Отчёт по домашнему заданию номер 1 по курсу Applied statistics

Исполнительница: Смолкина Ю.А

Группа: Адбм 2021

Используемая среда: гугл коллаб (язык – питон)

Задание:

1. Выбрать одну из шести бактерий и один эукариотический организм
2. Проиллюстрировать различия между распределениями длин белков двух выбранных организмов (бактерии и эукариота) тремя способами:
   a. Совместной гистограммой
   b. Парой “ящиков с усами” (box plots)
   c. Графиками эмпирических функций распределения (два графика на одном рисунке)

1. Для сравнения были взяты данные длины белков двух организмов (длина измеряется в пептидах). Одиниз домена бактерий (Bacilllus subtulls, сокращённо называемая BACSU) и другой из домена эукариотов (Plasmodium berghei, сокращённо называемый PLABA).

Предобработка и статистические значения:

Запишем в разные переменные наши данные

```
BACSU = Bac['Length.1']
PLABA = Euc['Length.4']
```

```
[43] BACSU # выбранная бактерия
```

```
0       394.0
1       539.0
2       544.0
3       424.0
4       231.0
        ...
4386      NaN
4387      NaN
4388      NaN
4389      NaN
4390      NaN
Name: Length.1, Length: 4391, dtype: float64
```

```
[68] BACSU_df = pd.DataFrame(BACSU)
     BACSU_df = BACSU_df.dropna()
```

```
[ ] BACSU_df_mean = BACSU_df['Length.1'].mean #394.0
    BACSU_df_mean = 394.0
```

Для каждой (аналогичным способом) посмотри средние значение, а потом разницу между ними.

```
[78] #PLABA_df_mean - BACSU_df_mean
     print('difference of mean length is ',PLABA_df_mean - BACSU_df_mean)
```

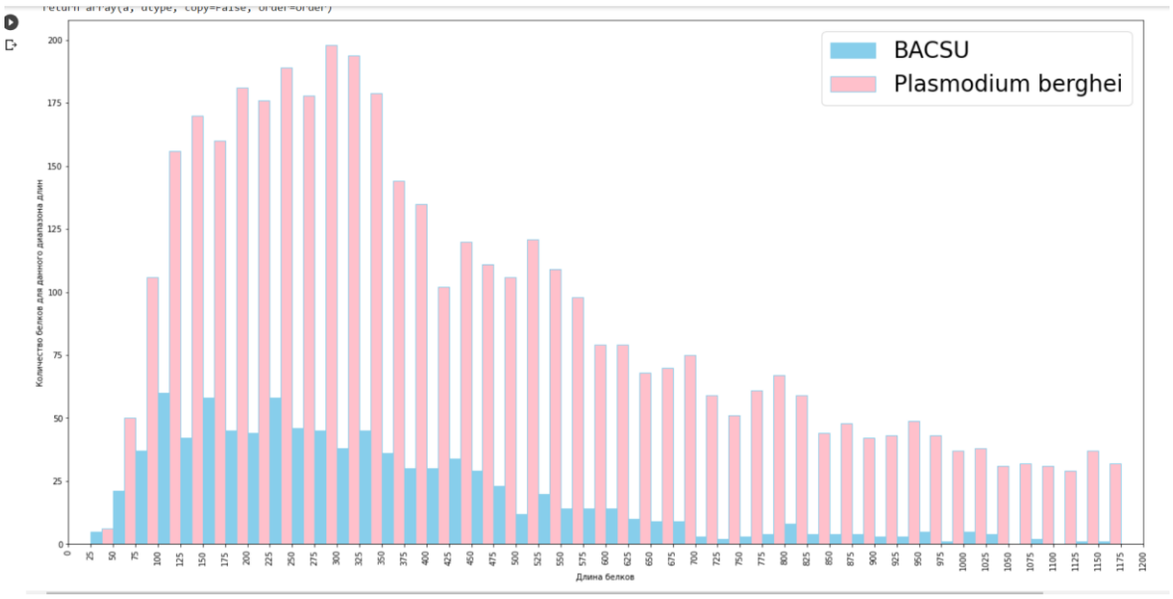
```
difference of mean length is  201.0
```

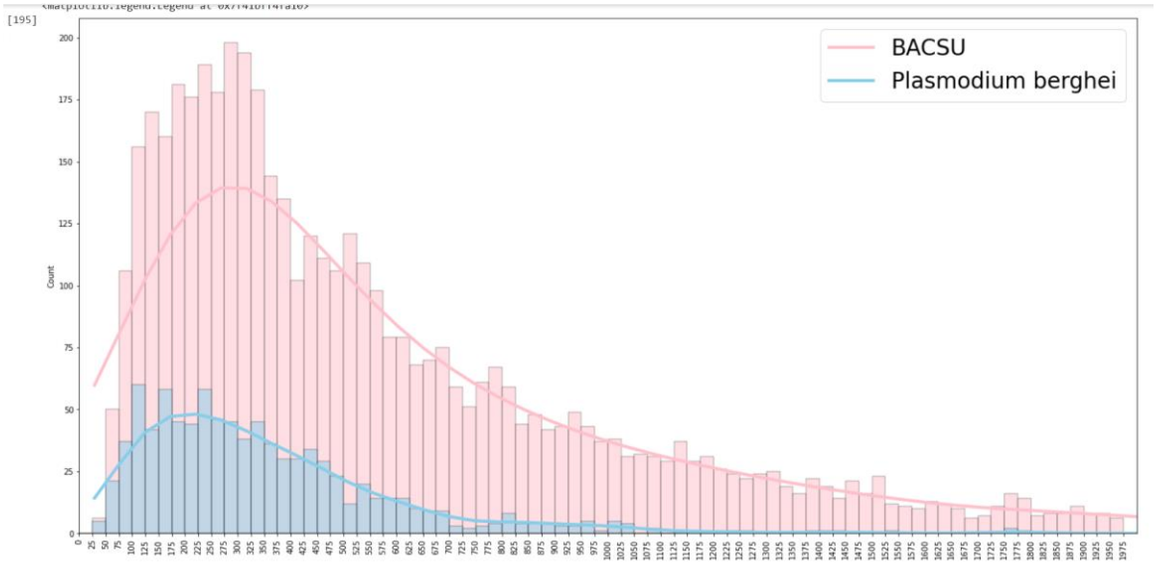
```
[88] count = 0
     for i in PLABA_df['Length.4']:
       if (i >= BACSU_df_mean):
         count = count+1
     print('число значений длины PLABA больше средней длины  BACSU = ',count)
     #PLABA_df['Length.4'].size # 4927
     print('процент кол-ва значений в PLABA , которые длиннее среднего в BACSU',count*100/PLABA_df['Length.4'].size,'%')
```

```
число значений длины PLABA больше средней длины  BACSU =  2738
процент кол-ва значений в PLABA , которые длиннее среднего в BACSU 55.571341587172725 %
```

Для обоих организмов была построена совместная гистограмма числа белков в зависимости от их длины:
С помощью команды

```
plt.hist([BACSU_df, PLABA_df], color = ["skyblue",'pink'], ec="skyblue",
         bins = [x*25 for x in range(0, int(1200/25))], rwidth = 2)
```

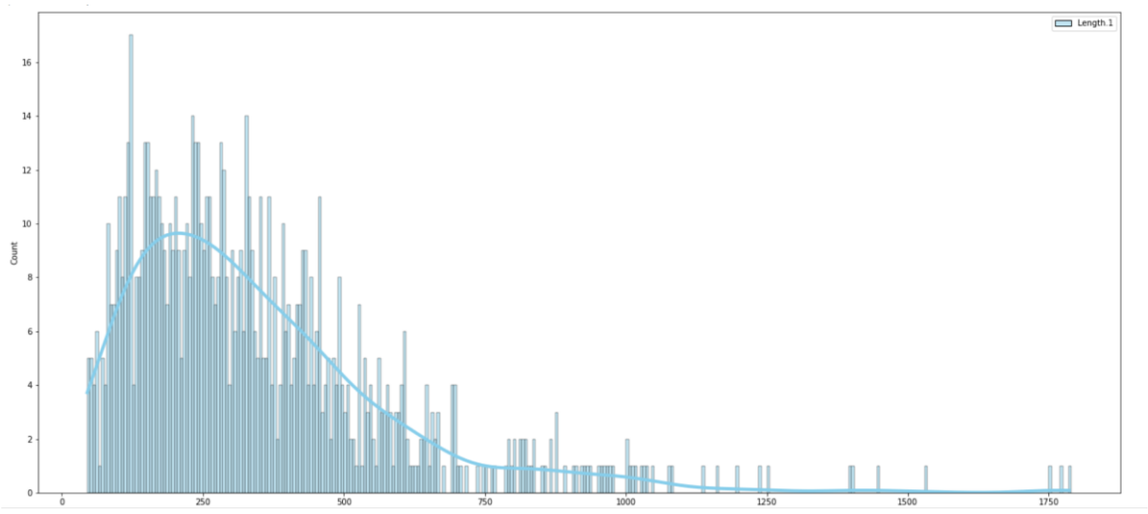




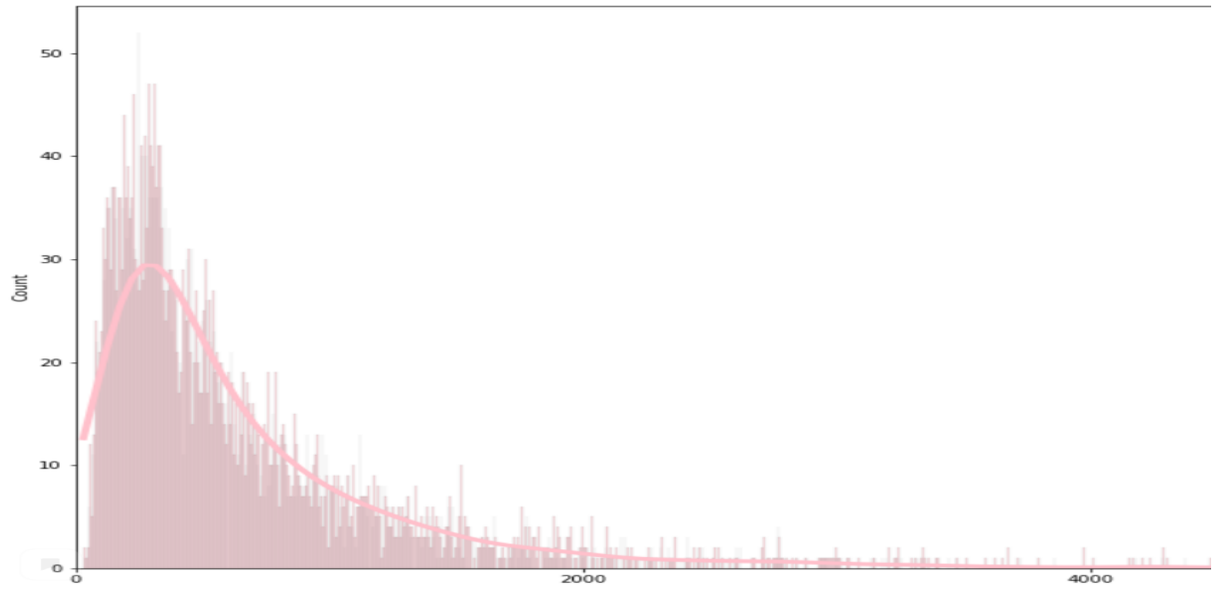
Распределение длин в Bacsu построение осуществлялось с помощью

```
g = sns.histplot(BACSU_df,kde=True,binwidth=5,

                 palette=["skyblue"],
                 ax = axs[0],
                 line_kws={'color': 'crimson', 'lw': 4},
                 bins = range(0,2001,25))#.set_xlim(0,2000)
```



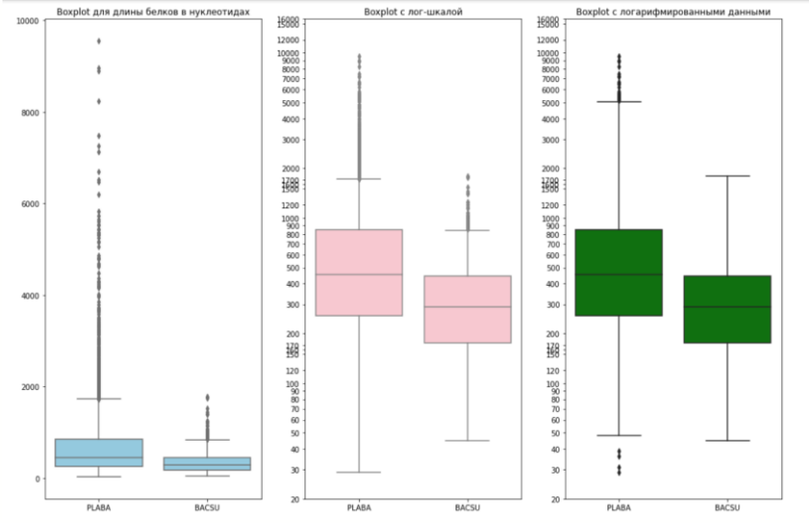
Plaba (аналогичные команды)



2. Теперь перейдем к построению ящиков с усами:

Основная команда :

```
sns.boxplot(data = data_for_boxplot,color = "skyblue", ax = axs[0]).set_title('Boxplot для длины белков в нуклеотидах')
```



В данном случаи

Усы Boxplot в данном случае соответствуют 1.5 интерквартального интервала (1.5 IQR).

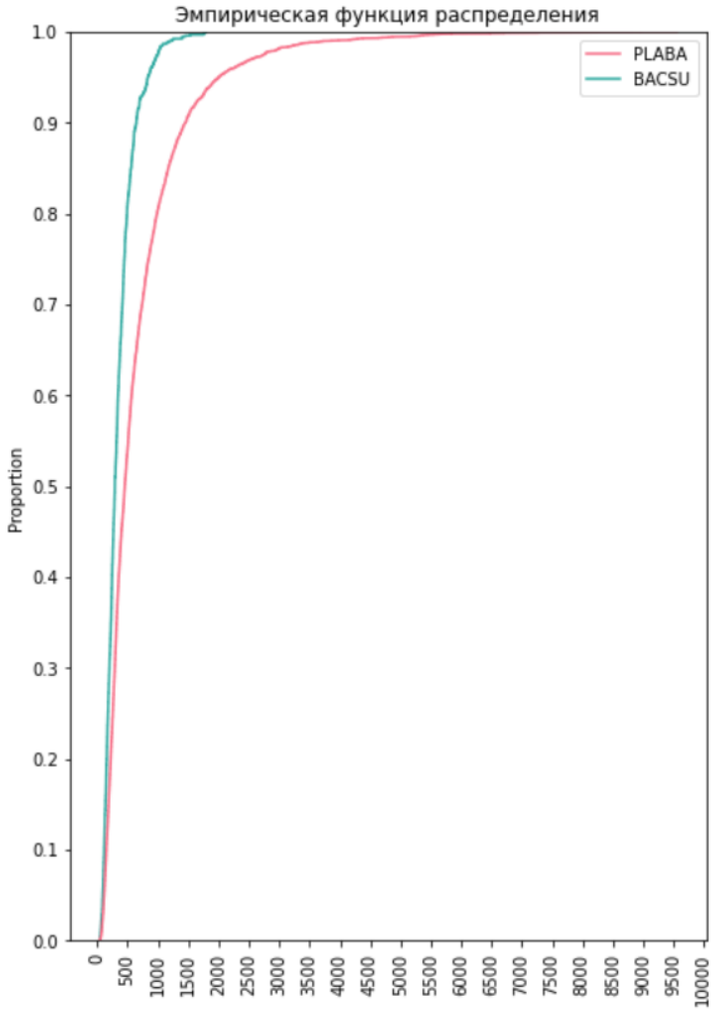
Линия посередине – среднее значение.

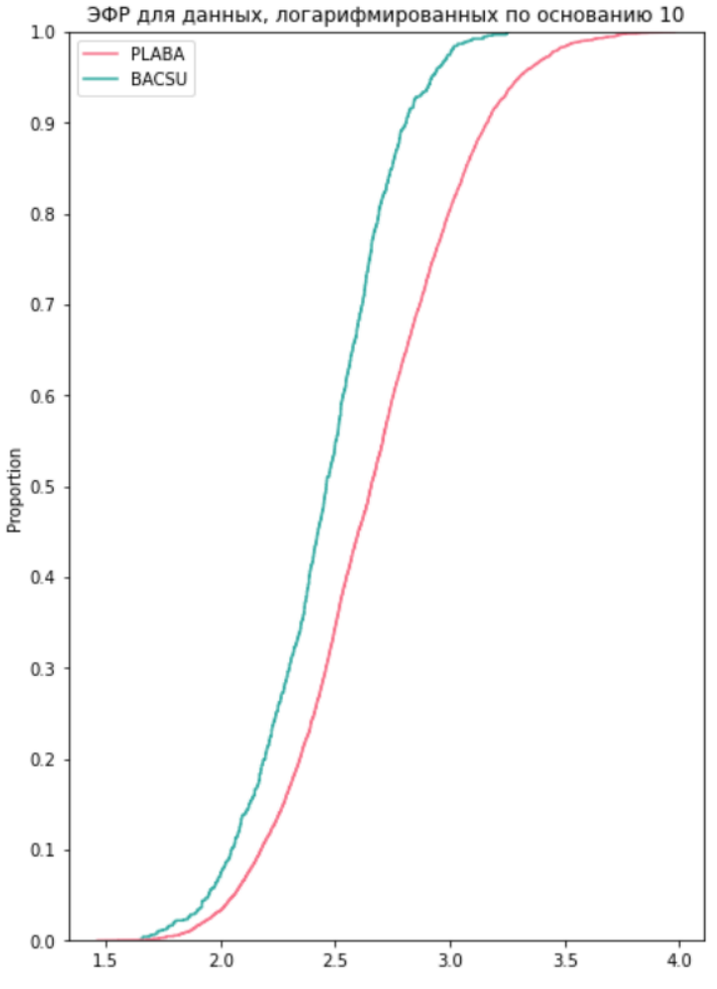
Использовано 3 разные шкалы для наглядности распределения длин в нуклеотидах,

логарифмической шкале и в логарифмированных данных

3. Эмпирическая функция распределения и ЭФР для данных, логарифмированных по основанию 10
Команда в питоне:

```
sns.ecdfplot(data_for_boxplot,palette = sns.color_palette("husl", 2), ax = axs[0]).set_titl
e('Эмпирическая функция распределения')
axs[0].set_xticks(range(0,10001,500))
```





4. Выводы

Как уже было показано, длина PLABA в среднем больше, чем у BACSU, если быть точнее то более 55% значений в эукариоте длиннее среднего значения в бактерии. Это очень хорошо видно на совместных графиках.

Графически можно заметить, что нет длины меньше 25 нуклеотидов (у бактерий и эукариотов) и нет длины больше(10 000) это можно подтвердить точным анализом столбцов длин :

Минимальная длина в нуклеотидах для эукариота = 29, для бактерии = 45.

Максимальная длина в нуклеотидах для эукариота = 9556, для бактерии = 1786.

Вычислительная часть (программа) доступна по ссылке

https://colab.research.google.com/drive/1MGTTXoRM_HKevWq9bXv910zgM7C_yyJH?usp=sharing