Дифференциально экспрессированные гены

**1. Применение t-критерия Стьюдента**

* Проверьте, что вы используете правильную версию критерия Стьюдента.
* Обратите внимание, что в этом задании **не требуется** делать множественную поправку.

**2. Поправка методом Холма**

* Для сдачи задания необходимо использовать функцию **smm.multipletests**из модуля**statsmodels.stats.multitest.**
* Обратите внимание на уровень значимости, который вы используете. Чему он равен?
* Внимательно проверьте, как вы реализовали функцию **fold\_change**. Для ускорения работы функции можете попробовать её векторизовать – сделайте так, чтобы на вход передавались массивы средних уровней экспрессии.

**3. Поправка методом Бенджамини-Хохберга**

* Решение этого задания отличается от решения предыдущего одним параметром функции **smm.multipletests**.
* Итак, я считываю данные
* data = pd.read\_csv('gene\_high\_throughput\_sequencing.csv')
* потом создаю датафрейм со столбцами как у исходной, но без айди пациента и диагноза - т.е. одни гены, а строки ее равняются уровням значимости для проверки гипотезы о равенстве средних у двух выборок. В первом случае это
* data[data['Diagnosis'] == 'normal' и data[data['Diagnosis'] == 'early neoplasia']
* а во втором:
* data[data['Diagnosis'] == 'early neoplasia'] и data[data['Diagnosis'] == 'cancer']
* Заполнение произвожу вот так:
* p\_values = pd.DataFrame(index = [0, 1], columns = data.columns[2:])
* for column in p\_values.columns:
* p\_values[column][0] = scipy.stats.ttest\_ind(data[data['Diagnosis'] == 'normal'][column], data[data['Diagnosis'] == 'early neoplasia'][column], equal\_var = False).pvalue
* ну и вторую строчку аналогично, только с соответствующими диагнозами.
* на первое задание ответы пишу вот так:
* write\_answer((p\_values.iloc[0] < 0.05).sum(), "answer1\_1.txt")
* Они подходят
* на второе провожу поправку Холма вот так
* import statsmodels.stats.multitest as smm
* reject\_0, p\_corrected\_0 = smm.multipletests(p\_values.iloc[0], alpha=0.025, method='h')[:2]
* И вот отвергнутых гипотез получается совсем мало, а когда я нахожу их номера и вычисляю fold\_change для средних по этим генам в выборках normal и early neoplasia
* c = data[data['Diagnosis'] == 'normal'].iloc[:, num].mean()
* t = data[data['Diagnosis'] == 'early neoplasia'].iloc[:, num].mean()
* то ни одна из них по модулю не выходит за 1.5
* fold change определяю вот как:
* def fold\_change(C, T):
* if T > C:
* r = T/float(C)
* else:
* r = -C/float(T)
* return r
* Может, кто понимает - в чем у меня проблема? (извиняюсь, что много кода, но словами очень долго писать бы это было, да и не до конца понятно). (Еще встречал у некоторых на форуме - что в fold change подставляют уровни значимости, а не средние - может тут что-то зарыто?)

Короче я к уровням значимости то когда поправку применял - получал номера отвергнутых гипотез, а потом их применял к начальному dataframe, а в нем ведь еще два столбца есть, не относящихся к генам - вот фигня и считалась. Сделал в индексах плюс 2 и все срослось.

Долго возился с 2.2 и 3.2 , причем 2.1 и 3.1 ответ подошел практически сразу. На самом деле пришлось все переделать - в fold change нужно подставлять среднее по соответствующим столбцам и все покатит

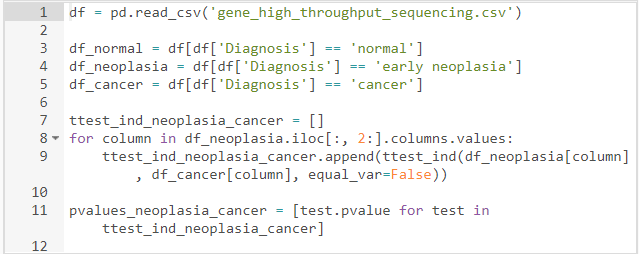
а какие столбы имеется в виду? я нашел список столбцов у которых reject == True и p\_corrected<0.025 и что-то не возьму в толк какие средние значения тут нужно брать

там где reject == True у тебя есть имена ген, вот по ним и берешь среднее из соответствующих выборок normal, cancer, early neoplasia

Не могу справиться с 3.2

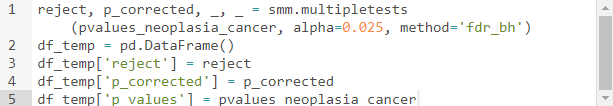
Я делаю так:

Получаем массив из p\_values для неоплазии и рака:



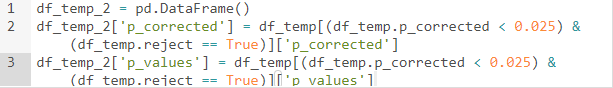
Длина pvalues\_neoplasia\_cancer: 15748

Далее делаем поправку методом Бенджамини-Хохберга и заполняем временный датафрейм

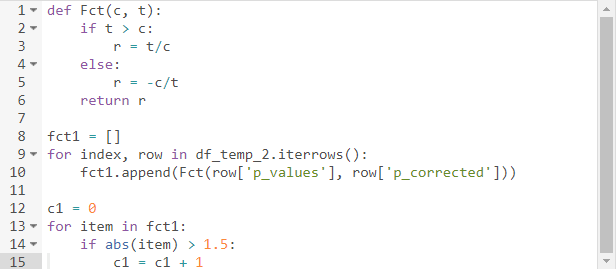


оличество reject == True: 832

Далее формируем темповый датафрейм с p\_corrected < 0.025 и rejected == True



Далее получаем массив из метрик и считаем количество abs(fold change) > 1.5:



в чем может причина, задание 2.2. не принимается (2.1 принято)

количество считаю так:

если reject = True + p\_corrected < 0.025 + fold\_change(p\_corrected, p) > 1.5

число получается двухзначным в отличие от 2,1

Когда я делал:



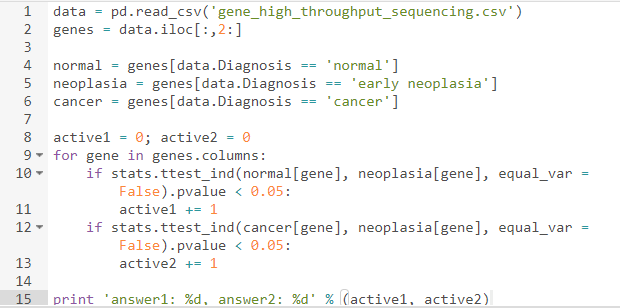
Я получал неправильный ответ. Он был завышен. Как только я сделал



мой ответ неожиданно стал правильным. Убил на этот танец с бубном несколько часов. Прекрасно понимаю, что первый вариант это извращение. Откуда берётся разница в значениях я так и не понял.

Нужно записывать через f.write(str(res)), ответы правильные!!!

Не принимает задания 1.1 и 1.2

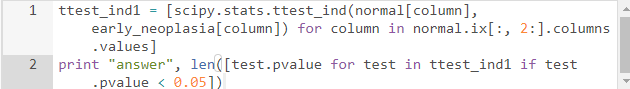


В файлах. Код медленный, но верный. Проверьте, чтобы число байтов было строго равно числу символов, или запишите файлы питоном:



Не принимает 1.1 и 1.2 задания, хотя задание вроде простое.

Я разделил выборку на 3 части:



1) Обрати внимание на параметры функции ttest\_ind...

2) Ты уверен что параметры распределения одинаковы для обоих пар генов?

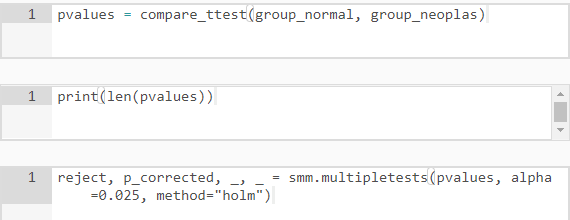
Подсказали что я забыл equal\_var=False

Часть 1 решил, а в частях 2-3 где-то ошибка. Альфа везде 0.025, метод соответствующий части в Multipletests.

При этом в задании 2.1, в сравнении normal и early\_neoplasia всего несколько генов после поправки имеет pvalue<0.025, и посчитав просто по-отдельности fold\_rate для каждого получаю тот же самый ответ что и в общем решении, и грейдер его не принимает.

Вроде все примечания что выше указаны сделал, непонятно уже где искать ошибку

В Задании 2.1 я передаю в multipletest p-value от сравнения normal и early\_neoplasia (целиком, т.е. длина 15000+), с альфа 0.025



Длина pvalues - 15748;

compare\_ttest возвращает pvalues без коррекции, она использовалась для решения Части 1, ответы к которой были зачтены.

В p\_corrected всего несколько значения меньше 0.025

Fold change нужно считать по значениям reject. Ответом здесь будет количество значимых отличий c abs(fold change) > 1.5 при сравнении normal и early neoplasia

Fold change рассчитывается по приведенной в задании формуле для каждой из групп: **normal (control)** и **early neoplasia (treatment);** **early neoplasia (control)** и **cancer (treatment)**. Где С,T (в формуле) — средние значения экспрессии гена в control и treatment группах соответственно. В результате вы должны получить список значений fold change, каждое из которых сравните с 1.5 и посчитаете количество значимых отличий

Уточните, пожалуйста, какие ряды вы берете в качестве control и treatment в каждом из случаев

два ряда pval\_1 и pval\_2... по ним ответы 1 и 2 правильные

во время построения рядов, считаю средние значения по каждому гену

if (Fct(means1[i], means2[i]) > 1.5 ) :

Функция Fct возвращает abs() значения fold change? - Да