION FOR CONSERVATIONISTS

- Mammals and birds are the categories with more species in protected status
- Given limited resources, more consideration should be given to conservation efforts to species in these two categories
- Conservationists could also consider whether there is any bias in initial assignment of conservation status, which may favor animals over plants, and favor more familiar (or more similar to human) animal categories over the lesser known or less similar

PART II: DISCOVERING LOCATIONS AND COUNT OF SPECIFIC SPECIES – SHEEP OBSERVATIONS

PART II: DISCOVERING LOCATIONS AND COUNT OF SPECIFIC SPECIES – OBSERVATIONS OF SPECIES FROM SEVERAL NATIONAL PARKS

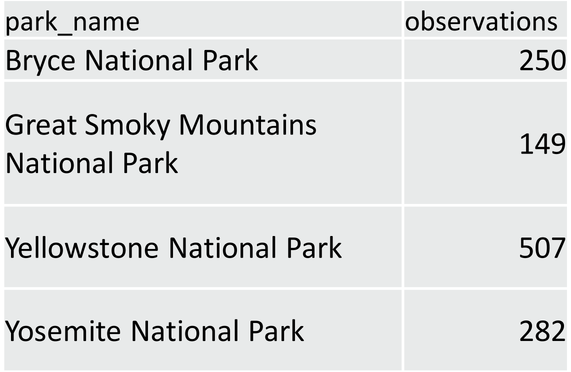
- Conservationists have been recording sightings of different species at several national parks for the past 7 days, recording their data in the file “observations.csv.” Sample data from this file is below:

	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

- Observations data file contains the following fields of data:
  - Scientific name (scientific_name), unique to species
  - National Park in which observation occurred (park_name)
  - Number of observations in 7 day period (observations)

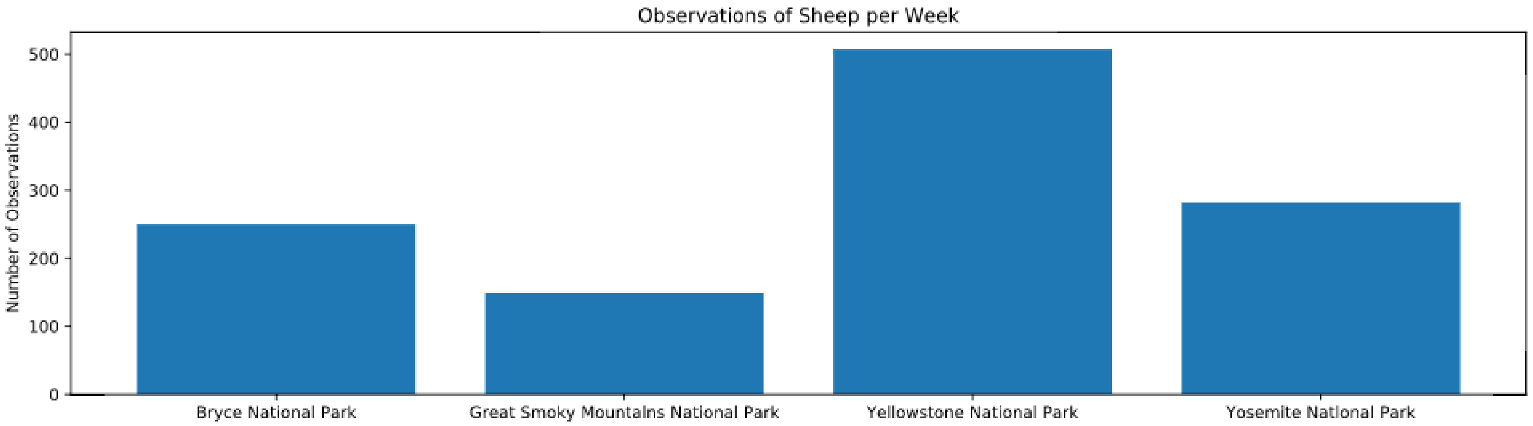
PART II: DISCOVERING LOCATIONS AND COUNT OF SPECIFIC SPECIES – USING THE FILES TO DISCOVER THE LOCATION AND COUNT OF SHEEP

- A team of scientists is interested in the count and location of various species of sheep across different national parks.
- Using the observation data from the new observations file and the common name data from the species file, we can find the counts of all sheep in the various observed National Parks for the last 7 days.
- Why need both the species and the observation files? Since only the scientific name is available in the observations file, selected and flagged the sheep species based on the common name containing the word “sheep” and the category “mammal” in the species file. Note that using the category in the search was needed to exclude other categories that had the word sheep in the common name, such as the plant “common sheep sorrel.”
- Since the scientific name is unique to a species, and is common to both files, we merged the data from the sheep species selection (from the species file) to the observation data on the same species to get the list of observations for sheep only.
- By grouping and summing the observation counts by National Park name, we get the distribution of sheep across the National parks in the last 7 days.



park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

PART II: DISCOVERING LOCATIONS AND COUNT OF SPECIFIC SPECIES – COUNT OF SHEEP SPECIES BY NATIONAL PARK



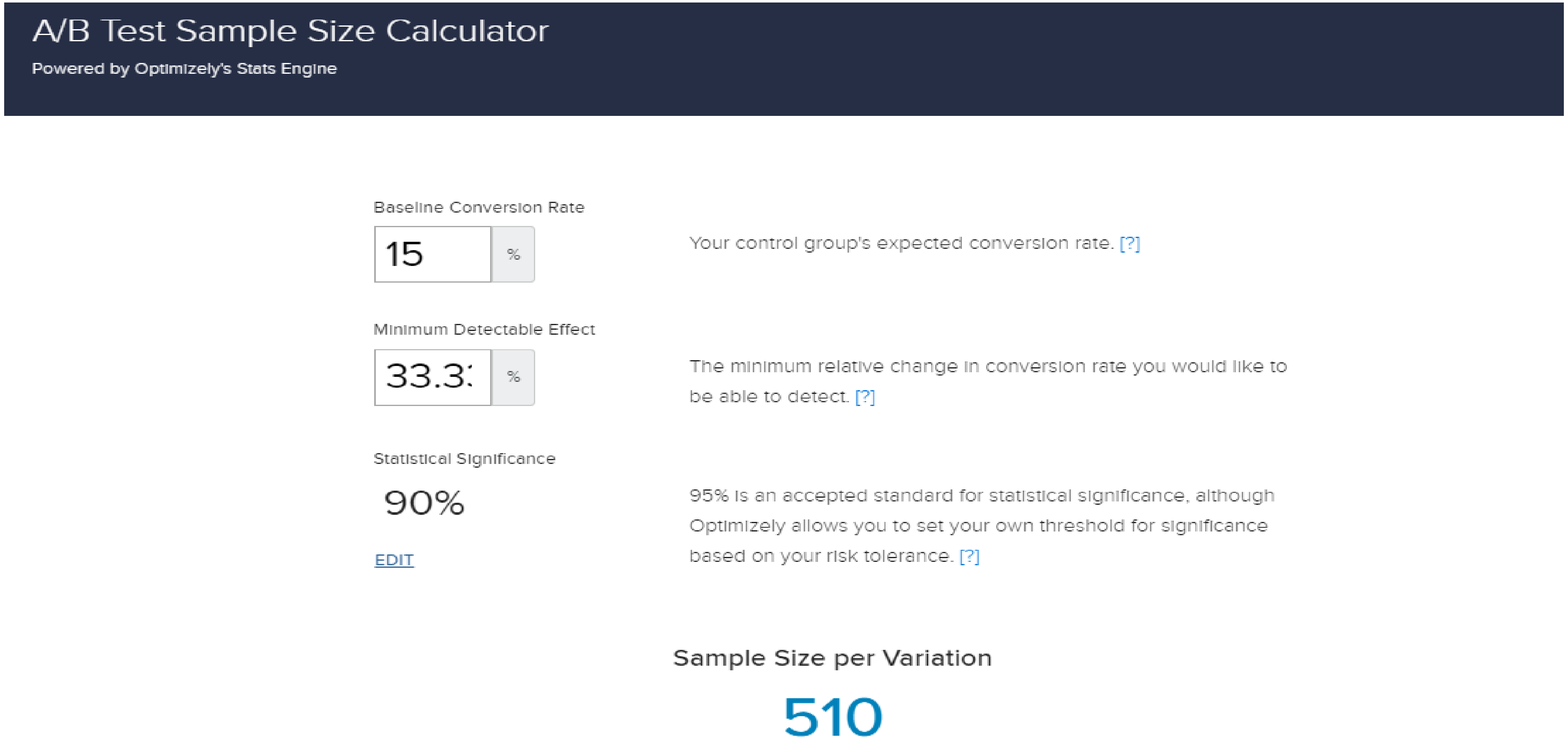
PART III: ASSISTING PROGRAM EVALUATION OF FOOT AND MOUTH DISEASE REDUCTION IN NATIONAL PARKS

www.alamy.com - AG27HR

PART III: ASSISTING PROGRAM EVALUATION OF FOOT AND MOUTH DISEASE REDUCTION IN NATIONAL PARKS – OVERVIEW

- Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park. The scientists want to test whether or not this program is working. Effectiveness has been defined as at least 5% reduction in disease.
- Is reduction effort working? What is the sample size of sheep observations – and time needed – to investigate efficacy of foot and mouth treatment?
- For sample size determination, need:
  - *Baseline data.* Do not have Yellowstone NP data, but can approximate using known value of 15% incidence at Bryce NP
  - *Minimum detectable effect*. How much reduction is needed? Use 33.3% (calculated from desired 5% reduction from 15%)
  - *Significance level*. Use default 90%
  - *Sample size calculator.* Use optimizely online sample size calculator (see next slide) https://www.optimizely.com/sample-size-calculator/?conversion=15&effect=33.333&significance=90

PART III: ASSISTING PROGRAM EVALUATION OF FOOT AND MOUTH DISEASE REDUCTION IN NATIONAL PARKS – SAMPLE SIZE CALCULATION



A/B Test Sample Size Calculator

Powered by Optimizely's Stats Engine

Baseline Conversion Rate

15 %

Your control group's expected conversion rate. [?]

Minimum Detectable Effect

33.3: %

The minimum relative change in conversion rate you would like to be able to detect. [?]

Statistical Significance

90%

EDIT

95% is an accepted standard for statistical significance, although Optimizely allows you to set your own threshold for significance based on your risk tolerance. [?]

Sample Size per Variation

510

PART III: ASSISTING PROGRAM EVALUATION OF FOOT AND MOUTH DISEASE REDUCTION IN NATIONAL PARKS – WEEKS NEEDED TO OBSERVE

- Scientists need to observe a sample of 510 sheep in each park in order to detect if the foot and mouth reduction program is working.
- Based on the previously observed number of sheep per week per park, we can calculate how long we will need to observe sheep at each relevant park in order to detect if the program is working.
- For Yellowstone NP, with 507 observed sheep per week, it will take a little over 1 week (1.006 weeks) to obtain the needed sample size (510 sample size/ 507 observations per week).
- For Bryce NP, with 250 observed sheep per week, it will take a little over 2 weeks (2.04 weeks) to obtain the needed sample size (510 sample size/ 250 observations per week).