Лабораторная работа №4

Обнаружение статистически значимых отличий в уровнях экспрессии генов больных раком

Данные для исследования

- Данные 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 **72** человека:
  - 24 человек, у которых не было рака груди (***normal***),
  - 25 человек, у которых это заболевание было диагностировано на ранней стадии (***early neoplasia***),
  - 23 человека с сильно выраженными симптомами заболевания (***cancer***).

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

- Секвенирование — это определение степени активности генов в анализируемом образце с помощью подсчёта количества соответствующей каждому гену РНК.
- Генами называют неразрывные протяженные участки линейных молекул нуклеиновых кислот (ДНК, редко – РНК), которые несут информацию об определенных признаках или функциях организма.

Цель — 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* (F_c, 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, F_c показывает, во сколько раз отличаются средние двух выборок.

Задание 1

Примените критерий Стьюдента для проверки гипотезы о равенстве средних в двух независимых выборках.

Применить критерий для каждого гена нужно будет дважды:

- для групп **normal** и **early neoplasia**,
- **early neoplasia** и **cancer**.

В качестве ответа на это задание назовите 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 (0.05).

Задание 2

Примените поправку Холма для получившихся двух наборов достигаемых уровней значимости из задания 1. 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 назвать количество значимых отличий в каждой группе после того, как произведена коррекция Холма-Бонферрони. Причем это число нужно назвать 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.

Задание 3

Примените поправку методом 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 назвать количество значимых отличий в каждой группе после того, как произведена коррекция Бенджамини-Хохберга, причем так же, как и во втором задании, считать только такие отличия, у которых abs(fold change) > 1.5.

Задача множественной проверки гипотез

Пусть имеется m выборок, каждая своего размера, и из своего распределения. Каждой выборке соответствует своя нулевая гипотеза H_i и альтернатива H'_i. Каждая из гипотез проверяется своей статистикой T_i. Для каждой из статистик известно свое нулевое распределение. Таким образом, можно вычислить достигаемые уровни значимости всех гипотез: $p_i, \quad i = 1, \ldots, m.$

Для этого вводятся следующие обозначения:

Пусть $\mathbf{M}$ — это множество индексов: $\mathbf{M} = \{1, 2, \ldots, m\}$;

$\mathbf{M}_0$ — это множество индексов верных нулевых гипотез, пусть его мощность равна m_0:

$$\mathbf{M_0} = \{i\colon H_i \text{ верна}\}, \quad |\mathbf{M_0}| = m_0.$$

Естественно, это множество неизвестно, потому что иначе не было бы смысла проверять гипотезы. Пусть $\mathbf{R}$ — это множество индексов отвергаемых гипотез, а его мощность равна R:

$$\mathbf{R} = \{i\colon H_i \text{ отвергнута}\}, \quad |\mathbf{R}| = R$$

Задача множественной проверки гипотез

Тогда пересечение множеств $\mathbf{R}$ и $\mathbf{M_0}$ состоит из неверно отвергнутых гипотез. Мощность этого множества обозначается $\mathbf{V}$, это есть число ошибок первого рода:

$$V = |\mathbf{M_0} \cap \mathbf{R}| .$$

	# верных H_i	# неверных H_i	Σ
# принятых H_i	U	T	$m - R$
# отвергнутых H_i	V	S	R
Σ	m_0	$m - m_0$	m

По аналогии с задачей однократной проверки гипотез можно составить таблицу 2×2, в которой будет стоят количество верных и неверных, принятых и отвергнутых гипотез Из всех величин, записанных в таблице, известна только m — общее число гипотез. А единственный параметр, которым можно управлять, — это R, количество отвергаемых гипотез. При этом самая пугающая величина — это V, количество ошибок первого рода. Хочется совершать мало ошибок первого рода, но при этом единственное, что можно делать, — это перераспределять по этой таблице гипотезы из второй строки в первую. То есть, чтобы совершать мало ошибок первого рода, нужно отвергать меньше гипотез.

Задача множественной проверки гипотез поставлена, теперь нужно её решить. Интерес представляет некоторая статистическая процедура, которая дает гарантии на значение V – оно не должно быть слишком большим.

Напрямую с V работать не очень удобно, поэтому, как правило, берут некоторые меры, определенные над V, и работают с ними. Одна из самых распространенных таких мер — это групповая вероятность ошибки первого рода (**familywise error rate**). По определению **это вероятность совершить хотя бы одну ошибку первого** $\mathrm{FWER} = P(V > 0)$.

Эту величину хочется контролировать на уровне α: $\mathrm{FWER} = P(V > 0) \leq \alpha.$

То есть, хочется построить такую статистическую процедуру, что вероятность совершить хотя бы одну ошибку первого рода будет не больше, чем α.

Единственный имеющийся в распоряжении инструмент — это уровни значимости $\alpha_1,\ldots,\alpha_m$, на которых проверяются гипотезы $H_1,\ldots,H_m$. Никаких других параметров в проверке гипотез нет. Ставится задача выбрать эти уровни так, чтобы обеспечить ограничение $\mathrm{FWER} \leq \alpha.$

Поправка Бонферрони

Самый простой способ решить поставленную выше задачу — это использовать поправку Бонферрони. В методе Бонферрони достигаемые уровни значимости всех гипотез сравниваются с величиной $\frac{\alpha}{m}$:

$$\alpha_1 = \ldots = \alpha_m = \frac{\alpha}{m}.$$

Альтернативный способ — преобразовать все достигаемые уровни значимости (p-value): $\tilde{p}_i = \min(1, mp_i)$.

Эти модифицированные достигаемые уровни значимости и будут сравниваться с исходным порогом α: H_i отвергается при $\tilde{p}_i \leq \alpha$. При такой процедуре точно так же контролируется величина FWER, как и при изменении порога.

Недостаток использования поправки Бонферрони: при использовании метода Бонферрони FWER не просто меньше, чем α, а намного меньше, чем α. В идеале хочется, чтобы вероятность совершить хотя бы одну ошибку первого рода была в точности равна α. При использовании метода Бонферрони эта вероятность ограничивается гораздо более низкой величиной, чем α. Это плохо, потому что перестраховываясь в отношении ошибки первого рода, мы неизбежно совершаем больше ошибок второго рода, то есть мощность такой статистической процедуры снижается.

В методе Бонферрони уровни значимости для всех гипотез выбираются одинаковыми: $\alpha_1 = \ldots = \alpha_m = \frac{\alpha}{m}$.

Оказывается, если значения a_i брать не одинаковыми, а разными, можно достичь лучшего результата. Для того, чтобы это сделать, необходимо использовать *нисходящую процедуру множественной проверки гипотез*. В общем виде она выглядит так: из достигаемых уровней значимости составляется вариационный ряд: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$,

а все гипотезы переобозначаются так, чтобы их номера соответствовали номерам достигаемых уровней значимости в этом вариационном ряду:

$$H_{(1)}, H_{(2)}, \ldots, H_{(m)}.$$

Дальше нужно самый маленький достигаемый уровень значимости $p_{(1)}$ сравнить с уровнем значимости α_1. Если $p_{(1)} \geq \alpha_1$, то принимаются все нулевые гипотезы $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$, и процесс останавливается. $p_{(1)} < \alpha_1$, то отклоняется гипотеза $H_{(1)}$, и процедура продолжается. На втором шаге сравниваются $p_{(2)}$ и α_2 и т.д.

Так в общем виде выглядит *нисходящая процедура множественной проверки гипотез*.

Метод Холма

Метод Холма — это нисходящая процедура множественной проверки гипотез со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m},\ \alpha_2 = \frac{\alpha}{m-1},\ \ldots,\ \alpha_i = \frac{\alpha}{m-i+1},\ \ldots,\ \alpha_m = \alpha.$$

Этот метод обеспечивает безусловный контроль над FWER. Это показать немного сложнее, чем для метода Бонферрони, поэтому доказательство здесь приведено не будет. Вместо того, чтобы сравнить исходные достигаемые уровни значимости с модифицированными α_i, можно их модифицировать и сравнивать с исходным порогом α. Так выглядит формула для модифицированных достигаемых уровней значимости метода Холма:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-i+1)\,p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

Метод Холма всегда мощнее, чем метод Бонферрони, то есть, он всегда отвергает не меньше гипотез, чем метод Бонферрони, потому что его уровни значимости всегда не меньше, чем из метода Бонферрони.

В описанных ранее поправках при множественном проверке гипотез контролировалась величина групповой вероятности ошибки, то есть ограничивалась вероятность совершить хотя бы одну ошибку первого рода:

$$\mathrm{FWER} = P\left(V > 0\right).$$

В некоторых ситуациях, например, когда проверяются десятки тысяч или миллионы гипотез, можно допустить какое-то количество ошибок первого рода ради того, чтобы увеличить мощность процедуры и отвергнуть больше неверных гипотез, то есть совершить меньше ошибок второго рода. В таких ситуациях выгоднее использовать другую меру: не familywise error rate, а **false discovery rate**, **ожидаемую долю ложных отклонений**:

$$\mathrm{FDR} = \mathbb{E}\left(\frac{V}{\max\left(R, 1\right)}\right).$$

Для любой фиксированной процедуры множественной проверки гипотез FDR ≤ FWER. За счет этого, если контролировать FDR, а не FWER, получается более мощная процедура, поскольку она позволяет отвергать больше гипотез.

Методы, которые контролируют FDR, как правило, восходящие. В каком-то смысле это противоположность нисходящих методов (таких как метод Холма), которые рассматривались до этого. Восходящие методы работают с тем же самым вариационным рядом достигаемых уровней значимости, что и нисходящие: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.

Отличие заключается в том, что процедура начинается с другого конца этого ряда. На первом шаге самый большой p-value p_m сравнивается с соответствующей ему константой α_m. Если $p_{(m)} \leq \alpha_m$, то все нулевые гипотезы $H_{(1)}$, $H_{(2)}$, ..., $H_{(m)}$ отвергаются, и процедура останавливается. Иначе гипотеза $H_{(m)}$ принимается, и процедура продолжается. На следующем шаге сравниваются $p_{(m\text{-}1)}$ и $\alpha_{m\text{-}1}$. Если $p_{(m\text{-}1)} \leq \alpha_{m\text{-}1}$, то все нулевые гипотезы $H_{(1)}$, $H_{(2)}$, ..., $H_{(m\text{-}1)}$ отвергаются, и процедура останавливается. Иначе принимается гипотеза $H_{(m\text{-}1)}$, процедура продолжается. И так далее.

Такие методы называются *восходящими*.

Метод Бенджамини-Хохберга

Для контроля над FDR чаще всего используется метод Бенджамини-Хохберга. Это восходящая процедура с уровнями значимости

$$\alpha_1 = \frac{\alpha}{m}, \ \ldots, \ \alpha_i = \frac{\alpha i}{m}, \ \ldots, \ \alpha_m = \alpha.$$

Крайние уровни значимости точно также же ,как и в методе Холма, а вот между ними — абсолютно другие. В методе Бенджамини-Хохберга уровни значимости между α_1 и α_m меняются линейно, в то время как в методе Холма — по гиперболе. Модифицированные достигаемые уровни значимости для метода Бенджамини-Хохберга выглядят следующим образом:

$$\tilde{p}_{(i)} = \min\left(1, \frac{mp_{(i)}}{i}, \tilde{p}_{(i+1)}\right).$$

Процедура восходящая, и каждый следующий p-value в ней не должен стать больше, чем предыдущий, поэтому берётся минимум из $\frac{mp_{(i)}}{i}$ и $\tilde{p}_{(i+1)}$ (а также 1, поскольку это вероятность). Метод Бенджамини-Хохберга обеспечивает контроль над FDR на уровне α только при условии независимости статистик, которые проверяют гипотезы. Это требование достаточно сильное. Иногда его можно ослабить, и в некоторых задачах выполняется ослабленное требование. Тем не менее, важно подчеркнуть, что процедура Бенджамини-Хохберга не является универсальной и она не применима безусловно, в отличие метода Холма.

Реализация в Python

- t-критерий Стьюдента

- *независимые выборки*:

scipy.stats.ttest_ind(*a*, *b*, *axis* = 0, *equal_var* = True)

- *a*, *b* – массивы данных,
- *axis* – ось,
- *equal_var* = True – при равных дисперсиях, False – при различных.

- *зависимые выборки*:

scipy.stats.ttest_rel(*a*, *b*, *axis* = 0)

• Поправки на множественную проверку

statsmodels.stats.multitest.multipletests(*pvals, alpha = 0,05, method = 'holm'*)

- *pvals* – массив значений p-value (нескорректированных),
- *alpha* – уровень значимости,
- *method – метод:*
  - *‘bonferroni’ – Бонферрони,*
  - *‘holm’* – Холма,
  - *‘fdr_bh’ – Бенджамини-Хохберга,*

  и др.

Функция возвращает:

reject – верна ли гипотеза для заданного уровня значимости,

pvals_corrected – скорректированные значения p-value.

Задание 1

1) Загрузка данных

```
import pandas as pd
df = pd.read_csv('gen.csv')
df.head()
```

	Patient_id	Diagnosis	LOC643837	LOC100130417	SAMD11	NOC2L	KLHL17	PLEKHN1	C1orf170	HES4	...
0	STT5425_Breast_001_normal	normal	1.257614	2.408148	13.368622	9.494779	20.880435	12.722017	9.494779	54.349694	...
1	STT5427_Breast_023_normal	normal	4.567931	16.602734	42.477752	25.562376	23.221137	11.622386	14.330573	72.445474	...
2	STT5430_Breast_002_normal	normal	2.077597	3.978294	12.863214	13.728915	14.543176	14.141907	6.232790	57.011005	...
3	STT5439_Breast_003_normal	normal	2.066576	8.520713	14.466035	7.823932	8.520713	2.066576	10.870009	53.292034	...
4	STT5441_Breast_004_normal	normal	2.613616	3.434965	12.682222	10.543189	26.688686	12.484822	1.364917	67.140393	...

5 rows × 15750 columns

2) Разделим данные на 3 группы по диагнозу (столбец Diagnosis):

- normal (нет рака),
- early neoplasia (заболевание на ранней стадии),
- cancer (сильно выраженные симптомы рака)

```
normal = df.loc[df.Diagnosis == 'normal']
early_neoplasia = df.loc[df.Diagnosis == 'early neoplasia']
cancer = df.loc[df.Diagnosis == 'cancer']
```

3) Для дальнейшего удобства создадим массив с идентификаторами всех генов, присутствующих в выборке

```
import numpy as np
genes = np.array(df.columns)[2:]
genes
```

```
array(['LOC643837', 'LOC100130417', 'SAMD11', ..., 'CYorf15B', 'KDM5D',
       'EIF1AY'], dtype=object)
```

4) Примерим t-критерий Стьюдента

```
from scipy import stats
ttest_1 = [] # В этом списке будут храниться значения p-value для каждого гена групп normal и early neoplasia
ttest_2 = [] # В этом списке будут храниться значения p-value для каждого гена групп early neoplasia и cancer

# Для каждого гена применяем t-критерий
for gene in genes:
    ttest_1.append(stats.ttest_ind(normal[gene], early_neoplasia[gene], equal_var = False).pvalue)
    ttest_2.append(stats.ttest_ind(early_neoplasia[gene], cancer[gene], equal_var = False).pvalue)
```

5) Определим число генов, у которых p-value оказался меньше, чем уровень значимости (0.05)

```
alpha = 0.05 # Уровень значимости

# Преобразуем списки в питру-массивы
ttest_1 = np.array(ttest_1)
ttest_2 = np.array(ttest_2)

print(np.sum(ttest_1 < alpha))
print(np.sum(ttest_2 < alpha))
```

```
1575
3490
```

Задание 2

6) Поправка методом Холма

```
import statsmodels.stats.multitest as smm
rej_1, correct_1 = smm.multipletests(ttest_1, alpha=alpha/2, method='holm')[:2]
rej_2, correct_2 = smm.multipletests(ttest_2, alpha=alpha/2, method='holm')[:2]
```

6) Функция для рассчета fold change

```
def fold_change(C, T):
    res = 0
    if T > C:
        res = T/C
    elif C > T:
        res = C/T
    return res
```

7) Вычислим кол-во значимых отличий в каждой группе

```
c_1 = 0
c_2 = 0

for i in range(len(genes)):
    gene = genes[i]

    # Считаем только такие отличия, у которых abs(fold change) > 1.5
    if abs(fold_change(np.mean(normal[gene]),np.mean(early_neoplasia[gene]))) > 1.5 and rej_1[i]:
        c_1+=1
    if abs(fold_change(np.mean(early_neoplasia[gene]),np.mean(cancer[gene]))) > 1.5 and rej_2[i]:
        c_2+=1

c_1, c_2
```

```
(2, 77)
```

Задание 3 выполняется аналогично заданию 2.