Анализ данных при помощи языка Python

Горин Н.А.

МГТУ им. Н.Э. Баумана

6 июля 2023 г.

Аннотация

Для анализа данных были выбраны два исследования по наблюдению за морфологией Дарвиновых вьюрков (лат. *Geospizinae*), проведённые в 1975 г. и 2012 г. Питером и Розмари Грант (*Peter and Rosemary Grant*). Выбранная работа является классическим научным трудом, позволяющим проследить эволюцию живых организмов (вьюрков вида *Geospiza fortis* и *Geospiza scandens*) на коротком временном отрезке.

Введение

В данной работе представлен статистический анализ данных, сделанный при помощи языка Python. При выполнении работы использовались библиотеки *matplotlib, numpy, pandas* и *seaborn*.

Основная часть

Импортируем библиотеки и подключаем выборки:

```
1 import seaborn as sns
2 import pandas as pd
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 sns.set_theme()
7
8 beaks_1975 = pd.read_csv('1975.csv')
9 beaks_2012 = pd.read_csv('2012.csv')
```

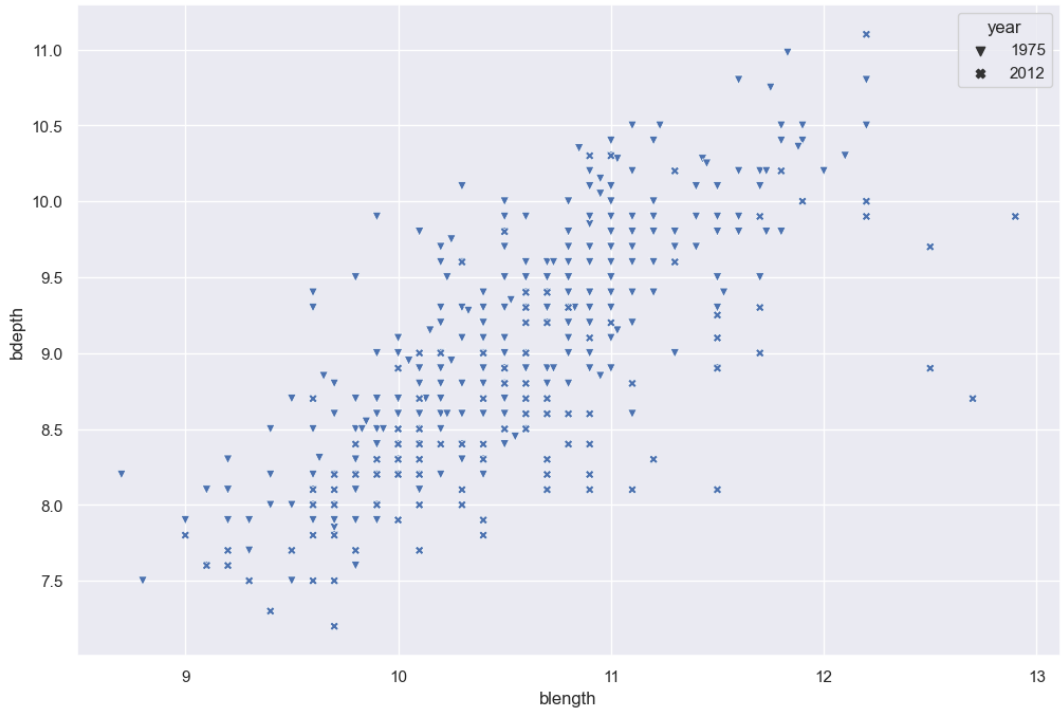
Редактируем файлы для более удобного анализа:

```
beaks_1975['year'] = "1975"
beaks_2012['year'] = "2012"
beaks_1975.rename(columns={'Beak depth, mm' : 'bdepth','Beak length, mm' :
    'blength'}, inplace=True)
data = pd.concat([beaks_1975,beaks_2012]).reset_index(drop=True)
fortis = data[data.species == 'fortis'].reset_index(drop=True)
scandens = data[data.species == 'scandens'].reset_index(drop=True)
```

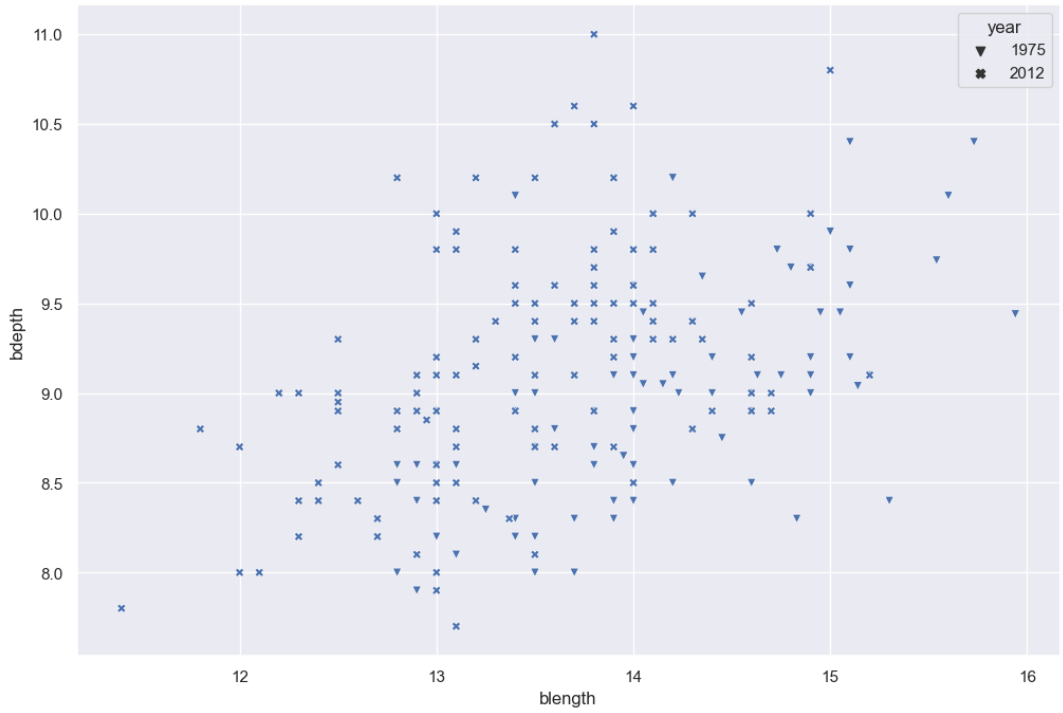
Маркируем данные и строим графики типа **ScatterPlot**:

```
markers = {'1975': "v", '2012': "X"}
fig = plt.figure(figsize=(12, 8))
sns.scatterplot(data=fortis, x='blength', y='bdepth', style='year', markers
    =markers)
plt.show()

markers = {'1975': "v", '2012': "X"}
fig = plt.figure(figsize=(12, 8))
sns.scatterplot(data=scandens, x='blength', y='bdepth', style='year',
    markers=markers)

plt.show()
```

Длина и ширина клюва (mm) вьюрков вида *fortis*



Длина и ширина клюва (mm) вьюрков вида *scandens*

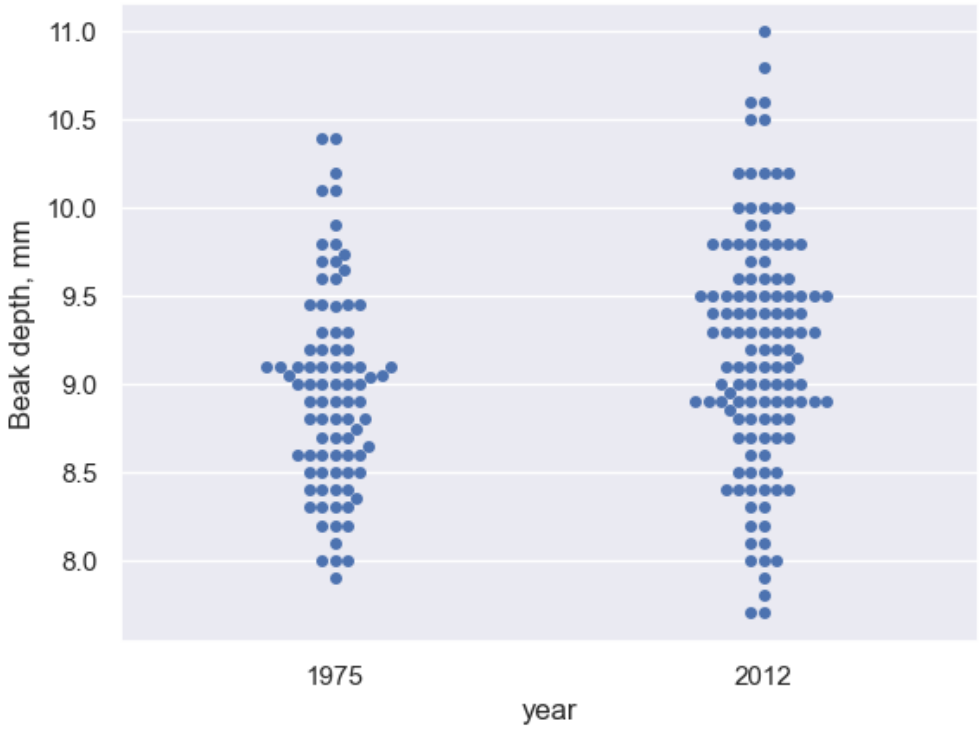


Из графиков видно заметное различие в параметрах клюва у вьюрков вида *scandens*. Дальнейшие данные будем систематизирвоать по особям этого вида.

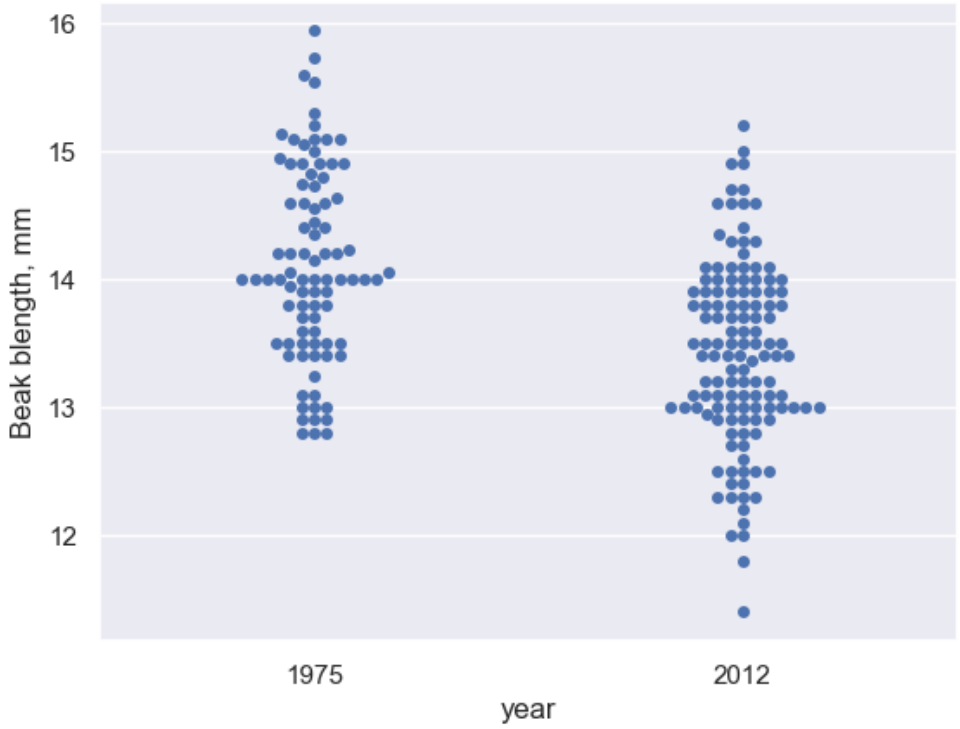
Разделим особей вида *scandens* по временным промежуткам проведённых исследований и построим графики типа **SwarmPlot**.

```
scandens_1975 = beaks_1975[beaks_1975['species'] == 'scandens']
candens_2012 = beaks_2012[beaks_2012['species'] == 'scandens']
scandens_bdepth_1975 = scandens_1975['bdepth'].reset_index(drop=True)
scandens_bdepth_2012 = scandens_2012['bdepth'].reset_index(drop=True)

sns.swarmplot(scandens, x='year', y='bdepth')
plt.xlabel('year')
plt.ylabel('Beak depth, mm')

scandens_blength_1975 = scandens_1975['blength'].reset_index(drop=True)
scandens_blength_2012 = scandens_2012['blength'].reset_index(drop=True)
sns.swarmplot(scandens, x='year', y='blength')
plt.xlabel('year')
plt.ylabel('Beak blength, mm')

plt.show()
```

Ширина клюва вьюрков вида *scandens* за 1975 и 2012 гг.



Длина клюва вьюрков вида *scandens* за 1975 и 2012 гг.

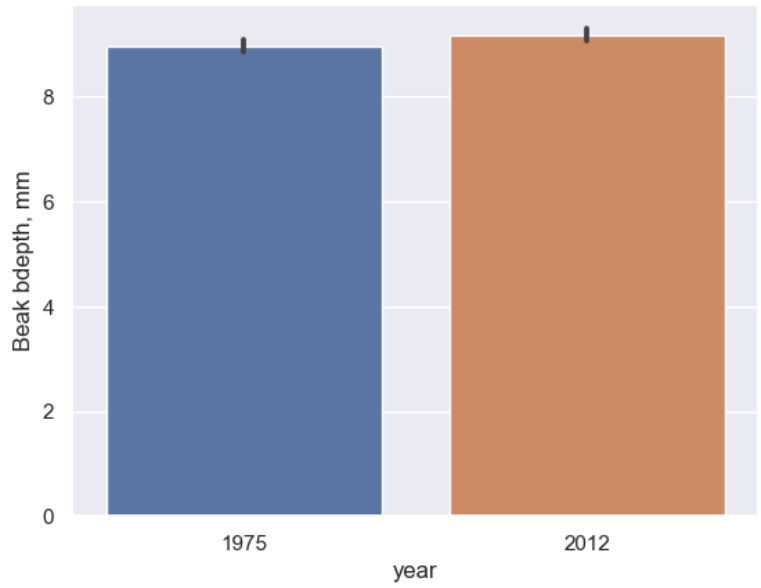


Построенные графики позволяют выявить явные статистические изменение в морфологии строения клюва вида *scandens* за 37 лет.

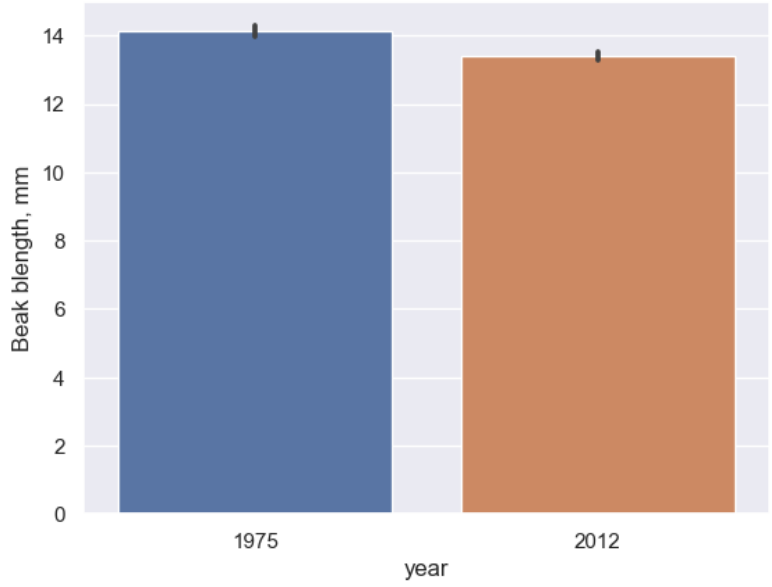
Выполним проверку результатов на графиках другого типа: **BarPlot**.

```
sns.barplot(scandens, x="year", y="bdepth")
plt.xlabel('year')
plt.ylabel('Beak bdepth, mm')

sns.barplot(scandens, x="year", y="blength")
plt.xlabel('year')
plt.ylabel('Beak blength, mm')

plt.show()
```

Ширина клюва вьюрков вида *scandens* за 1975 и 2012 гг.



Длина клюва вьюрков вида *scandens* за 1975 и 2012 гг.

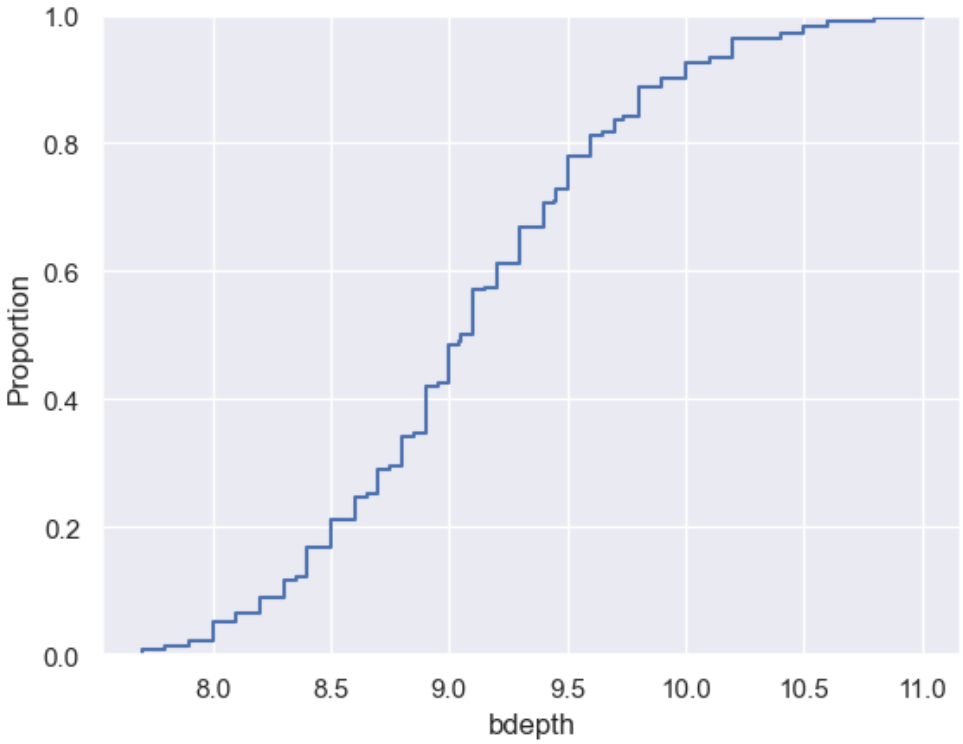


В завершение работы посмотрим, как менялась ширина и длина клюва вьюрков вида *scandens* в совокупности за 1975 и 2012 гг. Выполнить анализ помогут графики типа **ECDFPlot**.

Функция *ECDF* (эмпирическая кумулятивная функция распределения) представляет долю или количество наблюдений, попадающих ниже каждого уникального значения в наборе данных. Преимущество по сравнению с гистограммой или графиком заключается в том, что каждое наблюдение визуализируется напрямую, а это означает, что нет необходимости корректировать параметры группирования или сглаживания.

```
sns.ecdfplot(scandens, x="bdepth")
plt.xlabel('bdepth')
plt.ylabel('Proportion')

sns.ecdfplot(scandens, x="blength")
plt.xlabel('blength')
plt.ylabel('Proportion')

plt.show()
```

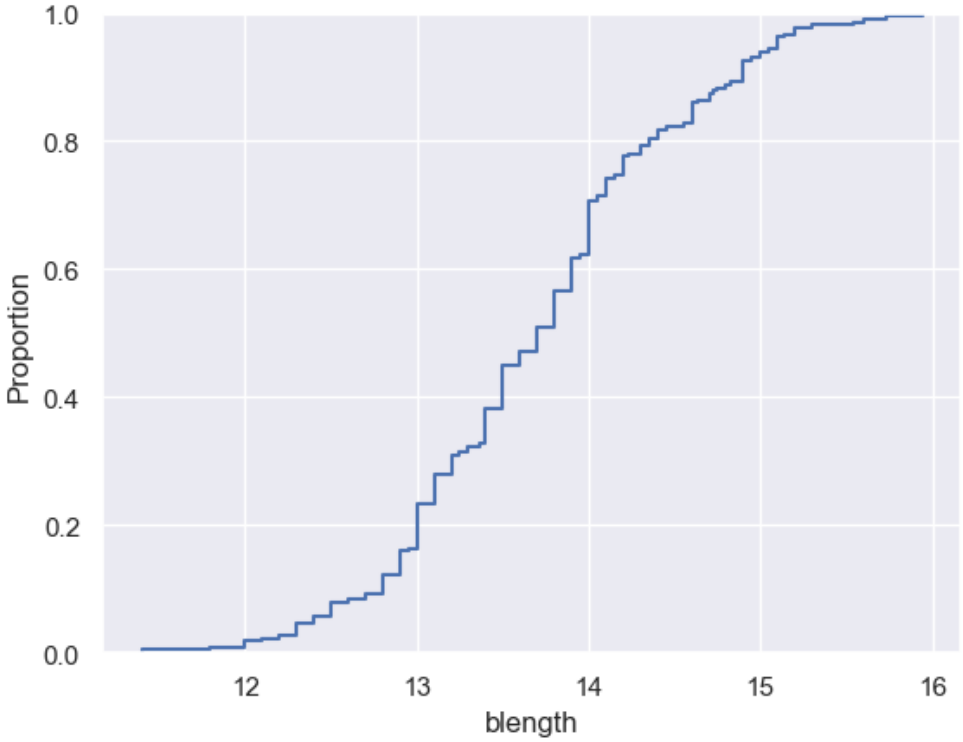
Ширина клюва вьюрков вида *scandens*



Заключение

Проведённый статистический анализ является прямым свидетельством морфологической изменчивости вида *scandens*, произошедшими за

Длина клюва вьюрков вида *scandens*



37 лет между исследованиями.

Результатом работы служит новое доказательство эволюционной теории Ч. Дарвина.