Иерархическая процедура множественной проверки гипотез

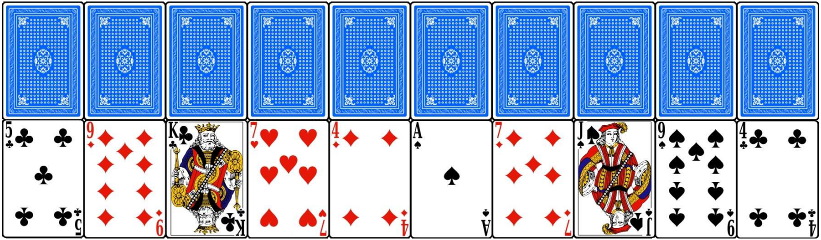
Рябенко Евгений, аспирант ММП

Семинар «Байесовские методы машинного обучения»
ВМК МГУ • 16 ноября 2011 г.

Поиск экстрасенсов

Joseph Rhine, 1950: исследования экстрасенсорного восприятия. Первый этап — поиск экстрасенсов.

Испытуемому предлагалось угадать цвет 10 карт.



H_0: испытуемый выбирает цвет карт наугад.

H_1: испытуемый может предсказывать цвет карт.

Статистика t — число угаданных цветов.

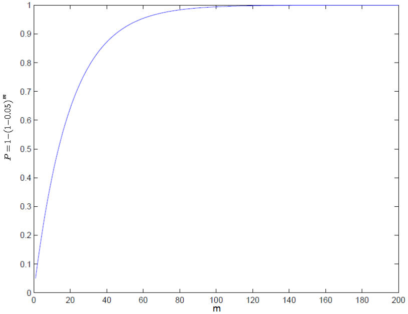
$$P\left(t \geq 9 \,|\, H_0\right) = 11 \cdot \frac{1}{2}^{10} = 0.0107421875,$$

т.е. при $t = 9$ достигаемый уровень значимости $p \approx 0.01$, событие достаточно редкое, можно отклонять H_0 и признавать испытуемого экстрасенсом.

Поиск экстрасенсов

Процедуру отбора прошли 1000 человек.
Девять из них угадали цвета 9 из 10 карт, двое — цвета всех 10 карт.
Ни один из них в последующих экспериментах не подтвердил своих способностей.

Вероятность того, что из 1000 человек хотя бы один случайно угадает цвета 9 или 10 из 10 карт: $1-\left(1-11\cdot\frac{1}{2}^{10}\right)^{1000}\approx 0.9999796.$



Математическая формулировка

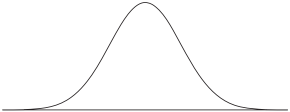
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim P \in \Omega;$

нулевая гипотеза: $H_0: P \in \omega, \quad \omega \in \Omega;$

альтернатива: $H_1: P \notin \omega;$

статистика: $T(\mathbf{X}), \quad T(\mathbf{X}) \sim F(x)$ при $P \in \omega;$

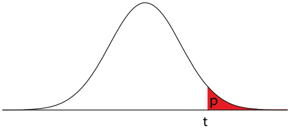
$T(\mathbf{X}) \nsim F(x)$ при $P \notin \omega;$



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\};$

реализация статистики: $t = T(\mathbf{x});$

достигаемый уровень значимости: $p(\mathbf{x})$— вероятность при H_0 получить $T(\mathbf{X}) = t$ или ещё более экстремальное;



Гипотеза отвергается при $p(\mathbf{x}) \leq \alpha, \quad \alpha$— уровень значимости.

Правило проверки гипотезы

Простейший пример

Требуется проверить симметричность монеты за 20 подбрасываний.

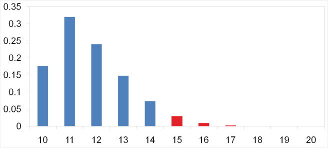
выборка: $\mathbf{X} = \{X_1, \dots, X_{20}\} - i.i.d. Bern\,(20, q)$;

нулевая гипотеза: $H_0: P = Bern\,(20, 0.5)$;

альтернатива: $H_1: P \neq Bern\,(20, 0.5)$;

статистика: $T\,(\mathbf{X}) = \max\left(\sum X_i, 20 - \sum X_i\right)$;

большие значения статистики свидетельствуют в пользу H_1;



реализация выборки: $\mathbf{x} = \{x_1, \dots, x_{20}\}$;

реализация статистики: $t = T\,(\mathbf{x})$;

достигаемый уровень значимости: при $t = 20$ $p = \frac{2}{1048576} \approx 1.9 \times 10^{-6}$,

при $t = 19$ $p = \frac{42}{1048576} \approx 4 \times 10^{-5}$,

при $t = 15$ $p = \frac{43400}{1048576} \approx 0.04$.

При $t \geq 15$ гипотеза отвергается на уровне значимости $\alpha = 0.05$.

Несимметричность задачи проверки гипотез

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода
H_0 отвергается	Ошибка первого рода	H_0 верно отвергнута

Вероятность ошибки первого рода жёстко ограничивается достаточно малой наперёд заданной величиной — $P\left(p\left(\mathbf{x}\right) \leq \alpha \,|H_0\,\right) \leq \alpha$.

Вероятность ошибки второго рода минимизируется путём выбора достаточно мощного критерия.

Усложнение примера

Требуется проверить симметричность 1000 монет.

Пусть все монеты симметричны.
Вероятность того, что хотя бы одна не менее 15 раз за серию упадёт одной и той же стороной, равна

$$1 - \left(1 - \frac{43400}{1048576}\right)^{1000} \approx 0.9999999999999999995.$$

t	15	16	17	18	19	20
P	$1-5\times10^{-19}$	$1-7\times10^{-6}$	0.92	0.33	0.039	0.002

При проведении статистического анализа данных по большому количеству гипотез необходимо ограничивать не только вероятность каждой ошибки первого рода, но и некую глобальную меру ошибки, учитывающую число гипотез.

Математическая постановка

данные:	$\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_m\} \sim P \in \Omega;$
нулевые гипотезы:	$H_i\colon\ P \in \omega_i, \quad \omega_i \in \Omega;$
альтернативы:	$H_i'\colon\ P \notin \omega_i;$
статистики:	$T_i = T(\mathbf{X}_i)$ проверяет H_i против H_i';
реализации статистики:	$t_i = t(\mathbf{x}_i);$
достигаемые уровни значимости:	$p_i = p(\mathbf{x}_i), \quad i = 1, \dots, m;$

$\mathbf{M} = \{1, 2, \dots, m\};$
$\mathbf{M_0} = \mathbf{M_0}(P) = \{i\colon H_i$ верна$\}$ — индексы верных гипотез, $|\mathbf{M_0}| = m_0;$
$\mathbf{R} = \mathbf{R}(P, \alpha) = \{i\colon H_i$ отвергнута$\}$ — индексы отвергаемых гипотез,
$|\mathbf{R}| = R;$
$V = |\mathbf{M_0} \cap \mathbf{R}|$ — число ошибок первого рода.

	Число верных H_0	Число неверных H_0	Всего
Число принятых H_0	U	T	$m - R$
Число отвергнутых H_0	V	S	R
Всего	m_0	$m - m_0$	m

FWER

Групповая вероятность ошибки первого рода (familywise error rate):

$$FWER = P\left(V \geq 1\right).$$

Контроль над групповой вероятностью ошибки на уровне α означает

$$FWER = P\left(V \geq 1\right) \leq \alpha \quad \forall P.$$

$\alpha_1, \ldots, \alpha_m$ — уровни значимости, на которых необходимо проверять гипотезы $H_1, \ldots, H_m$; задача — выбрать их так, чтобы обеспечить $FWER \leq \alpha$.

Метод Бонферрони

Метод Бонферрони: $\alpha_1 = \ldots = \alpha_m = \alpha/m$.

Theorem

Если гипотезы $H_i, \;\; i = 1, \ldots, m$, *отвергаются при* $p_i \leq \alpha/m$, *то* $FWER \leq \alpha$.

Доказательство.

$$FWER = P\left(V \geq 1\right) \leq P\left(\bigcup_{i=1}^{m_0} \{p_i \leq \alpha/m\}\right) \leq \sum_{i=1}^{m_0} P\left(p_i \leq \alpha/m\right) \leq$$

$$\leq \sum_{i=1}^{m_0} \alpha/m = \frac{m_0}{m}\alpha \leq \alpha.$$

□

Поправка Бонферрони

При увеличении m в результате применения поправки Бонферрони мощность статистической процедуры резко уменьшается — шансы отклонить неверные гипотезы падают.

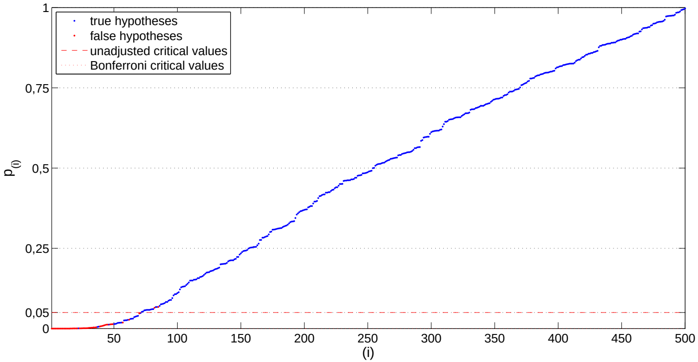
Пример: критерий Стьюдента для независимых выборок с одинаковой дисперсией $\sigma^2 = 1$ и разностью средних $\mu_1 - \mu_2 = 1$ при проверке гипотезы однородности на уровне значимости 0.05 имеет мощность 0.9 уже при длине выборок $n = 23$.

m	1	10		100		1000	
n	23	23	36	23	49	23	62
power	0.9	0.67	0.9	0.37	0.9	0.16	0.9

Если проверяется одновременно 1000000 гипотез, при размере выборок $n = 10$ мощность 0.9 достигается при расстоянии между средними выборок в пять стандартных отклонений.

Пример

$m = 500, \quad m_0 = 50$



Без поправки на множественные сравнения: $R = 71, \quad V = 23.$

С поправкой Бонферрони: $R = 9, \quad V = 0.$

Метод Холма

Вариационный ряд достигаемых уровней значимости:
$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$,
$H_{(1)}, H_{(2)}, \dots, H_{(m)}$ — соответствующие нулевые гипотезы.

Нисходящая процедура Холма:

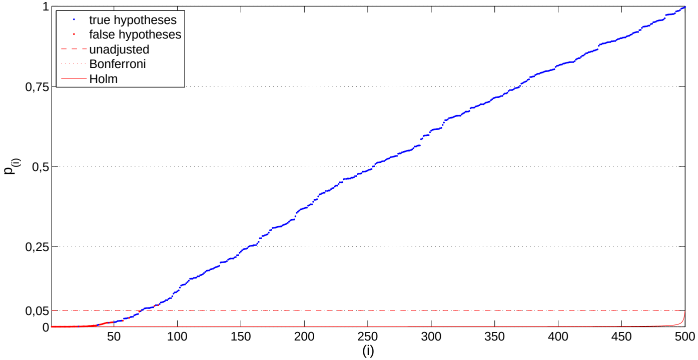
1. Если $p_{(1)} \geq \alpha_1 = \alpha/m$, принять все нулевые гипотезы $H_{(1)}, H_{(2)}, \dots, H_{(m)}$ и остановиться; иначе отвергнуть $H_{(1)}$ и продолжать.
2. Если $p_{(2)} \geq \alpha_2 = \alpha/\left(m-1\right)$, принять нулевые гипотезы $H_{(2)}, H_{(3)}, \dots, H_{(m)}$ и остановиться; иначе отвергнуть $H_{(2)}$ и продолжать.
3. ...

$$\alpha_{(1)} = \alpha/m, \alpha_{(2)} = \alpha/(m-1), \dots, \alpha_{(m)} = \alpha.$$

$FWER \leq \alpha$ обеспечивается при любых p_i.

Метод Холма

Метод Холма равномерно мощнее поправки Бонферрони, однако на практике различия между ними невелики.



Без поправки на множественные сравнения: $R = 71, \quad V = 23.$
С поправкой Бонферрони: $R = 9, \quad V = 0.$
С поправкой методом Холма: $R = 10, \quad V = 0.$

Другие методы

- Если совместное нулевое распределение статистик $T_1, \dots, T_m$ известно, константы $\alpha_{(i)}$ могут быть найдены точно:

$$P_{0,\dots,0}\left(\max\left(T_1, \dots, T_i\right) \geq c_{(m-i+1)}\right) = \alpha.$$

- Если статистики независимы, можно брать

$$\alpha_{(i)} = 1 - (1-\alpha)^{\frac{1}{m-i+1}}.$$

- Если между статистиками положительная регрессионная зависимость, то $\alpha_{(i)}$ лежат в диапазоне

$$1 - (1-\alpha)^{\frac{1}{m-i+1}} \leq \alpha_{(i)} \leq \alpha.$$

- Если выполняется свойство subset pivotality (нулевое распределение любого подмножества статистик T_i не зависит от того, верны или неверны соответствующие оставшимся статистикам гипотезы), то можно найти α_i по нулевому распределению максимальной статистики $M_T = \max_i T_i$: $c = F^{-1}_{M_T \left| \cup_{i \in \mathbf{M}} H_i \right.} (1-\alpha)$.
- Если статистики можно представить как реализацию случайного поля, используются топологические методы.

FDR

Ожидаемая доля ложных отклонений гипотез среди всех отклонений (false discovery rate):

$$FDR = \mathbb{E}\left(\frac{V}{R} \cdot I_{\{R>0\}}\right) = \mathbb{E}\,(FDP)$$

(FDP — доля ложных отклонений, false discovery proportion).

Контроль над ожидаемой долей ложных отклонений на уровне q означает

$$FDR = \mathbb{E}\left(\frac{V}{R} \cdot I_{\{R>0\}}\right) \leq q \quad \forall P.$$

$\alpha_1, \dots, \alpha_m$ — уровни значимости, на которых необходимо проверять гипотезы $H_1, \dots, H_m$; задача — выбрать их так, чтобы обеспечить $FDR \leq q$.

Байесовский вариант

Пусть $T_i \sim F_0(t),\ i \in \mathbf{M_0}, \quad T_i \sim F_1(t),\ i \in \mathbf{M} \setminus \mathbf{M_0}$.

Функция распределения смеси:

$$F(t) = \frac{m_0}{m} F_0(t) + \frac{m - m_0}{m} F_1(t).$$

$$FDR(t) = \frac{m_0}{m} F_0(t)/F_1(t).$$

Методы контроля FDR

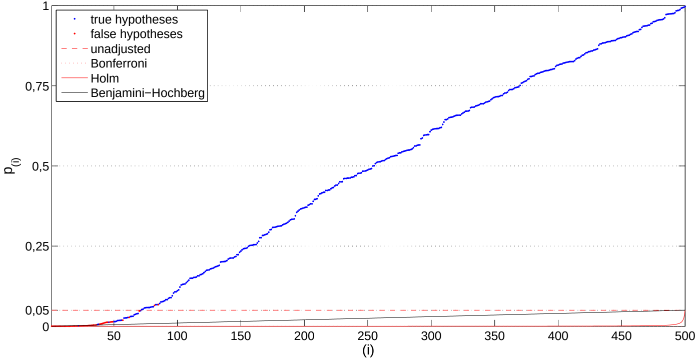
Вариационный ряд достигаемых уровней значимости:
$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$,
$H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ — соответствующие нулевые гипотезы.

Нисходящие методы:

1. Если $p_{(1)} \geq \frac{1}{m} \frac{q}{c(m)}$, принять все нулевые гипотезы $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ и остановиться; иначе отвергнуть $H_{(1)}$ и продолжать.
2. Если $p_{(2)} \geq \frac{2}{m} \frac{q}{c(m)}$, принять нулевые гипотезы $H_{(2)}, H_{(3)}, \ldots, H_{(m)}$ и остановиться; иначе отвергнуть $H_{(2)}$ и продолжать.
3. ...

При $c(m) = 1$ получаем метод Бенджамини-Хохберга; $FDR \leq m_0/mq \leq q$, если между статистиками $T_i,\ i \in \mathbf{M_0}$, положительная регрессионная зависимость (или они независимы).
При $c(m) = \sum_{i=1}^{m} \frac{1}{i}$ получаем метод Бенджамини-Йекутиели;
$FDR \leq m_0/mq \leq q$ безусловно.

Метод Бенджамини-Хохберга



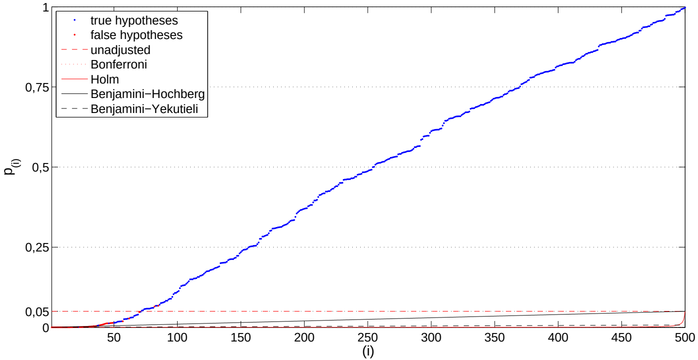
Без поправки на множественные сравнения: $R = 71, \quad V = 23.$

С поправкой Бонферрони: $R = 9, \quad V = 0.$

С поправкой методом Холма: $R = 10, \quad V = 0.$

С поправкой Бенджамини-Хохберга: $R = 33, \quad V = 1.$

Метод Бенджамини-Йекутиели



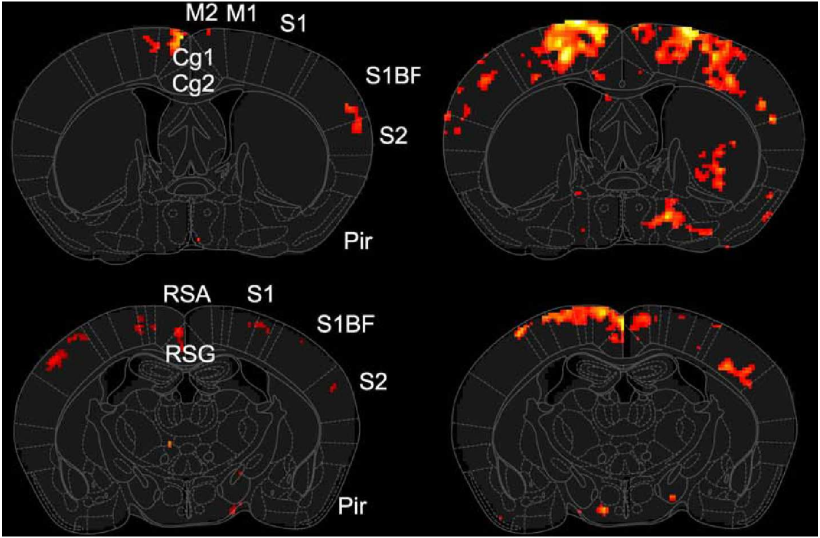
Без поправки на множественные сравнения: $R = 71, \quad V = 23.$
С поправкой Бонферрони: $R = 9, \quad V = 0.$
С поправкой методом Холма: $R = 10, \quad V = 0.$
С поправкой Бенджамини-Хохберга: $R = 33, \quad V = 1.$
С поправкой Бенджамини-Йекутиели: $R = 16, \quad V = 0.$

Способы улучшения

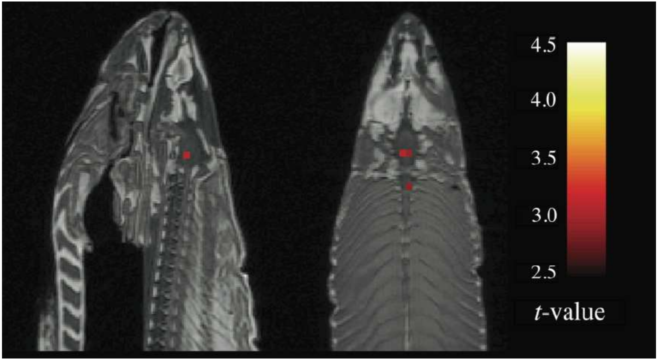
Если ожидается, что значение m_0 невелико, можно оценить его и увеличить критические константы.

- $\hat{m}_0 = 2 \times \sharp \{p_i > 0.5\}$;
- $\hat{m}_0$ —число гипотез, не отвергаемых методом Бенджамини-Хохберга при $q' = q / (q+1)$;
- ...

Экспрессия на срезах мозга

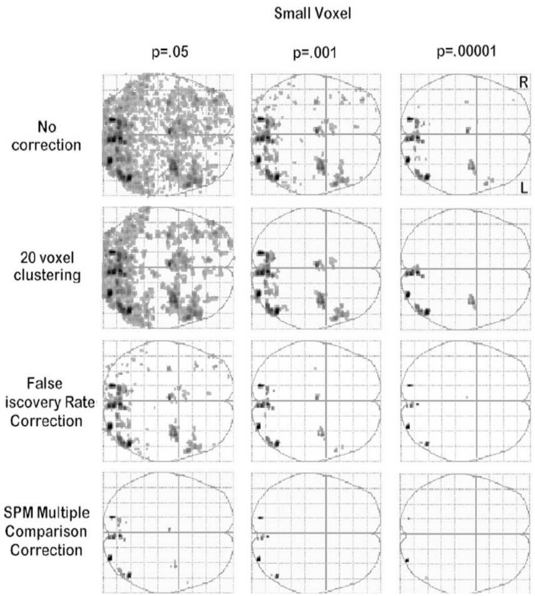


Мёртвый лосось

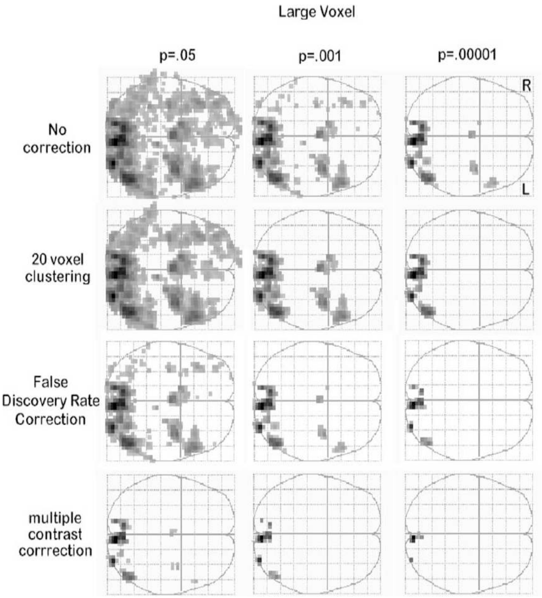


(Использовано отсечение по p с порогом $\alpha = 0.001$)

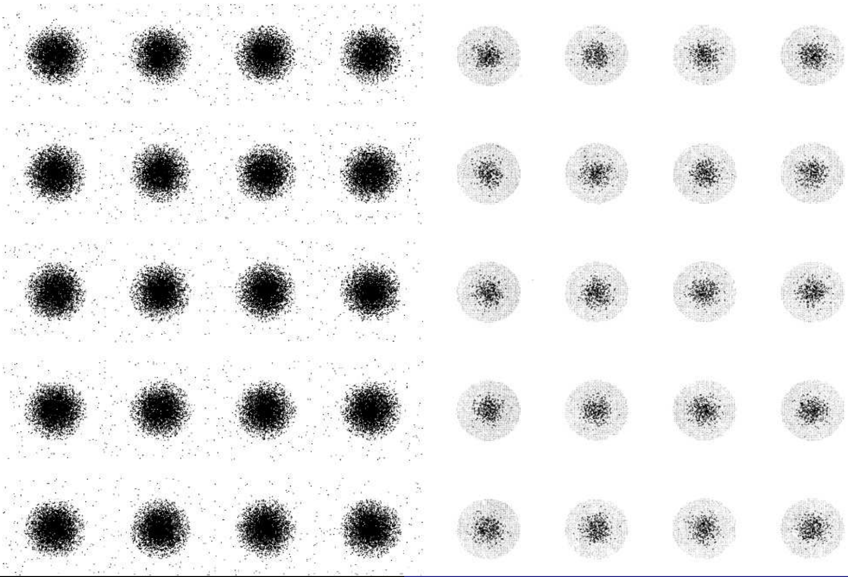
Now you see it, now you don't



Now you see it, now you don't

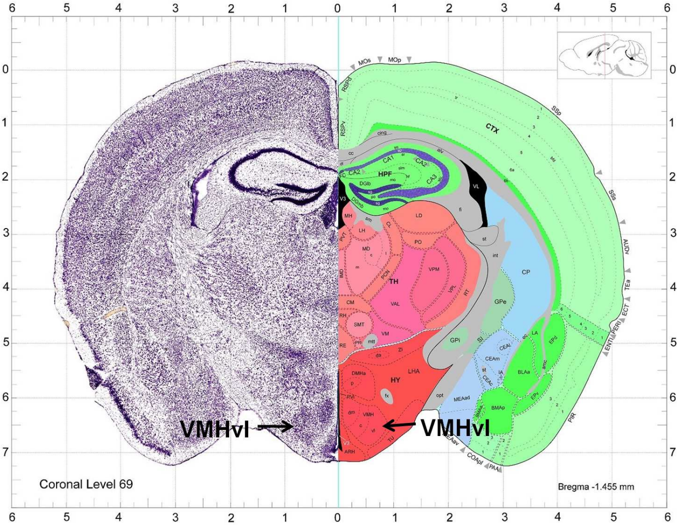


FDR vs. FWER



Семейства гипотез

Гипотезы могут быть объединены в семейства.



Типы контроля при наличии семейств

Далее рассматриваются меры ошибок первого рода, которые можно представить в виде $\mathbb{E}(C)$ для некоторой случайной меры ошибок первого рода C.

$$FWER = P\,(V \geq 1) = \mathbb{E}\left(I_{\{V\geq 1\}}\right)$$
$$FDR = \mathbb{E}\left(\frac{V}{R} \cdot I_{\{R>0\}}\right) = \mathbb{E}\,(FDP)$$

Контроль на уровне семейств	$\mathbb{E}\left(C_{fam}\right) \leq q$
Контроль внутри каждого семейства	$\mathbb{E}\left(C_i\right) \leq q, \quad i = 1, \ldots, m$
Контроль в среднем по отобранным семействам	S — индексы отобранных семейств, $$\mathbb{E}\left(\frac{\sum_{i\in S} C_i}{\max\left(\lvert S\rvert, 1\right)}\right) \leq q$$
Глобальный контроль	$\mathbb{E}\left(C_{comb}\right) \leq q$

Два экстремальных подхода

Агрегированная проверка:

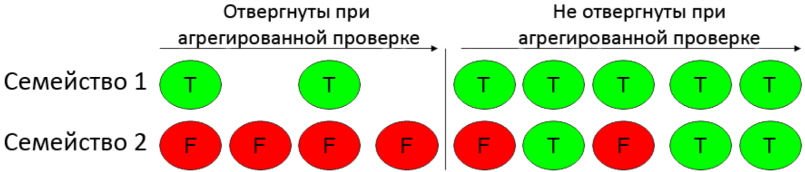
- проверяется множество индивидуальных гипотез, игнорируя семейства;
- обеспечен глобальный контроль;
- агрегированная проверка с глобальным контролем FDR не обеспечивает контроля FDR внутри семейств.

Раздельная проверка:

- гипотезы каждого семейства проверяются независимо;
- обеспечен контроль внутри каждого семейства;
- мощность выше, чем при агрегированной проверке;
- не гарантируется глобальный контроль.

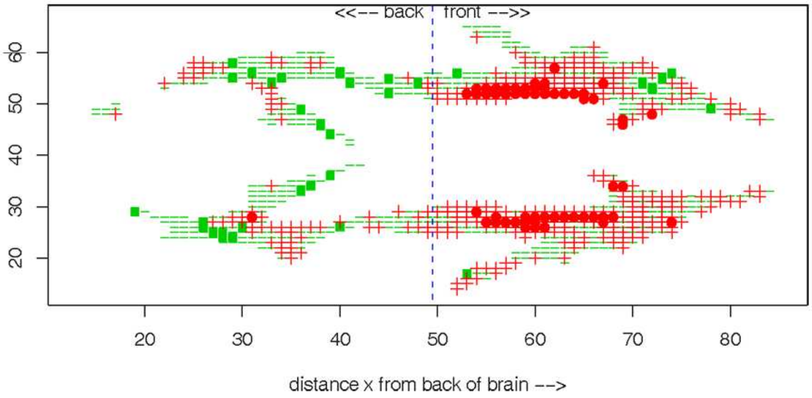
Проблемы агрегированной проверки

Внутри семейств выводы могут быть искажены в обоих направлениях (в каких-то семействах получаем высокий уровень FDR, в каких-то — низкую мощность.



Проблемы агрегированной проверки

Внутри семейств выводы могут быть искажены в обоих направлениях (в каких-то семействах получаем высокий уровень FDR, в каких-то — низкую мощность.



Случай однородных семейств

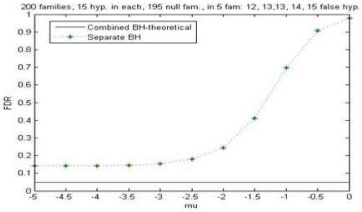
Если распределение статистик $F(t) = \frac{m_0}{m} F_0(t) + \frac{m-m_0}{m} F_1(t)$ во всех семействах одинаково, то:

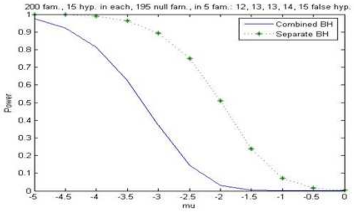
- агрегированная проверка $\approx$ раздельная проверка;
- контроль FDR обеспечен глобально и внутри каждого семейства;
- улучшать мощность некуда.

Случай неоднородных семейств

Если для разных семейств либо $m_{oi} \approx 0$, либо $m_{oi} \approx 1$, то:

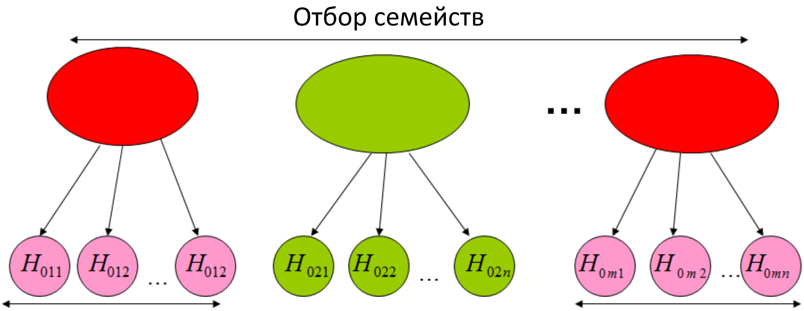
- ни агрегированная, ни раздельная проверка методом Бенджамини-Хохберга не гарантирует контроль над FDR одновременно глобально и внутри каждого семейства;
- при большой размерности раздельная проверка будет иметь значительно большую мощность.





Идея

- На первом этапе отобрать подающие надежды семейства (в которых есть признаки содержания ложных нулевых гипотез).
- На втором этапе сделать раздельную проверку гипотез в отобранных семействах.



Раздельная проверка	Проверка с отбором
В *каждом* семействе $i = 1, \ldots, m$ применяется процедура, обеспечивающая контроль $\mathbb{E}\left(C\right)$	В *каждом отобранном* семействе $i \in S$ применяется процедура, обеспечивающая контроль $\mathbb{E}\left(C\right)$
$\mathbb{E}\left(C_i\right) \leq q, \quad i = 1, \ldots, m$	$\mathbb{E}\left(C_i \,\middle\vert i \in S\right) \leq q, \quad i = 1, \ldots, m$
Контроль в среднем по всем семействам: $$\mathbb{E}\left(\frac{\sum_{i=1}^m C_i}{m}\right) \leq q$$	Контроль в среднем по отобранным семействам: $$\mathbb{E}\left(\frac{\sum_{i \in S} C_i}{\max\left(\lvert S\rvert, 1\right)}\right) \leq q$$

Контроль в среднем по отобранным семействам

$$\mathbb{E}\left(C_S\right) = \mathbb{E}\left(\frac{\sum_{i \in S} C_i}{\max\left(|S|, 1\right)}\right)$$

$\mathbb{E}\left(C\right)$	$\mathbb{E}\left(C_S\right)$	
$FWER$ $P\left(V \geq 1\right)$ $\mathbb{E}\left(I_{\{V \geq 1\}}\right)$	$\mathbb{E}\left(\frac{\sum_{i \in S} I_{\{V \geq 1\}}}{\max\left(\lvert S\rvert, 1\right)}\right)$	$\mathbb{E}\left(\frac{\sharp \text{ семейств с } V \geq 1}{\sharp \text{ отобранных семейств}}\right)$
$FDR = \mathbb{E}\left(FDP\right)$	$\mathbb{E}\left(\frac{\sum_{i \in S} FDP_i}{\max\left(\lvert S\rvert, 1\right)}\right)$	Ожидаемое среднее значение FDP по отобранным семействам

Проблема

Проверка с отбором не обеспечивает контроль в среднем:

- рассмотрим m семейств по n верных гипотез;
- семейство отбирается, если p_i в нём меньше 0.05;
- каждое отобранное семейство проверяется с поправкой Бонферрони на уровне $\alpha = 0.05$;
- $\mathbb{E}(C) = FWER$.

m	n	$\mathbb{E}(\sharp$ отобранных$/m)$	$\mathbb{E}(C_S)$
20	100	0.99	0.049
100	20	0.64	0.076
100	10	0.40	0.122
100	2	0.10	0.506

Исправленная процедура проверки с отбором

- На первом этапе к множеству достигаемых уровней значимости P применяется правило отбора S; $S(P)$ — множество отобранных семейств.
- На втором этапе в каждом отобранном семействе сделать раздельную проверку, применяя $\mathbb{E}(C)$-контролирующую процедуру на уровне $q|S(P)|/m$.

Контроль на уровне семейств	Нет
Контроль внутри каждого семейства	Да
Контроль в среднем по отобранным семействам	Да
Глобальный контроль	Нет

Исправленная процедура проверки с отбором

Требования.

- Правило отбора семейств должно быть **простым**: для каждого отбираемого семейства i при фиксированных достигаемых уровнях значимости всех гипотез все семейства и любом распределении достигаемых уровней значимости внутри семейства, при котором $i \in S(P)$, число отобранных семейств не меняется.
- Ограничения на структуру зависимости достигаемых уровней значимости.

Иерархическая процедура проверки гипотез

Чтобы добавить контроль на уровне семейств, отбор нужно тоже проводить в соответствии с какой-либо процедурой множественной проверки гипотез.

- Каждому семейству поставить в соответствие его полную нулевую гипотезу.
- Скомбинировать пи-величины внутри каждого семейства, получить пи-величины для проверки полных нулевых гипотез.
- Выбрать и применить процедуру множественной проверки, контролирующую желаемую меру ошибки с учётом зависимости между пи-величинами полных нулевых гипотез.
- Отобрать семейства, для которых полная нулевая гипотеза была отвергнута.
- Сделать раздельную проверку, применяя $\mathbb{E}(C)$-контролирующую процедуру на уровне $q|S(P)|/m$.

Иерархическая процедура проверки гипотез

Контроль на уровне семейств	Да
Контроль внутри каждого семейства	Да
Контроль в среднем по отобранным семействам	Да
Глобальный контроль	Нет

Все одношаговые, нисходящие и восходящие методы контроля подходящих мер ошибок — простые.

1. Раздельная проверка гипотез на семействах не обеспечивает глобальный контроль над мерой ошибки.
2. При раздельной проверке гипотез только на предварительно отобранных семействах, не обеспечивается даже контроль в среднем.
3. Необходимо самостоятельно выбирать, какие виды контроля необходимы для интерпретации результатов.
4. Иерархическая процедура проверки может позволить увеличить мощность.