Programming Assignment: Дифференциально экспрессированные гены

You have not submitted. You must earn 10/12 points to pass.

It looks like this is your first programming assignment. Learn more

Deadline Pass this assignment by August 5, 11:59 PM PDT

Instructions

My submission

Discussions

Обнаружение статистически значимых отличий в уровнях экспрессии генов больных раком

Это задание поможет вам лучше разобраться в методах множественной проверки гипотез и позволит применить ваши знания на данных из реального биологического исследования.

В этом задании вы:

- вспомните, что такое t-критерий Стьюдента и для чего он применяется
- сможете применить технику множественной проверки гипотез и увидеть собственными глазами, как она работает на реальных данных
- почувствуете разницу в результатах применения различных методов поправки на множественную проверку

Основные библиотеки и используемые методы:

Библиотека scipy и основные статистические функции:http://docs.scipy.org/doc/scipy/reference/stats.html#statistical-functions

Библиотека statmodels для методов коррекции при множественном сравнении:

http://statsmodels.sourceforge.net/devel/stats.html

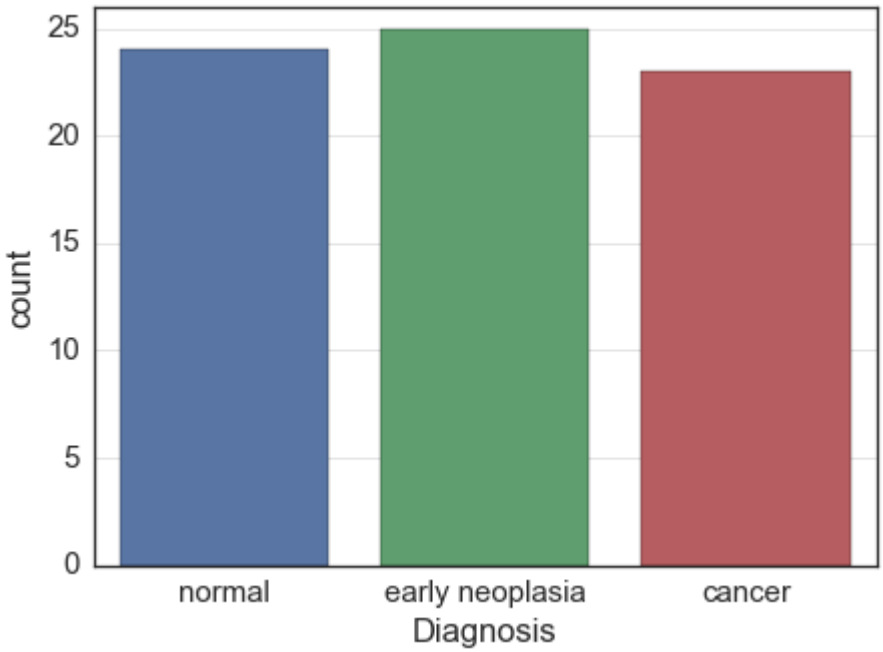
Статья, в которой рассматриваются примеры использования statsmodels для множественной проверки гипотез:

http://jpktd.blogspot.ru/2013/04/multiple-testing-p-value-corrections-in.html

Описание используемых данных

Данные для этой задачи взяты из исследования, проведенного в Stanford School of Medicine. В исследовании была предпринята попытка выявить набор генов, которые позволили бы более точно диагностировать возникновение рака груди на самых ранних стадиях.

В эксперименте принимали участие 24 человек, у которых не было рака груди (normal), 25 человек, у которых это заболевание было диагностировано на ранней стадии (early neoplasia), и 23 человека с сильно выраженными симптомами (cancer).



Ученые провели секвенирование биологического материала испытуемых, чтобы понять, какие из этих генов наиболее активны в клетках больных людей.

Секвенирование — это определение степени активности генов в анализируемом образце с помощью подсчёта количества соответствующей каждому гену РНК.

В данных для этого задания вы найдете именно эту количественную меру активности каждого из 15748 генов у каждого из 72 человек, принимавших участие в эксперименте.

Вам нужно будет определить те гены, активность которых у людей в разных стадиях заболевания отличается статистически значимо.

Кроме того, вам нужно будет оценить не только статистическую, но и практическую значимость этих результатов, которая часто используется в подобных исследованиях.

Диагноз человека содержится в столбце под названием "Diagnosis".

Практическая значимость изменения

Цель исследований — найти гены, средняя экспрессия которых отличается не только статистически значимо, но и достаточно сильно. В экспрессионных исследованиях для этого часто используется метрика, которая называется fold change (кратность изменения). Определяется она следующим образом:

$$F_c(C,T) = \begin{cases} \frac{T}{C}, & T > C \\ -\frac{C}{T}, & T < C \end{cases}$$

где C, T — средние значения экспрессии гена в control и treatment группах соответственно. По сути, fold change показывает, во сколько раз отличаются средние двух выборок.

Инструкции к решению задачи

Задание состоит из трёх частей. Если не сказано обратное, то уровень значимости нужно принять равным 0.05.

Часть 1: применение t-критерия Стьюдента

В первой части вам нужно будет применить критерий Стьюдента для проверки гипотезы о равенстве средних в двух независимых выборках. Применить критерий для каждого гена нужно будет дважды:

1. для групп **normal (control)** и **early neoplasia (treatment)**
2. для групп **early neoplasia (control)** и **cancer (treatment)**

В качестве ответа в этой части задания необходимо указать количество статистически значимых отличий, которые вы нашли с помощью t-критерия Стьюдента, то есть число генов, у которых p-value этого теста оказался меньше, чем уровень значимости.

Часть 2: поправка методом Холма

Для этой части задания вам понадобится модуль **multitest** из statsmodels.

```
1  import statsmodels.stats.multitest as smm
```

В этой части задания нужно будет применить поправку Холма для получившихся двух наборов достигаемых уровней значимости из предыдущей части. Обратите внимание, что поскольку вы будете делать поправку для каждого из двух наборов p-value отдельно, то проблема, связанная с множественной проверкой останется.

Для того, чтобы ее устранить, достаточно воспользоваться поправкой Бонферрони, то есть использовать уровень значимости 0.05 / 2 вместо 0.05 для дальнейшего уточнения значений p-value с помощью метода Холма.

В качестве ответа к этому заданию требуется ввести количество значимых отличий в каждой группе после того, как произведена коррекция Холма-Бонферрони. Причем это число нужно ввести с учетом практической значимости: посчитайте для каждого значимого изменения fold change и выпишите в ответ число таких значимых изменений, абсолютное значение fold change которых больше, чем 1.5.

Обратите внимание, что

- применять поправку на множественную проверку нужно **ко всем значениям достигаемых уровней значимости, а не только для тех, которые меньше значения уровня доверия.**
- при использовании поправки на уровне значимости 0.025 **меняются значения достигаемого уровня значимости, но не меняется значение уровня доверия** (то есть для отбора значимых изменений скорректированные значения уровня значимости нужно сравнивать с порогом 0.025, а не 0.05)!

Часть 3: поправка методом Бенджамини-Хохберга

Данная часть задания аналогична второй части за исключением того, что нужно будет использовать метод Бенджамини-Хохберга.

Обратите внимание, что методы коррекции, которые контролируют FDR, допускает больше ошибок первого рода и имеют большую мощность, чем методы, контролирующие FWER. Большая мощность означает, что эти методы будут совершать меньше ошибок второго рода (то есть будут лучше улавливать отклонения от H_0, когда они есть, и будут чаще отклонять H_0, когда отличий нет).

В качестве ответа к этому заданию требуется ввести количество значимых отличий в каждой группе после того, как произведена коррекция Бенджамини-Хохберга, причем так же, как и во второй части, считать только такие отличия, у которых abs(fold change) > 1.5.

Данные для выполнения задания

gene_high_throughput_sequencing.csv

P.S. Вспомните, какое значение имеет уровень значимости α в каждой из поправок: Холма и Бенджамини-Хохберга. Одинаковый ли смысл имеет уровень значимости в каждой из поправок?

How to submit

When you're ready to submit, you can upload files for each part of the assignment on the "My submission" tab.