

# Biodiversity for the National Parks

Alex Liu

# In “species\_info.csv”, I saw

- There are 4 columns in the table:
  - 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
5	Mammal	Odocoileus virginianus	White-Tailed Deer	nan

# In “species\_info.csv”, I saw(cont'd)

- There are 5541 species in the table
- There are 7 types of species(including animals and plant)  
Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant, Nonvascular Plant
- There are 5 types of conservation status:
  - Endangered
  - In recovery
  - Species of concern
  - Threatened
  - NA(not endangered at all)

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

# Which species are more endangered?

- From the data, we know how many percentages of species are endangered.
- Top 3 endangered species are “Mammal”(0.170), “Bird”(0.153), “Amphibian”(0.088)

Categories	Percentages
Amphibian	0.088608
Bird	0.153689
Fish	0.087302
Mammal	0.170455
Nonvascular Plant	0.015015
Reptile	0.064103
Vascular Plant	0.010793

# Which species are more endangered? (cont'd)

- Though we know the top 3 endangered species, can we really claim they rank like this way?
- Null hypothesis:  
The percentages of endangered possibility between the species are only by chance.
- Since the data is categorical, we use Chi-Square test
- We found P-value of “**mammal vs. bird**” is 0.688. 0.688 is way bigger than 0.05 so we can't reject the null hypothesis.  
Namely, the difference between mammal and bird could be by chance.

# Which species are more endangered? (cont'd)

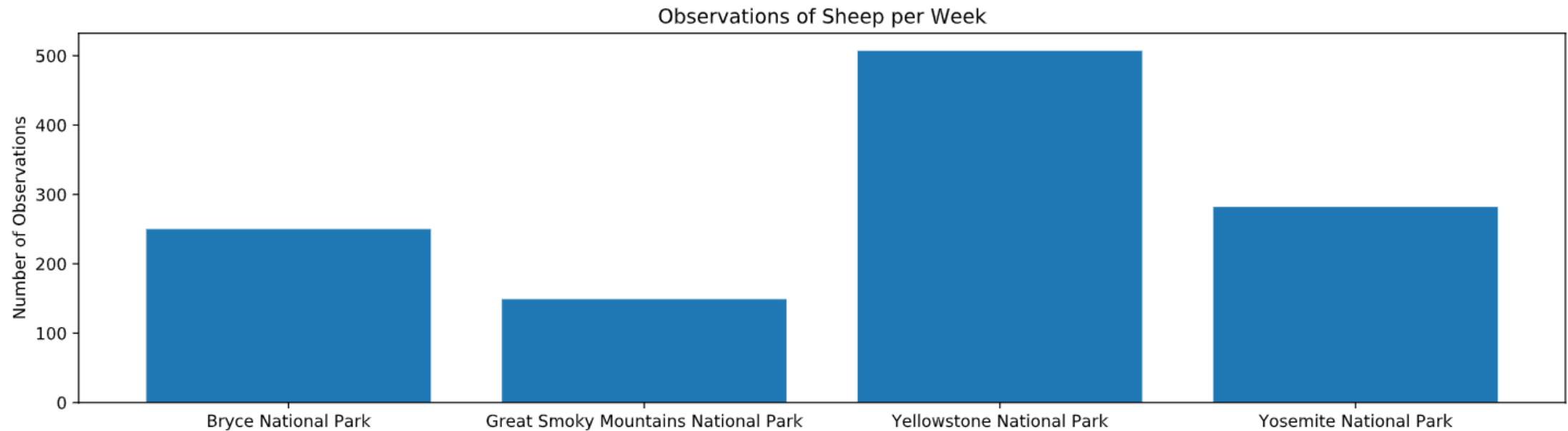
- We found P-value of “**mammal vs. amphibian**” is 0.126. 0.126 is still bigger than 0.05 so we can't reject the null hypothesis.  
Again, the difference between mammal and amphibian could be by chance.
- We found P-value of “**mammal vs. reptile**” is 0.038. 0.038 is smaller than 0.05 so we can reject the null hypothesis and claim: “the endangered possibility between mammal and reptile is significant”  
Namely, it's not by chance.

# Recommendation to the conservationists

- We can know the mammal species are more severely endangered than reptile species.
- We can't know the mammal species are more severely endangered than bird species and amphibian species.
- We recommend to protect mammal, bird and amphibian species at the same time first.

# sample size determination of sheep's disease

- Since we want to investigate the condition of the sheep, we make a filter to pick up the name including sheep and also the category belonging to mammal.
- Merging the species and observations dataframe, we can know the sheep being observed for how many times in certain park in a week:





# sample size determination of sheep's disease

- Since park ranger has been running the program to reduce the rate of foot and mouth disease at that park, scientists want to know the outcome but they need the figure to support the outcome.
- We know the ratio in last year is 15%
- We want to have 5% drop
- We use statistical significance at 90%
- The sample size should be more than 870 sheep
- How many week should scientists take to get enough samples:

Bryce National Park	Great Smoke Mountain National Park	Yellow Stone National Park	Yosemite National Park
4weeks	6weeks	2weeks	4weeks

The End