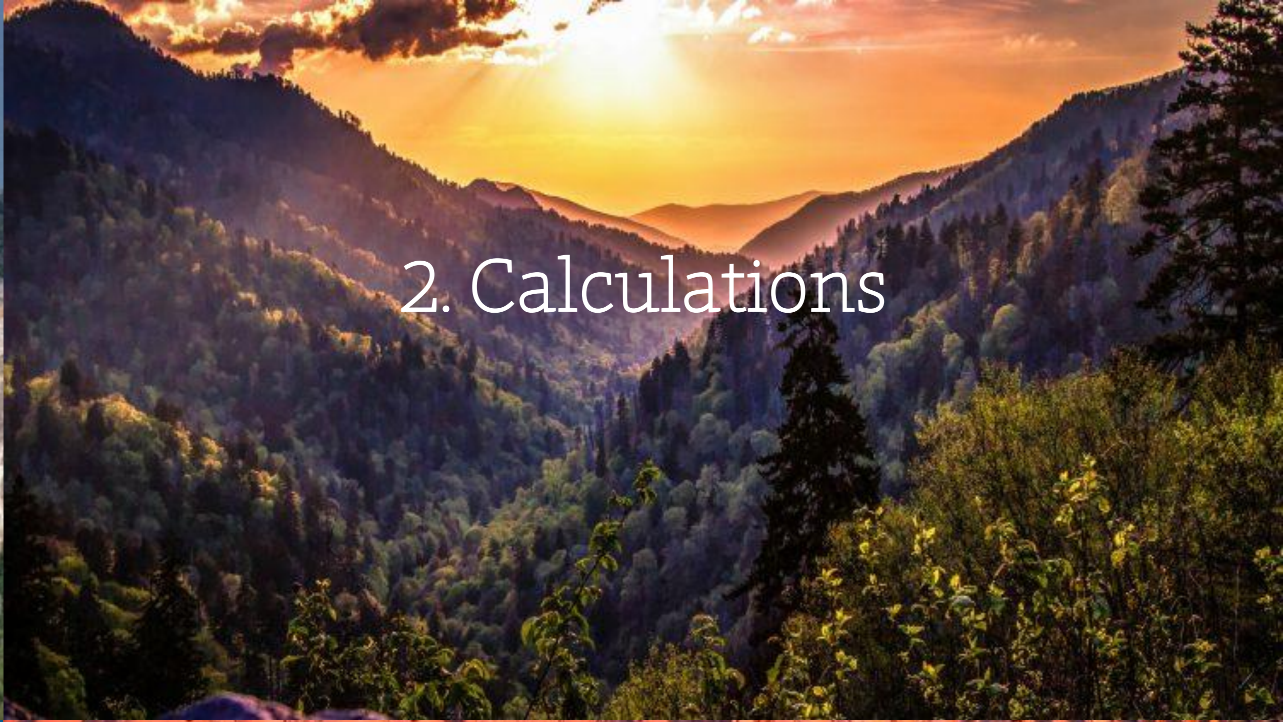
Biodiversity in the South Eastern National Parks

1. Data



2. Calculations

3. Recommendations

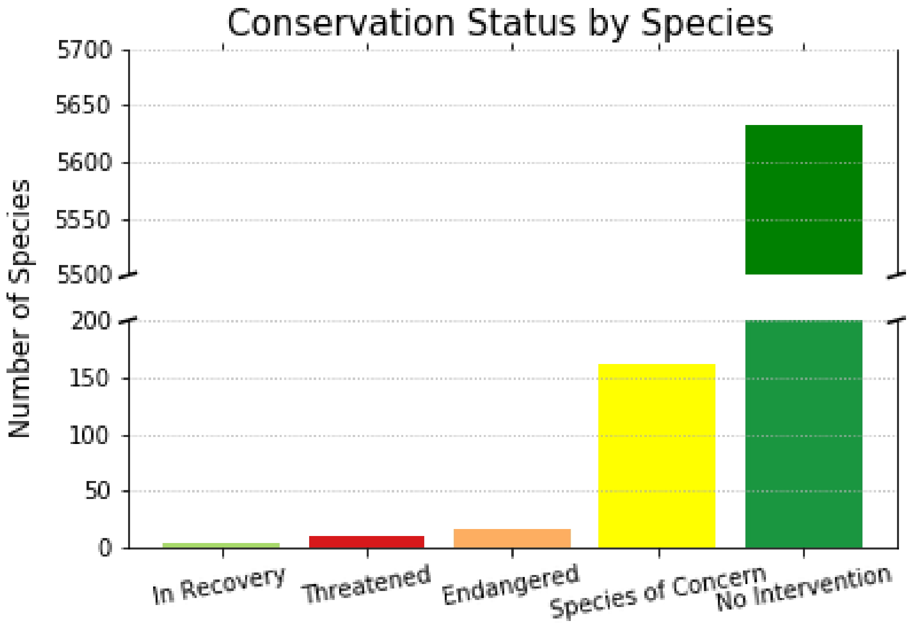
4. Sample size determination

Data

- The data used is inspired by real data, but is mostly fictional
- The dataset "species_info.csv" contains data of different species in the South East US National Parks
- species_info.csv contains data from 5,541 different species.
- The species are put into 7 categories: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant and Nonvascular Plant

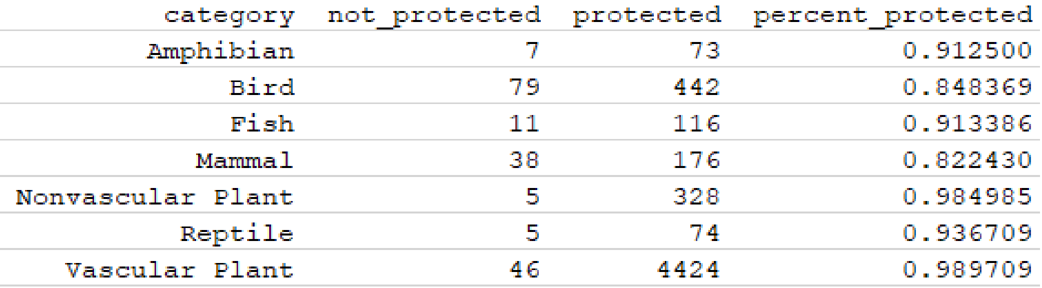
Calculations

- From all species, 191 require intervention for their preservation
- Which types of species are more likely to be endangered?



Calculations

- Mammals seems to have the highest likelihood of not being protected, but this difference is not significantly different than Birds (p = 0.45) and is significantly different to Reptiles (p = 0.02)



category	not_protected	protected	percent_protected
Amphibian	7	73	0.912500
Bird	79	442	0.848369
Fish	11	116	0.913386
Mammal	38	176	0.822430
Nonvascular Plant	5	328	0.984985
Reptile	5	74	0.936709
Vascular Plant	46	4424	0.989709

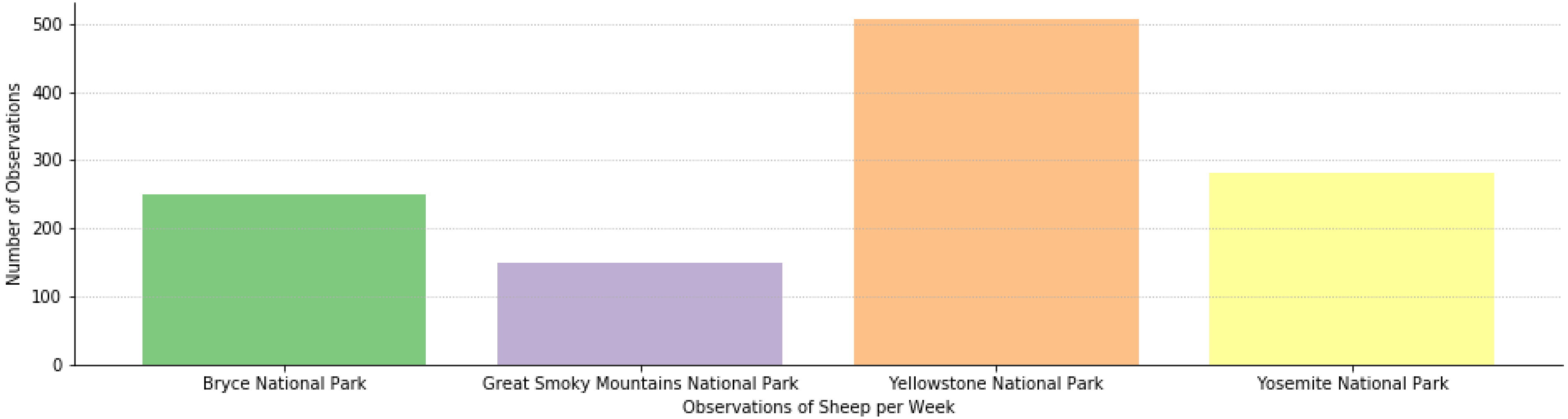
Data

- The dataset “observations.csv” contains recordings sightings of different species at several national parks for the past 7 days.
- observations.csv contains data from 23,296 observations.
- The observations distinguish the species’ scientific names, the number of observations and the park where its been observed.

Calculations

When looking at the number of sheep observations, we found that many Vascular Plants have "Sheep" in their scientific name.

- Within the 23,296 observations 1,118 were of Sheep (mammal). The division per park is as follows:



Recommendation

- To rate whether the rate of foot and mouth disease (fmd) is declining in the park due to the program I'd recommend a statistical siginificance Chi-squared test.
- 15% of sheep at Bryce National Park have fmd. To conclude a reduction of 5 percentage point with 90% confidence we'd recommend a minimum sample size of 870 which would require around 3.5 weeks with the current speed of observations.
- If you can conclude the program is successful, you can confidently expand the program to other parks

Recommendation

- While noting the observations, be aware of the possibility of double counts. This could greatly influence your conclusions and make them worth less. – Survivalship bias

Sample size determination

- The more observations, the smaller the confidence interval and the closer your observed mean tends to be to the real mean.
- However, observing the whole population is time consuming.
- Sample size determination allows us to predict the required sample size to give enough confidence that the conclusions are valid while also valuing the necessity of finding conclusions.

Sample size determination

- More observations are better if you have the time, but be sure to pre set the number of observations –or time you want to put into collecting the data.
- Changing the required number of observations or time spent while collecting the data could influence the validity of the data as it could (subconsciously) be influenced by outcome preferences – Confirmation bias

Sample size determination

- Baseline conversion rate: 15% - This is the ratio of positives you want to compare the sample size ratio with
  - Higher conversion rate requires a smaller sample as the detectable effect becomes absolutely larger
- Statistical significance: 90% - This is an estimate of the minimal confidence you wish to have to not have false-positives
  - Higher significance require a bigger sample to acquire the right level of confidence
- Minimum detectable effect: 33.3% - This is the percentage difference to the baseline that you'd wish to observe
  - Smaller detectable effects require a bigger sample to confidently conclude a difference