Capstone Project
Biodiversity

by **Runtao Wang**
bingowrt@gmail.com
(+1)562-473-1575

Biodiversity Project

1. Brief description of the data
2. Significance calculation for endangered status between different categories of species
3. Sample size determination for the foot and mouth disease study

1. Brief description of the Data

Data to be analyzed

- species_info.csv (5824 rows)
- observations.csv (23296 rows)

Data to be analyzed

- **species_info.csv**
    - Category
    - Scientific_name
    - Common_names
    - Conservation_status
- observations.csv

Data to be analyzed

- **species_info.csv**
    - Category

        There are 7 categories: 'Mammal', 'Bird', 'Reptile', 'Amphibian', 'Fish', 'Vascular Plant', 'Nonvascular Plant'

    - Scientific_name
    - Common_names
    - Conservation_status
- observations.csv

Data to be analyzed

- **species_info.csv**
    - Category
    - Scientific_name

        There are 5541 scientific names in total

    - Common_names
    - Conservation_status
- observations.csv

Data to be analyzed

- **species_info.csv**
    - Category
    - Scientific_name
    - Common_names

      There are 5504 unique common_names in total
    - Conservation_status
- observations.csv

Data to be analyzed

- **species_info.csv**
    - Category
    - Scientific_name
    - Common_names
    - Conservation_status

    There are 5 status of conservation: 'Species of Concern', 'Endangered', 'Threatened', 'In Recovery' and Null

- observations.csv

Data to be analyzed

- species_info.csv
- **observations.csv**
  - scientific_name

  There are 5541 scientific names in total

  - park_name
  - observations

Data to be analyzed

- species_info.csv
- **observations.csv**
    - scientific_name
    - park_name

    There 4 national parks: 'Great Smoky Mountains National Park', 'Yosemite National Park', 'Bryce National Park', 'Yellowstone National Park'

    - observations

Data to be analyzed

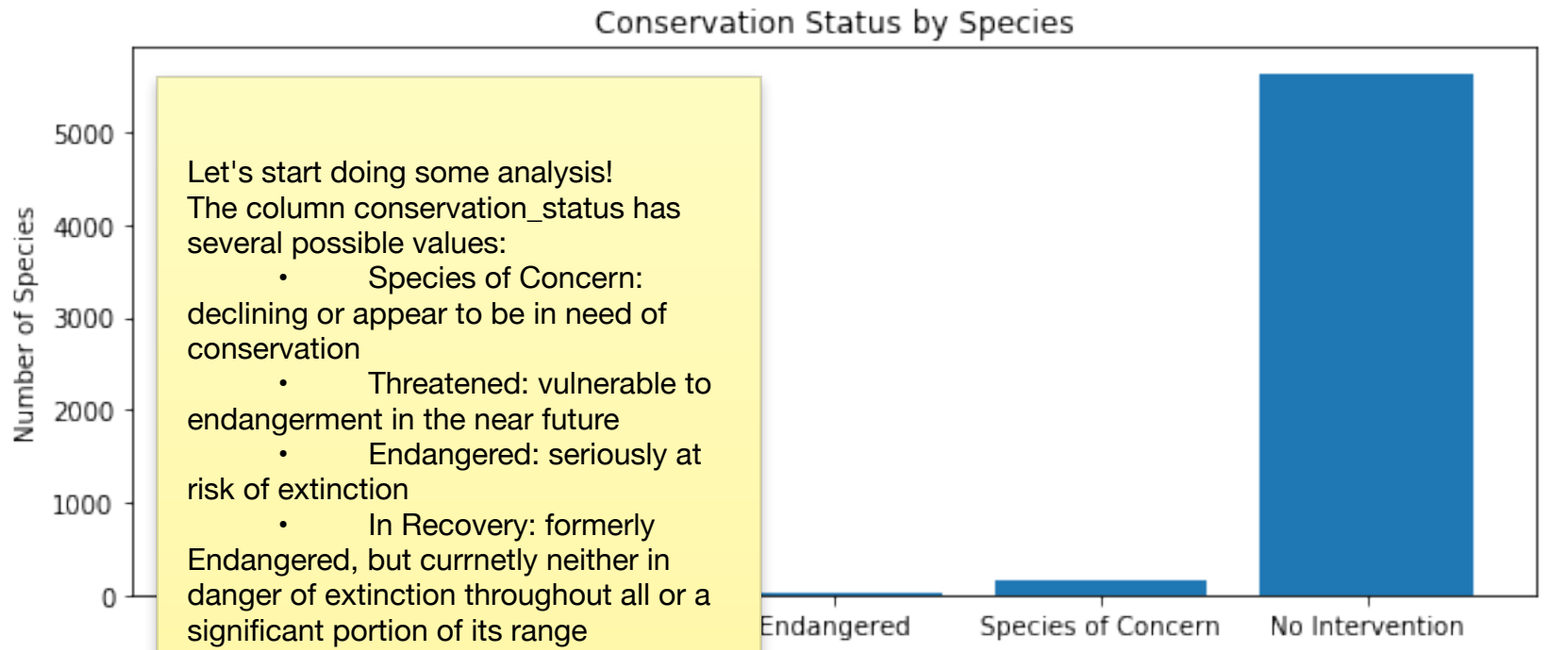
- species_info.csv
- **observations.csv**
  - scientific_name
  - park_name
  - observations

The number of animals observed

2. Significance calculation

Protection counts on different conservation status

conservation_status	scientific_name
In Recovery	4
Threatened	10
Endangered	16
Species of Concern	161
No Intervention	5633



Let's start doing some analysis!
The column conservation_status has several possible values:

- Species of Concern: declining or appear to be in need of conservation
- Threatened: vulnerable to endangerment in the near future
- Endangered: seriously at risk of extinction
- In Recovery: formerly Endangered, but currnetly neither in danger of extinction throughout all or a significant portion of its range

We'd like to count up how many species meet each of these criteria. Use groupby to count how many scientific_name meet each of these criteria.

Conservation Status by category

	category	is_protected	scientific_name
0	Amphibian	FALSE	72
1	Amphibian	TRUE	7
2	Bird	FALSE	413
3	Bird	TRUE	75
4	Fish	FALSE	115
5	Fish	TRUE	11
6	Mammal	FALSE	146
7	Mammal	TRUE	30
8	Nonvascular Plant	FALSE	328
9	Nonvascular Plant	TRUE	5
10	Reptile	FALSE	73
11	Reptile	TRUE	5
12	Vascular Plant	FALSE	4216
13	Vascular Plant	TRUE	46

Pivot

	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

Significance test of ‘Mammal’ and ‘Bird’

	protected	not protected	percent protected
Mammal	30	146	0.170455
Bird	75	413	0.153689

- Data is categorical
- There are 2 pieces of data to be compared

Chi Contingency Test
contingency = [[30, 146],
[75, 413]]

chi2,pval,dof,expected=(0.16170148
31654557,0.6875948096661336, 1,
[[27.8313253, 148.1686747],
[77.1686747, 410.8313253]]

Conclusion
1. Pval>5%
2. We can't reject the hypothesis Null
3. This difference isn't significant

Significance test of ‘Reptile’ and ‘Mammal’

	protected	not protected	percent protected
Mammal	30	146	0.170455
Reptile	5	73	0.064103

- Data is categorical
- There are 2 pieces of data to be compared

Chi Contingency Test
contingency = [[30, 146],
[5, 73]]

chi2,pval,dof,expected=(4.289183096203645, 0.038355590229698 98, 1, [[24.2519685, 151.7480315], [10.7480315, 67.2519685]]

Conclusion
1. Pval<5%
2. We can reject the hypothesis Null
3. This difference is significant

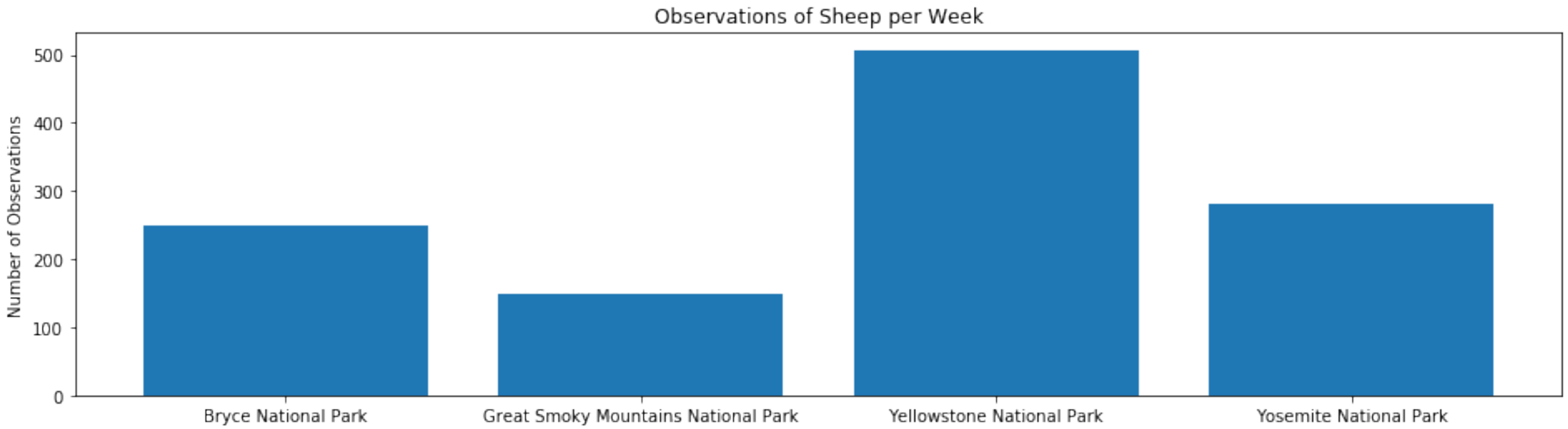
3. Sheep protection

Sheep observations in different national parks

	scientific name	park name	observations	category	common names	conservation status	is protected	is sheep
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
2	Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
3	Ovis canadensis	Great Smoky Mountains	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	TRUE	TRUE
4	Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	TRUE	TRUE
5	Ovis canadensis sierrae	Yosemite National Park	39	Mammal	Sierra Nevada Bighorn Sheep	Endangered	TRUE	TRUE
6	Ovis canadensis sierrae	Bryce National Park	22	Mammal	Sierra Nevada Bighorn Sheep	Endangered	TRUE	TRUE
7	Ovis canadensis sierrae	Great Smoky Mountains	25	Mammal	Sierra Nevada Bighorn Sheep	Endangered	TRUE	TRUE
8	Ovis aries	Yosemite National Park	126	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	FALSE	TRUE
9	Ovis aries	Great Smoky Mountains	76	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	FALSE	TRUE
10	Ovis aries	Bryce National Park	119	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	FALSE	TRUE
11	Ovis aries	Yellowstone National Park	221	Mammal	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	FALSE	TRUE

Observations of Sheep per Week in different national parks

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



Sample size determination for foot and mouth disease

15% of sheep at Bryce National Park have foot and mouth disease, there is a program to reduce the rate by at least 5%, and level of significance is 90%.

1. Minimum_detectable_effect=33.33%
2. Baseline=15%
3. Statistical_Significance=90%

Sample Size per Variation=510

bryce need 2 weeks while yellowstone need 1 week to observe enough sample

Tips for the conservationists

- According to the table of Conservation Status by category, ‘Mammal’ has the highest protected rate. However we can’t just say ‘Mammal’ is the most endangered species, because comparing with ‘Bird’, which has a tinny little bit lower protected ratio, there are high possibility that the result is due to a hypothesis Null.Here are the tips:

1. Pay attention to the species that have close protected rate, a significance calculation should be done to make sure there is a significant difference.

2. If there’s no significant difference, the assumption should be they are in the same conservation status

3. Calculate the right sample size before starting the program

4. ‘In recovery’ and ‘threatened' species should be the top priority