Estimating species abundance

Today you are going to use some of the new tools we have learned to explore how we might estimate the number of species in an ecosystem based on a random sampling of animals. As with the airplanes, you will know what the truth is which will enable you to assess how well an estimation works.

In [1]:
```
# Run this cell to set up the notebook, but please don't change it.

# These lines import the Numpy and Datascience modules.
import numpy as np
from datascience import *

# These lines do some fancy plotting magic.
import matplotlib
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
%matplotlib inline
import warnings
warnings.simplefilter('ignore', FutureWarning)
```

In [2]:
```
## First, a little for-loop practice.
## I want a for-loop that takes each of the items in an array of vegetables

veggie_array = make_array("broccoli", "spinach","potatoes")

for x in veggie_array:
    print("I like",x)
```

```
I like broccoli
I like spinach
I like potatoes
```

In [17]:
```
## Now, write a function using a for-loop that takes a string as an input ar
## appears in the string.

## Hint: To determine the number of times a letter appears in a string, you
## in its original form to its length with the given letter removed.

def vowel_counter(string):
    # string of vowels
    vowels = "aeiou"
    # casefold has been used to ignore cases
    string = string.casefold()

    # form a dictionary with key as a vowel and the value as 0
    count = {}.fromkeys(vowels, 0)

    # count the vowels
    for character in string:
        if character in count:
            count[character] += 1
    return count

vowel_counter("apple")
```

Out[17]:
```
{'a': 1, 'e': 1, 'i': 0, 'o': 0, 'u': 0}
```

Estimating species in ecosystem 1

The function below enables you to take a random sample of mammals living in an ecosystem in Australia. The function takes as an input the sample size

In [18]:
```
def mammal_sample(n):
    '''Returns a random sample of mammals from a section of the Australian c
    ecosystem = Table.read_table("/srv/data/DS_113_S23/Labs/Estimation_Pract
    return np.random.choice(ecosystem,n)

mammal_sample(1)
```

Out[18]:
```
array(['Bilby'],
      dtype='<U22')
```

In [19]:
```
## Use the function to take a sample of 30 animals.  Give it a name for furt
sample = mammal_sample(30)
sample
```

Out[19]:
```
array(['Koala', 'Bandicoot', 'Brush Tailed Possum', 'Wallaby', 'Kanagaroo',
       'Wallaby', 'Wallaroo', 'Wombat ', 'Wombat ', 'Koala', 'Wallaby',
       'Wallaby', 'Brush Tailed Possum', 'Spotted-Tail Quoll', 'Koala',
       'Wombat ', 'Wombat ', 'Grey Headed Flying Fox', 'Wombat ',
       'Wombat ', 'Wallaroo', 'Brush Tailed Possum', 'Rufous Bettong',
       'Bandicoot', 'Dingo', 'Spotted-Tail Quoll', 'Rufous Bettong',
       'Wallaroo', 'Dingo', 'Wallaroo'],
      dtype='<U22')
```

In [20]:
```
## Write a line of code that will determine for you the number of different
Table().with_column("species",sample).group("species").num_rows
```

Out[20]:
```
11
```

Suppose you wanted to estimate the total number of different mammal species in the ecosystem. Suppose we took the number of species in our sample as an estimate. How good an estimate is this? That is what we will explore below.

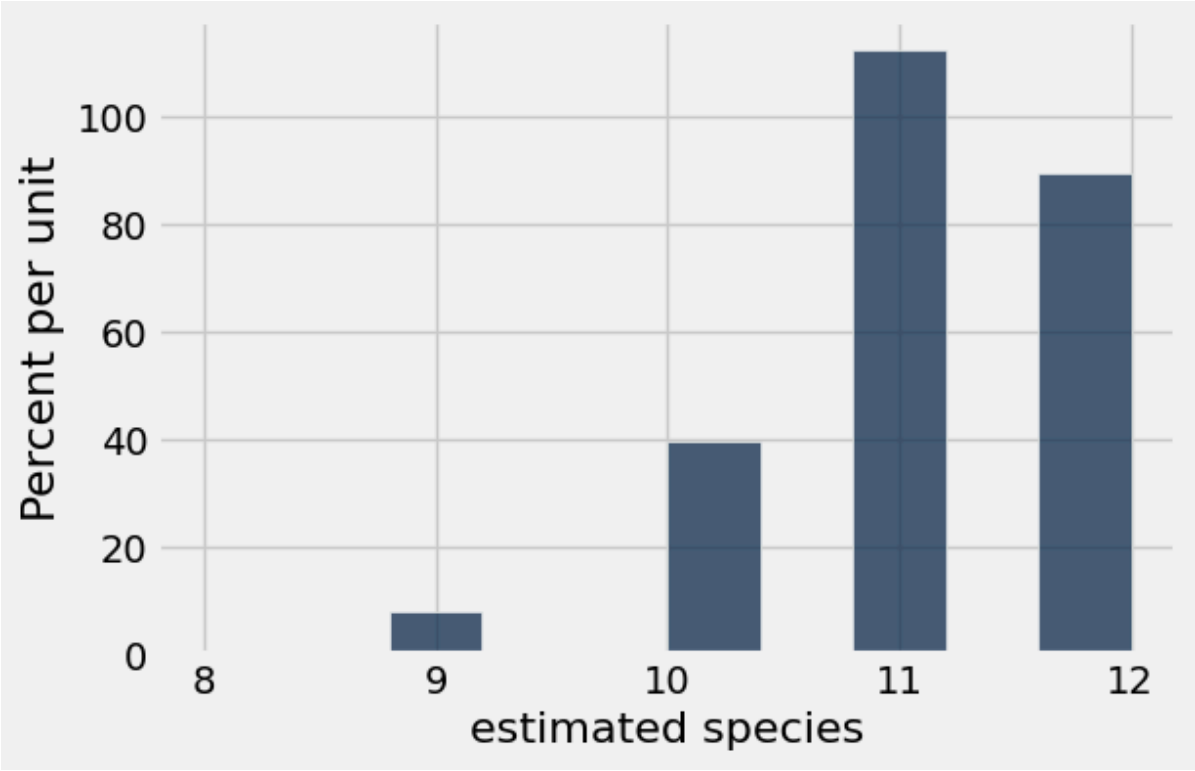
In [21]:
```
## Use a for-loop to take 1000 different samples of size 30.
## Consider each an estimate of the total number of mammal species.
## You should produce an array of estimates that I would like you to report

estimateArray = make_array()

for i in np.arange(1000):
    sampled_mammal = mammal_sample(30)
    estimateArray = np.append(estimateArray, Table().with_column("species",s

estimates = Table().with_column("estimated species", estimateArray)
estimates
```

Out[21]:

estimated species
11
10
10
10
11
12
11
12
11
12

... (990 rows omitted)

In [22]:
```
## Create an appropriate histogram of your estimates.
estimates.hist(bins=10)
```



In [23]:
```
## Run this code to look at the actual species in the ecosystem.
species = Table.read_table("/srv/data/DS_113_S23/Labs/Estimation_Practice_La
Table().with_columns(
    "Number", np.arange(1,len(species)+1),
    "Species", species
).show()
```

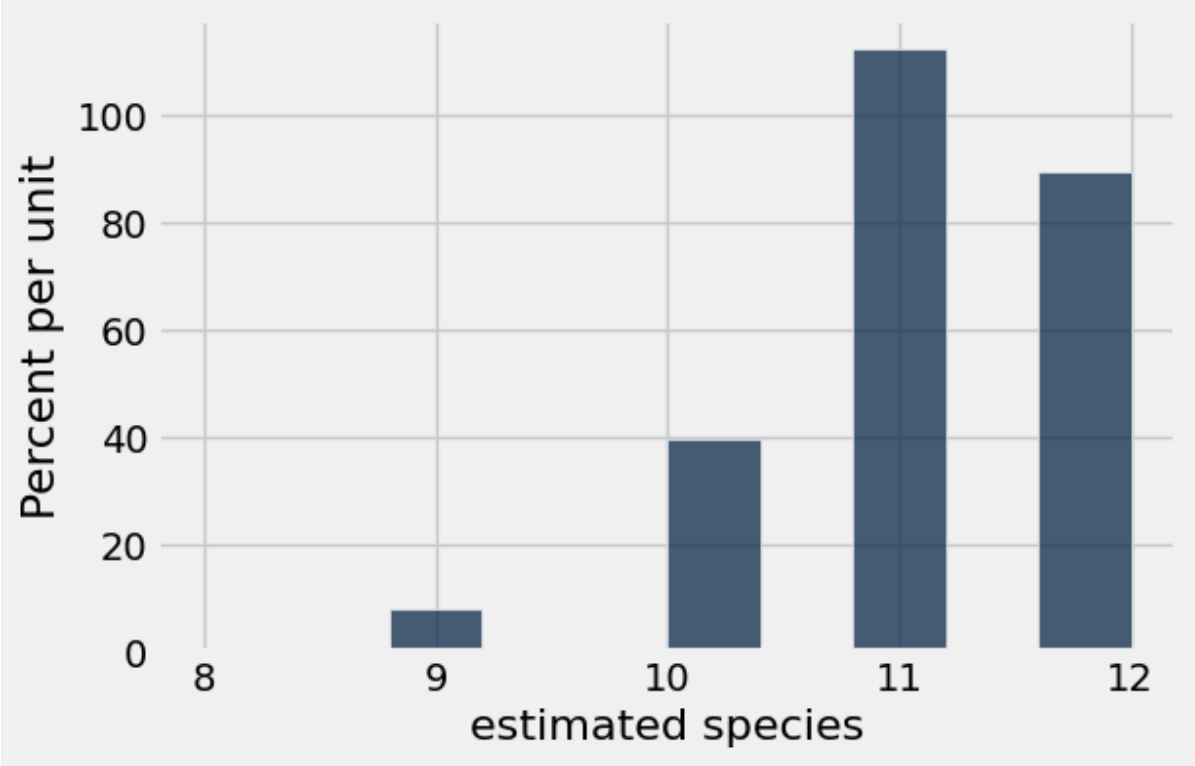
Number	Species
1	Dingo
2	Kanagaroo
3	Wallaby
4	Bandicoot
5	Bilby
6	Brush Tailed Possum
7	Grey Headed Flying Fox
8	Koala
9	Rufous Bettong
10	Spotted-Tail Quoll
11	Wombat
12	Wallaroo

Comment on the accuracy of this estimation method as well as any other characteristics that help you determine how useful it is.

Since the maximum number we have found from the estimate is 12, we can know that there will be 12 species in the ecosystem.

Repeat your work above, but now with a sample size of 100. Place your new code below and then comment on how this changed the accuracy of your method.

In [24]:
```
estimateArray100 = make_array()

for i in np.arange(1000):
    sampled_mammal_100 = mammal_sample(100)
    estimateArray100 = np.append(estimateArray100,
                                 Table().with_column("species",sampled_mamma

estimates100 = Table().with_column("estimated species", estimateArray)

# histogram of the table
estimates100.hist(bins=10)
```



There are no big changes in the estimation compared to the previous graph. However, since the sample size is larger than the previous data, we are more confident that the maximum number of species will be 12.

Estimating species in ecosystem 2

The function below enables you to take a random sample of reptiles and amphibians living in an ecosystem in Montana, USA. The function takes as an input the sample size. There is one significant difference for this ecosystem. In our Australian ecosystem, each species had the same chance of being found. This is not representative of reality. In this ecosystem, some species are more abundant than others.

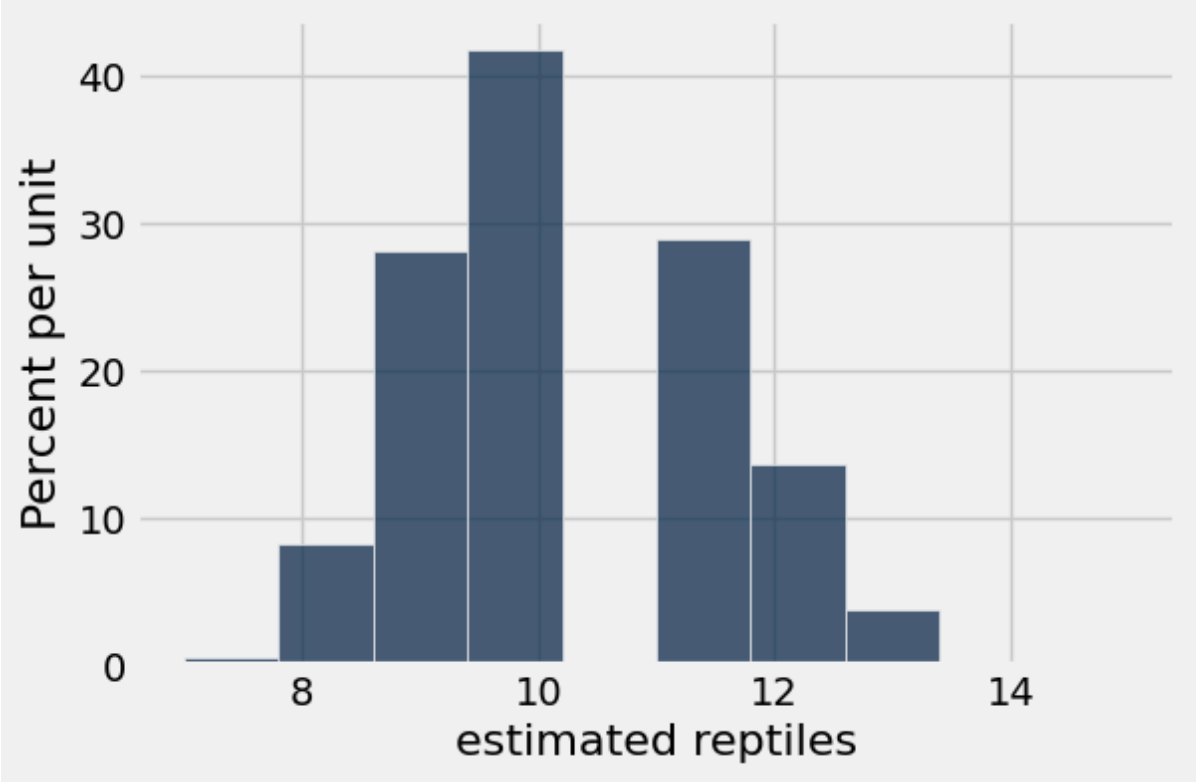
In [25]:
```
def herp_sample(n):
    '''Returns a random sample of reptiles and amphibians from Montana'''
    ecosystem = Table.read_table("/srv/data/DS_113_S23/Labs/Estimation_Pract
    x = make_array(200, 180, 150, 120, 100, 80, 60, 30, 10, 5, 3, 2, 2, 1, 1
    return np.random.choice(ecosystem,n, p = x/sum(x))

herp_sample(1)
```

Out[25]:
```
array(['Great Plains Toad'],
      dtype='<U21')
```

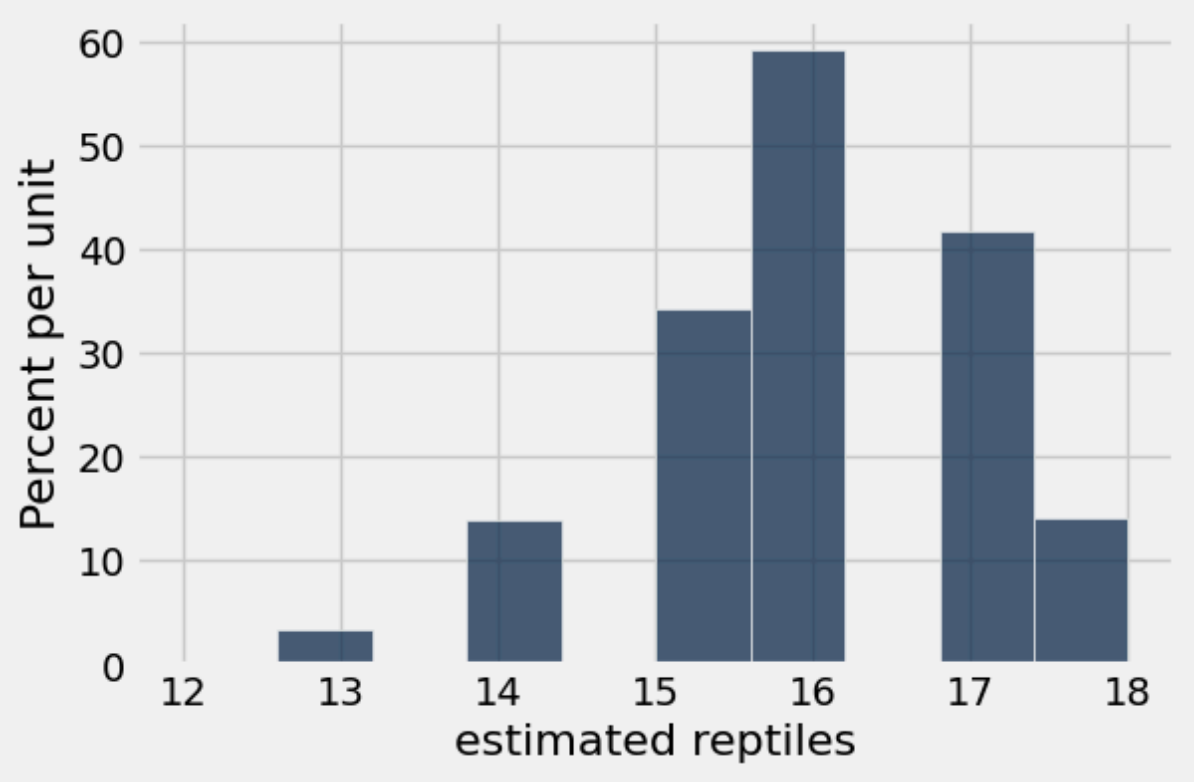
Please repeat the work you did for Ecosystem 1 except this time start with a sample size of 100 and increase to a sample size of 1000 for your second run through. When you are ready to observe the true number of species in this ecosystem, please run the code below.

In [26]:
```
estimateReptiles100 = make_array()

for i in np.arange(1000):
    sampled_Reptiles_100 = herp_sample(100)
```

```
    estimateReptiles100 = np.append(estimateReptiles100,
                                    Table().with_column("species",sampled_Repti

reptiles100 = Table().with_column("estimated reptiles", estimateReptiles100)

# histogram of the table
reptiles100.hist(bins=10)
```



In [27]:
```
estimateReptiles1000 = make_array()

for i in np.arange(1000):
    sampled_Reptiles_1000 = herp_sample(1000)
    estimateReptiles1000 = np.append(estimateReptiles1000,
                                    Table().with_column("species",sampled_Repti

reptiles1000 = Table().with_column("estimated reptiles", estimateReptiles100

# histogram of the table
reptiles1000.hist(bins=10)
```



In [28]:
```
## Run this code to look at the actual species in the ecosystem.
species = Table.read_table("/srv/data/DS_113_S23/Labs/Estimation_Practice_La
Table().with_columns(
    "Number", np.arange(1,len(species)+1),
    "Species", species
).show()
```

Number	Species
1	Tiger Salamander
2	Pacific Tree Frog
3	Boreal Chorus Frog
4	Western Toad
5	Woodhouse Toad
6	Great Plains Toad
7	American Bullfrog
8	Columbia Spotted Frog
9	Sagebrush Lizard
10	Gopher Snake
11	Eastern Racer
12	Western Skink
13	Milksnake
14	Snapping Turtle
15	Smooth Greensnake
16	Painted Turtle
17	Rubber Boa
18	Spiny Softshell

Please comment on the effectiveness of your method in this second ecosystem. How could you modify your method to increase your accuracy?

Compared to the estimation of the sample size of 100, the estimation with the sample size of 1000 showed more number of species. Since there are more species than the first ecosystem, the bigger sample size would help us to figure out the number of species more accurately.

Estimating some probabilities

Using your method with 1000 samples, please estimate the following:

1. The probability of your method producing the right answer.
2. The probability of your method being wrong by more than 3 species.

In [29]:
```
probability_right = (len(reptiles1000.where("estimated reptiles", are.equal_
probability_right
```

Out[29]:
```
0.085
```

In [30]:
```
probability_wrong_3 = (len(reptiles1000.where("estimated reptiles", are.belo
probability_wrong_3
```

Out[30]: 0.31

The probability to find the number of species in the ecosystem is 9.7%. The probability to find the number of species wrong by more than 3 species in the ecosystem is 30.5%.

If you have time, please ponder a more sophisticated approach to estimating the number of species. In particular, what might you look at in a sample to decide whether or not you are not likely to find any new species if you keep searching?