codecademy

Biodiversity for the National Parks

Introduction to Data Analysis

Jennifer Dennis

March 8, 2019

Table of Contents

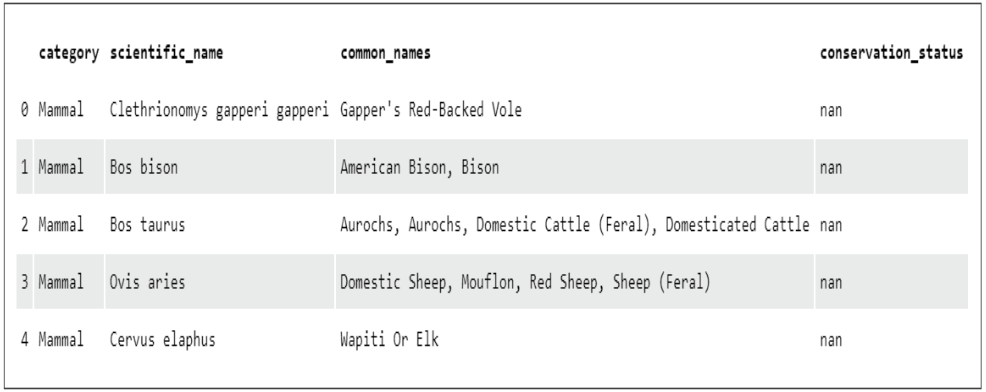
1. Describing data in species_info.csv
2. Significance calculations for endangered status species
3. Recommendation for conservationists
4. Sample size determination for foot and mouth disease

1. Describing data in species_info.csv

1.1 Describing data in species_info.csv

The data in species_info.csv includes:

- The scientific name of each species
- The common names of each species
- The species 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

Additionally, species_info.csv includes:

- 5541 different species
- 7 difference species types:
  - Mammal, bird, reptile, amphibian, fish, vascular plant, and nonvascular plant
- 5 different conservation statuses:
  - Nan (not a number), species of concern, endangered, threatened, and in recovery

2. Significance calculations for endangered species

2.1 Significance calculations for endangered species

Looking further into the conservation status aspect of species_info.csv, we can find out how many of each species falls into each conservation status. First, lets define the statuses:

- Species of Concern: declining population or appears 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 currently not in danger of extinction throughout all of a significant portion of its inhabitable range.

Using the code on the right, we are able to find how many of each species falls into each conservation status.

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

# Loading the Data
species = pd.read_csv('species_info.csv')

# print species.head()

species_count = species.scientific_name.nunique()

species_type = species.category.unique()

conservation_statuses =
species.conservation_status.unique()

conservation_counts =
species.groupby('conservation_status').scientific_nam
e.nunique().reset_index()

print conservation_counts
```

2.2 Significance calculations for endangered species (cont’d)

Using this new table, we can begin the analysis.

Unfortunatley, many species are in danger currently or will be soon. Only 4 species are in recovery, and I would urge better conservation efforts to increase the number of species in recovery while lowering the species that are endangered, a species of concern, or threatened.

	Conservation_status	Scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

# Loading the Data
species = pd.read_csv('species_info.csv')

# print species.head()

# Inspecting the DataFrame
species_count = species.scientific_name.nunique()

species_type = species.category.unique()

conservation_statuses =
species.conservation_status.unique()

# Analyze Species Conservation Status
conservation_counts =
species.groupby('conservation_status').scientific_name.nunique().reset_index()

print conservation_counts

species.fillna('No Intervention', inplace = True)

conservation_counts_fixed =
species.groupby('conservation_status').scientific_name.nunique().reset_index()

print(conservation_counts_fixed)
```

2.3 Significance calculations for endangered species (cont'd)

Fortunately, we can see though a tweak in coding that 5363 of the 5541 species in the DataFrame require no intervention. That's the vast majority at 96.8%.

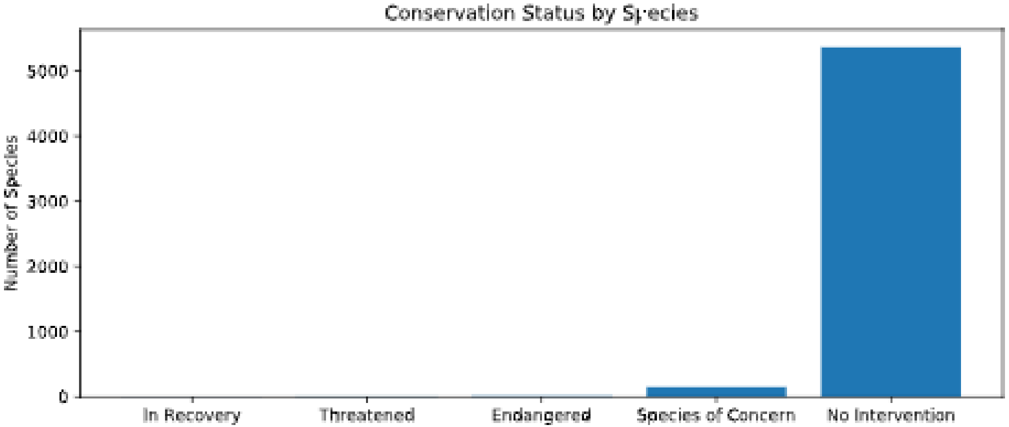
	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363*
3	Species of Concern	151
4	Threatened	10

*Please note the actual number is 5361, not 5363.

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

species = pd.read_csv('species_info.csv')

species_count = species.scientific_name.nunique()

species_type = species.category.unique()

conservation_statuses = species.conservation_status.unique()

conservation_counts =
species.groupby('conservation_status').scientific_name.nunique().reset_
index()

print conservation_counts

species.fillna('No Intervention', inplace = True)

conservation_counts_fixed =
species.groupby('conservation_status').scientific_name.nunique().reset_
index()

print(conservation_counts_fixed)
```

2.4 Significance calculations for endangered species (cont’d)

We can also use a plot to observe the data in a graph, which makes the data easier to understand.

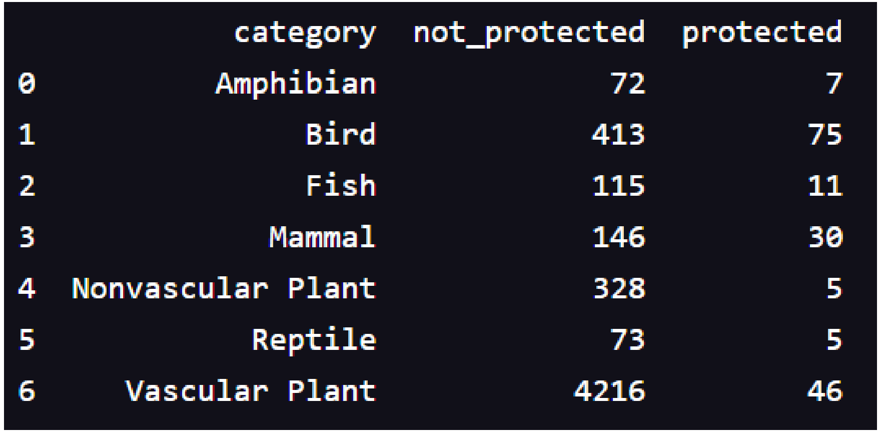


```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

species = pd.read_csv('species_info.csv')

species.fillna('No Intervention', inplace = True)

protection_counts =
species.groupby('conservation_status')\
    .scientific_name.nunique().reset_index()\
    .sort_values(by='scientific_name')

plt.figure(figsize=(10, 4))
ax = plt.subplot()
plt.bar(range(len(protection_counts)),protection_coun
ts.scientific_name.values)
ax.set_xticks(range(len(protection_counts)))
ax.set_xticklabels(protection_counts.conservation_sta
tus.values)
plt.ylabel('Number of Species')
plt.title('Conservation Status by Species')
labels = [e.get_text() for e in ax.get_xticklabels()]
plt.show()
```

2.5 Significance calculations for endangered species (cont’d)

By modifying the DataFrame (see code to the right), we are able to determine which types of species are more likely to be not protected and which are more likely to be endangered.

This results in the following table, which shows that bird species are the most in need or protection, followed by vascular plants, then mammals based on number of species, but what about percentage of total?



```
            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
```

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt

species = pd.read_csv('species_info.csv')

species.fillna('No Intervention', inplace = True)

species['is_protected'] = species.conservation_status !=
'No Intervention'

category_counts = species.groupby(['category',
'is_protected']).scientific_name.nunique().reset_index()

print category_counts.head()

category_pivot =
category_counts.pivot(columns='is_protected',
                      index='category',
                      values='scientific_name')\
                      .reset_index()

print category_pivot
```

2.6 Significance calculations for endangered species (cont'd)

To determine which category of species is more mathematically significance, we can do a significance test (Chi-Squared Test).

*

Significance is when the pval > 0.05

The result shows that mammal species are the most significant for endangered species at a .116 pval.

```
import codecademylib
import pandas as pd
from matplotlib import pyplot as plt
from scipy.stats import chi2_contingency

contingency = [[30, 146],
               [75, 413]]

pval = chi2_contingency(contingency)[1]
print(pval)
# No significant difference because pval > 0.05

contingency_reptile_mammal = [[30, 146],
                              [5, 73]]

pval_reptile_mammal =
chi2_contingency(contingency_reptile_mammal)[1]
print(pval_reptile_mammal)
# Significant difference! pval_reptile_mammal < 0.05
```

3. Recommendation for conservationists

3.1 Recommendation for conservationists

Based on the information discovered upon analysis of species_info.csv, I make the following recommendations to conservationists:

1. Determine what type of mammal is the highest endangered mammal, then look into possibilities of why
    2. For example, is it rodents and are they endangered because they have too many predators or is it a food shortage?
    3. For example, is it deer and are they endangered because there is illegal hunting going on in the parks or their food source is diseased?
4. Increase efforts to move more species from the species of concern, threatened, and endangered categories into the in recovery category.

4. Sample size determination for food and mouth disease

4.1 Sample size determination for foot and mouth disease

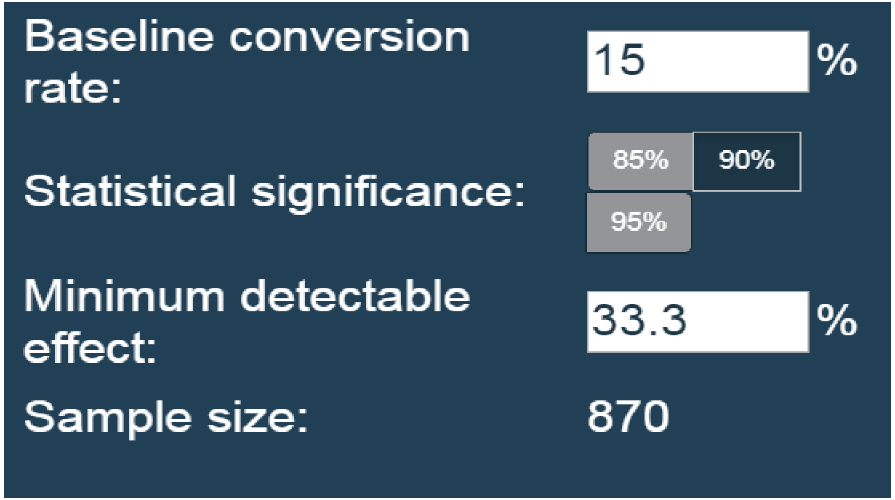
In order to determine a baseline percentage for a sample size, we can use the data already collected by the scientist last year.

- 15% of sheep at Bryce National Park have foot and mouth disease

We can combine data we know with data we want to know:

- Detect reductions of at least 5 percentage points
- Use the default level of significance at 90%

The sample size needed is 870. The time necessary to observe 870 sheep would vary depending on the park location being used.



Baseline conversion rate: 15 %

Statistical significance: 85% [✓ 90%] 95%

Minimum detectable effect: 33.3 %

Sample size: 870

4.1 Sample size determination for foot and mouth disease (cont’d)

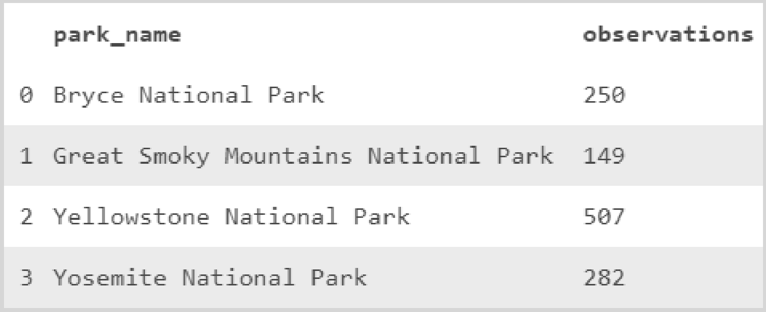
Based on the chart below, obtaining the sample size would take:

3.48 week in Bryce National Park

5.83 weeks in Great Smoky Mountains National Park

1.7 weeks in Yellowstone National Park

2.86 weeks in Yosemite National Park



	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