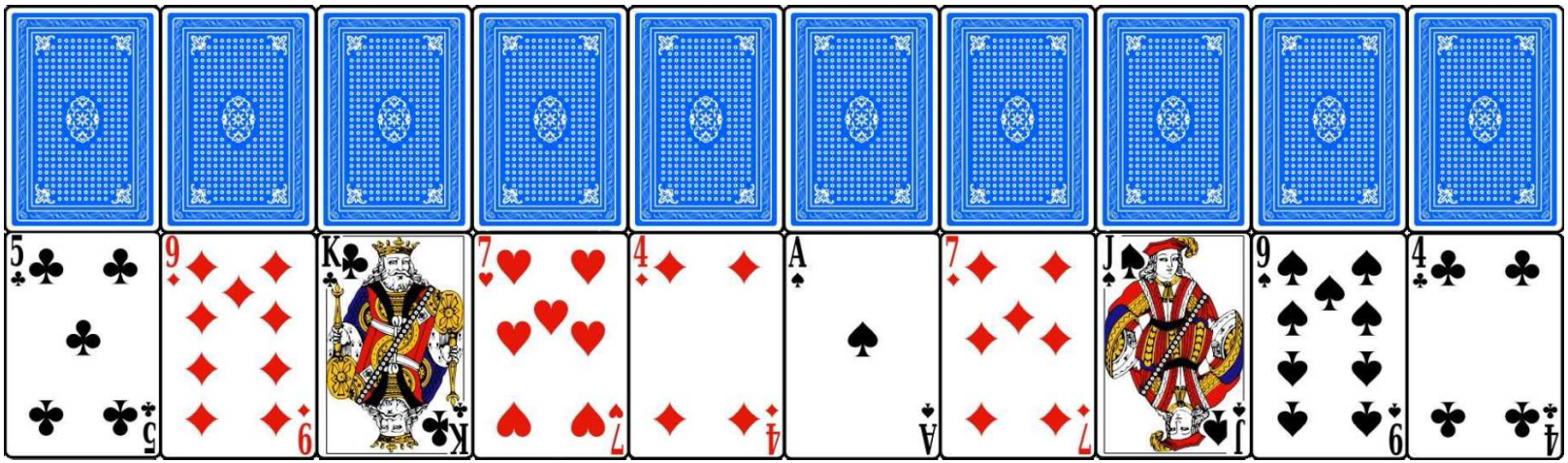
Прикладная статистика 5. Множественная проверка гипотез.

4 октября 2013 г.

Поиск экстрасенсов

Joseph Rhine, 1950: исследования возможности экстрасенсорного восприятия. Первый этап — поиск экстрасенсов.
Испытуемому предлагается угадать цвет 10 карт.



H_0: испытуемый выбирает ответ наугад.
H_1: испытуемый может предсказывать цвета карт.
Статистика t — число карт, цвета которых угаданы.

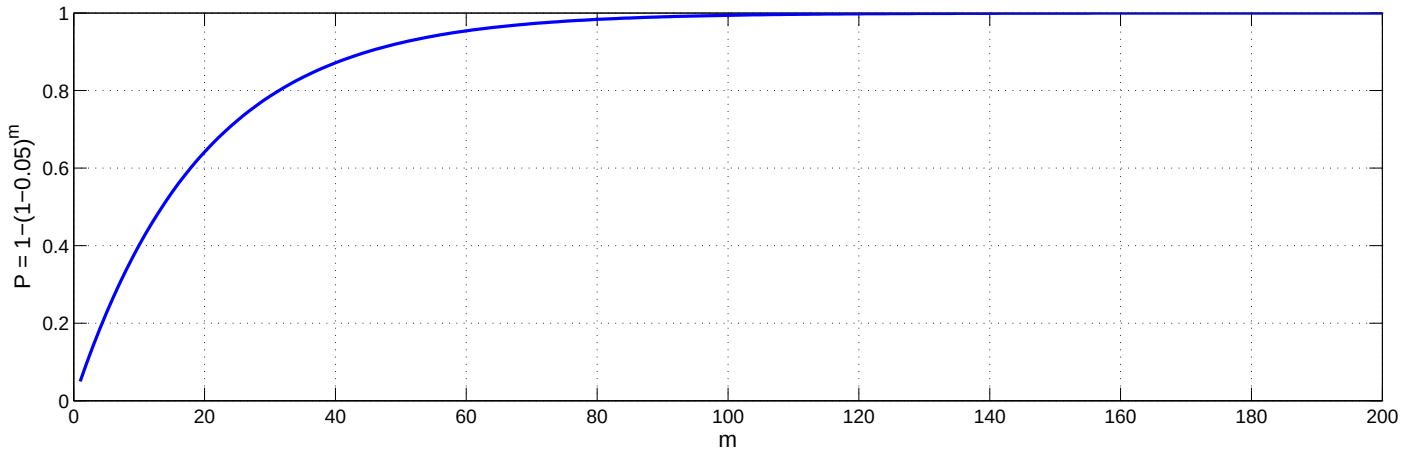
$$P\left(t \geq 9 \,|H_0\right) = 11 \cdot \frac{1}{2}^{10} = 0.0107421875,$$

т. е. при $t = 9$ получаем достигаемый уровень значимости $p \approx 0.01$ — можно отклонять H_0.

Поиск экстрасенсов

Процедуру отбора прошли 1000 человек.
Девять из них угадали цвета 9 из 10 карт, двое — цвета всех 10 карт. Ни один 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

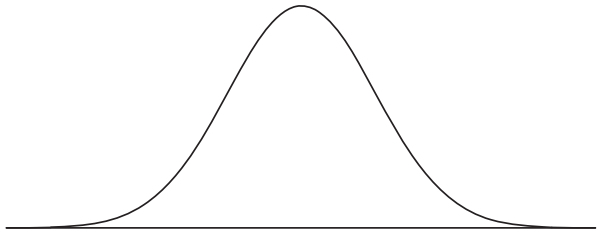
выборка: $X^n = (X_1, \ldots, X_n) \sim P \in \Omega;$

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

альтернатива: $H_1 : P \notin \omega;$

статистика: $T(X^n), \ \ T(X^n) \sim F(x)$ при $P \in \omega;$

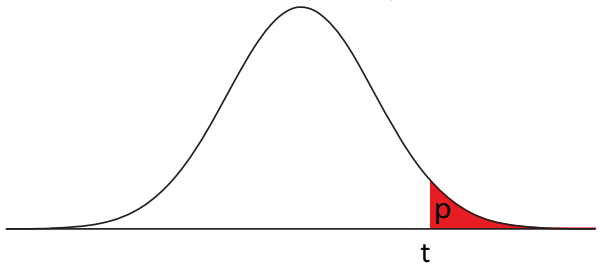
$T(X^n) \not\sim F(x)$ при $P \notin \omega;$



реализация выборки: $x^n = (x_1, \ldots, x_n);$

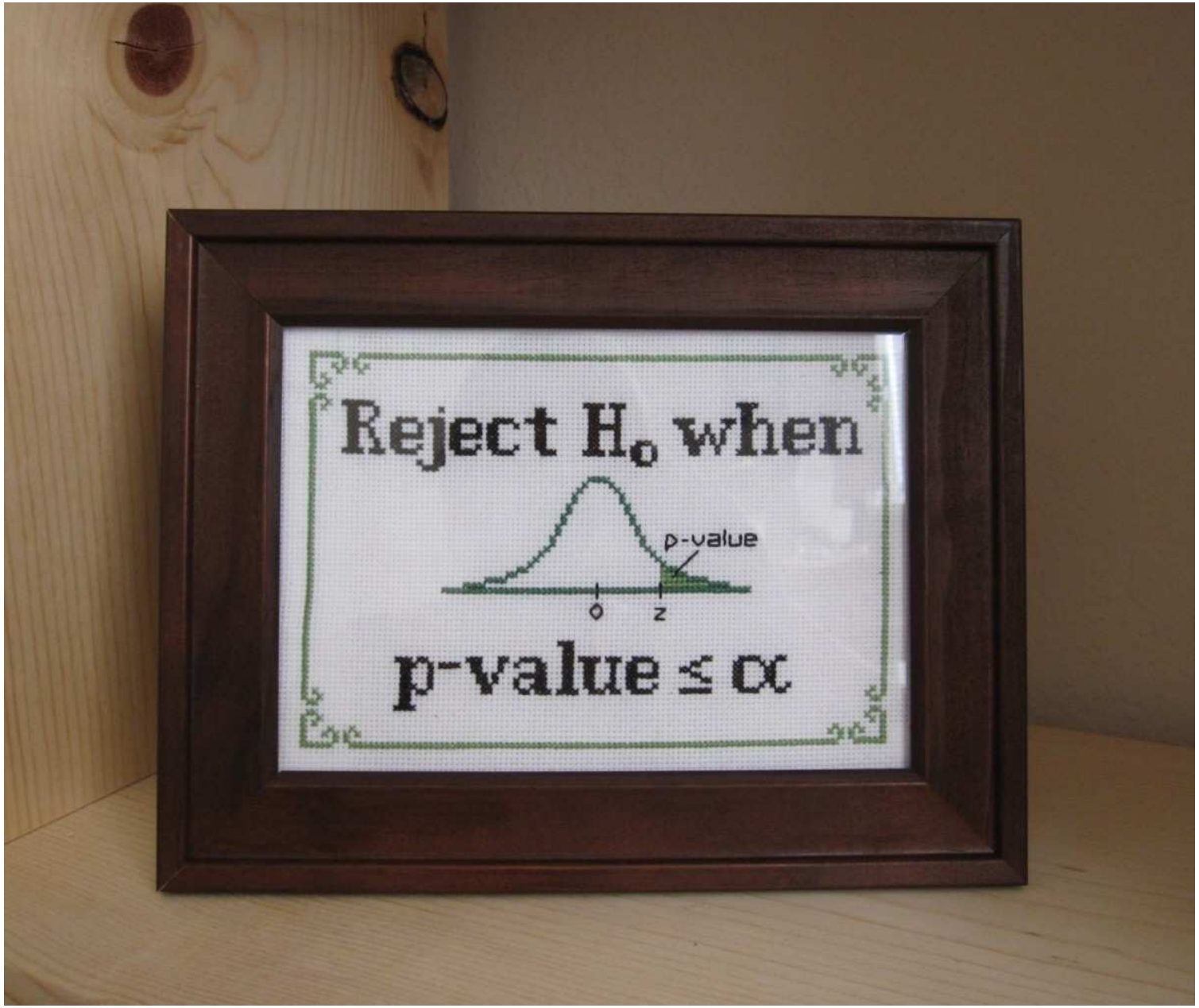
реализация статистики: $t = T(x^n);$

достигаемый уровень значимости: $p(x^n)$ — вероятность при H_0 получить $T(X^n) = t$ или ещё более экстремальное;



Гипотеза отвергается при $p(x^n) \leq \alpha, \ \ \alpha$ — уровень значимости.

Правило проверки гипотезы



Несимметричность задачи проверки гипотезы

	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 малой величиной:

$$p\left(x^n\right) = P\left(T\left(X^n\right) \leq t \,|H_0\right) = P\left(p\left(x^n\right) \leq \alpha \,|H_0\right) \leq \alpha.$$

Вероятность ошибки второго рода минимизируется путём выбора достаточно мощного критерия.

Математическая постановка

данные: $\mathbf{X} = \{X_1^{n_1}, \ldots, X_m^{n_m}\}, \quad X_i^{n_i} \sim P_i \in \Omega;$

нулевые гипотезы: $H_i \colon P_i \in \omega_i, \quad \omega_i \in \Omega;$

альтернативы: $H_i' \colon P_i \notin \omega_i;$

статистики: $T_i = T\left(X_i^{n_i}\right)$ проверяет H_i против $H_i';$

реализации статистик : $t_i = T\left(x_i^{n_i}\right);$

достигаемые уровни значимости: $p_i = p\left(x_i^{n_i}\right), \quad i = 1, \ldots, m;$

$\mathbf{M} = \{1, 2, \ldots, m\};$

$\mathbf{M_0} = \mathbf{M_0}\left(P\right) = \{i \colon H_i$ верна$\}$ — индексы верных гипотез, $|\mathbf{M_0}| = m_0;$

$\mathbf{R} = \mathbf{R}\left(P, \alpha\right) = \{i \colon H_i$ отвергнута$\}$ — индексы отвергаемых гипотез, $|\mathbf{R}| = R;$

$V = |\mathbf{M_0} \cap \mathbf{R}|$ — число ошибок первого рода.

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

Многомерные обобщения ошибки первого рода

Групповая вероятность ошибки первого рода (familywise error rate):

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

Контроль над групповой вероятностью ошибки на уровне α означает

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

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

Поправка Бонферрони

Метод Бонферрони:

$$\alpha_1 = \ldots = \alpha_m = \alpha/m.$$

Теорема

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

Доказательство.

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

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

Поправка Бонферрони

Альтернативный вид — переход к модифицированным достигаемым уровням значимости:

$$\tilde{p}_i = \min\left(1, mp_i\right).$$

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, $X_1^n \sim N(\mu_1, 1)$, $X_2^n \sim N(\mu_1, 1)$, $H_0 \colon \mu_1 = \mu_2$, $H_1 \colon \mu_1 \neq \mu_2$, $\mu_1 - \mu_2 = 1$:

m	n	Мощность
1	23	0.9
10	23	0.67
100	23	0.37
1000	23	0.16
1000	62	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.

Модельный эксперимент

$n = 20, \quad m = 200, \quad m_0 = 150;$
$X_{ij} \sim N(0,1), \ i = 1, \ldots, m_0, \ j = 1, \ldots, n;$
$X_{ij} \sim N(1,1), \ i = m_0 + 1, \ldots, m, \ j = 1, \ldots, n;$
$H_i : \mathbb{E}X_{ij} = 0, \quad H_i' : \mathbb{E}X_{ij} \neq 0;$
для проверки используем одновыборочный критерий Стьюдента.

Без поправок:

	Верных H_i	Неверных H_i	Всего
Принятых H_i	142	0	142
Отвергнутых H_i	8	50	58
Всего	150	50	200

Бонферрони:

	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	27	177
Отвергнутых H_i	0	23	23
Всего	150	50	200

Нисходящие методы множественной проверки гипотез

Составим вариационный ряд достигаемых уровней значимости:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)},$$

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

1. Если $p_{(1)} \geq \alpha_1$, принять все нулевые гипотезы $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ и остановиться; иначе отвергнуть $H_{(1)}$ и продолжить.
2. Если $p_{(2)} \geq \alpha_2$, принять все нулевые гипотезы $H_{(2)}, H_{(3)}, \ldots, H_{(m)}$ и остановиться; иначе отвергнуть $H_{(2)}$ и продолжить.
3. ...

Каждый достигаемый уровень значимости $p_{(i)}$ сравнивается со своим уровнем значимости α_i.

Метод Холма

Метод Холма — нисходящая процедура со следующими уровнями значимости:

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

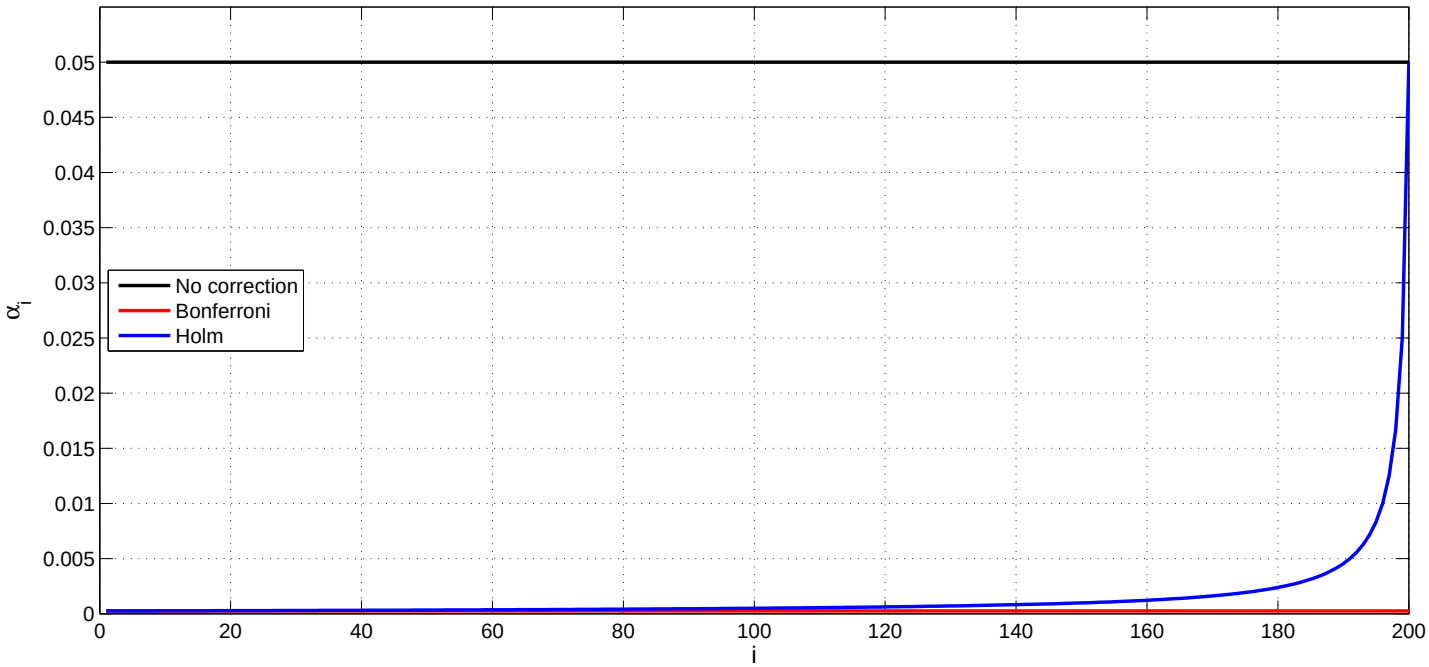
Метод обеспечивает контроль над $FWER$ на уровне α при любых p_i и T_i.

Альтернативный вид — переход к модифицированным достигаемым уровням значимости:

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

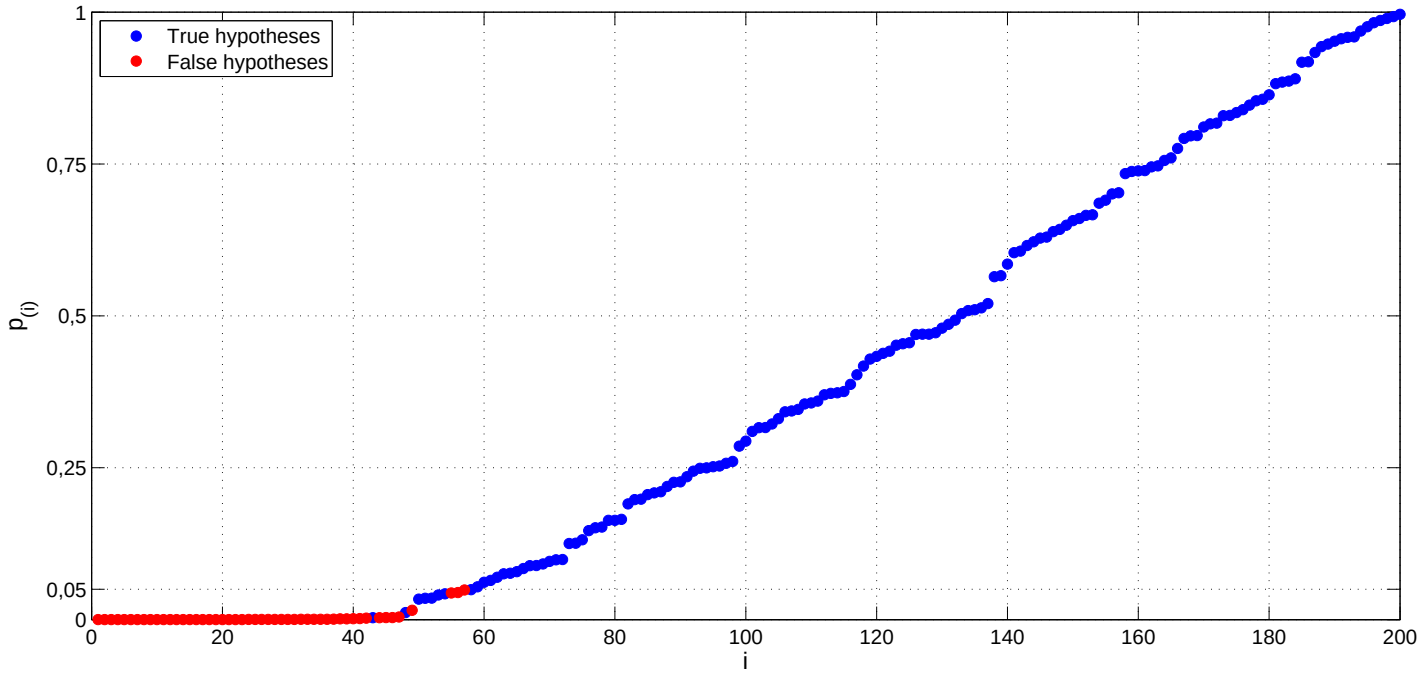
Метод Холма

Метод Холма равномерно мощнее поправки Бонферрони, поскольку все его уровни значимости α_i не меньше:



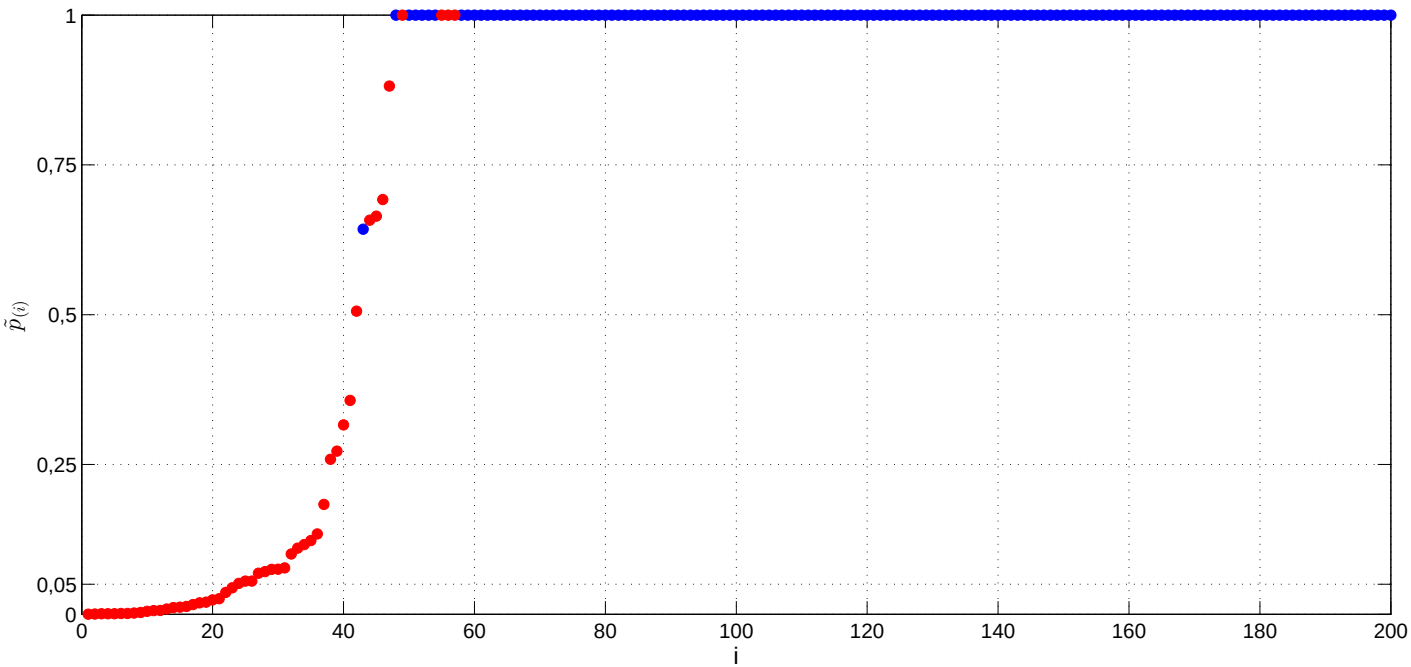
Модельный эксперимент

Отсортированные достигаемые уровни значимости:



Модельный эксперимент

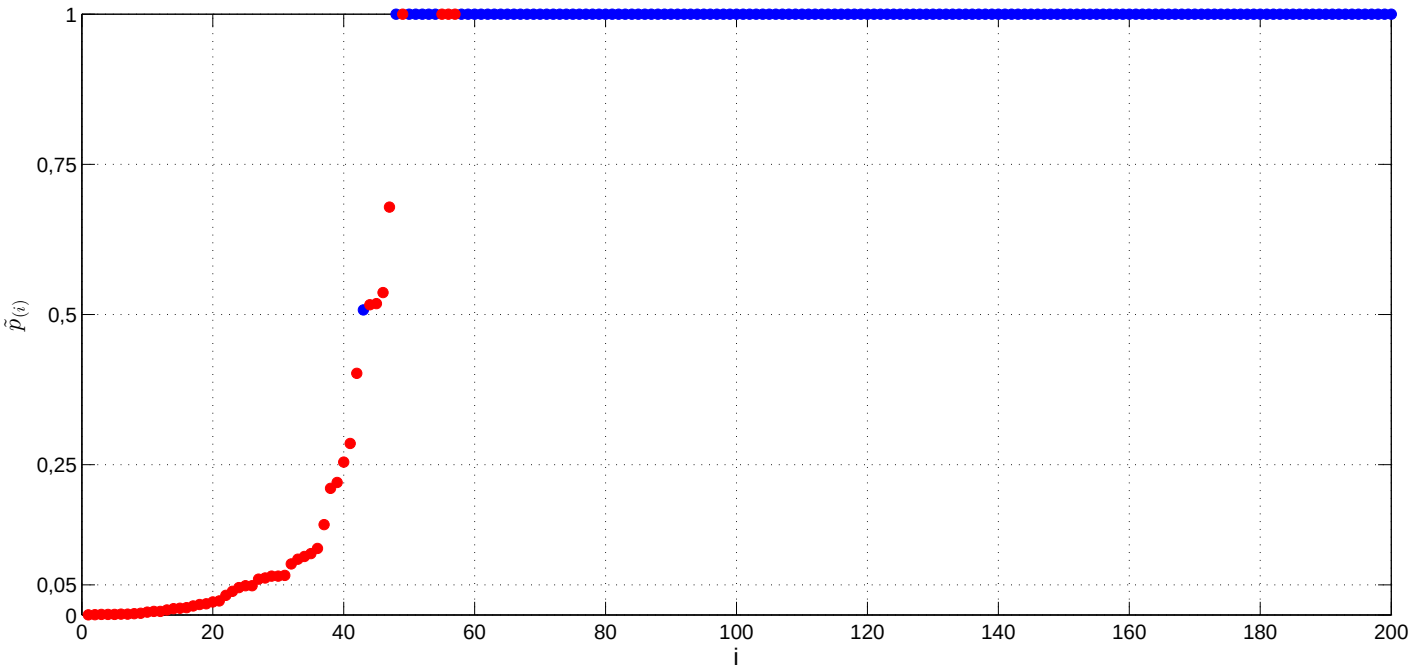
Модифицированные достигаемые уровни значимости, метод Бонферрони:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	27	177
Отвергнутых H_i	0	23	23
Всего	150	50	200

Модельный эксперимент

Модифицированные достигаемые уровни значимости, метод Холма:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	24	174
Отвергнутых H_i	0	26	26
Всего	150	50	200

Идеи для дальнейших улучшений

- Дополнительно оценить m_0.
- Сделать дополнительные предположения:
  - о характере зависимости между статистиками;
  - о совместном распределении статистик.
- Учесть зависимость между статистиками с помощью перестановочных методов.

Предварительное оценивание m_0

Из доказательства теоремы 1 следует, что метод Бонферрони контролирует $FWER$ на уровне $\frac{m_0}{m}\alpha$.

Примеры методов оценки m_0:

- метод Стори:

$$\hat{m}_0 = 2\sum_{i=1}^{m}[p_i \geq 0.5];$$

- метод наименьшего наклона Бенджамини-Хохберга:

$$\hat{m}_0 = \min\left(m, \frac{1}{S_{i_0}} + 1\right),$$

$$S_i = \frac{1 - p_{(i)}}{m - i + 1}, i = 1, \ldots, m,$$

$$i_0 = \min\{i : S_i < S_{i-1}\}.$$

Как правило, доказательств контроля $FWER$ для процедур с предварительной оценкой m_0 нет, но на практике они часто работают хорошо.

Одношаговый метод Шидака

Метод Шидака:

$$\alpha_1 = \alpha_2 = \ldots = \alpha_m = 1 - (1 - \alpha)^{\frac{1}{m}} .$$

Метод обеспечивает контроль над $FWER$ на уровне α при условии, что статистики T_i **независимы или отрицательно коррелированы**.

Альтернативный вид — переход к модифицированным достигаемым уровням значимости:

$$\tilde{p}_i = 1 - (1 - p_i)^m .$$

Нисходящая модификация

Нисходящий метод Шидака (метод Шидака-Холма) — нисходящая процедура со следующими уровнями значимости:

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

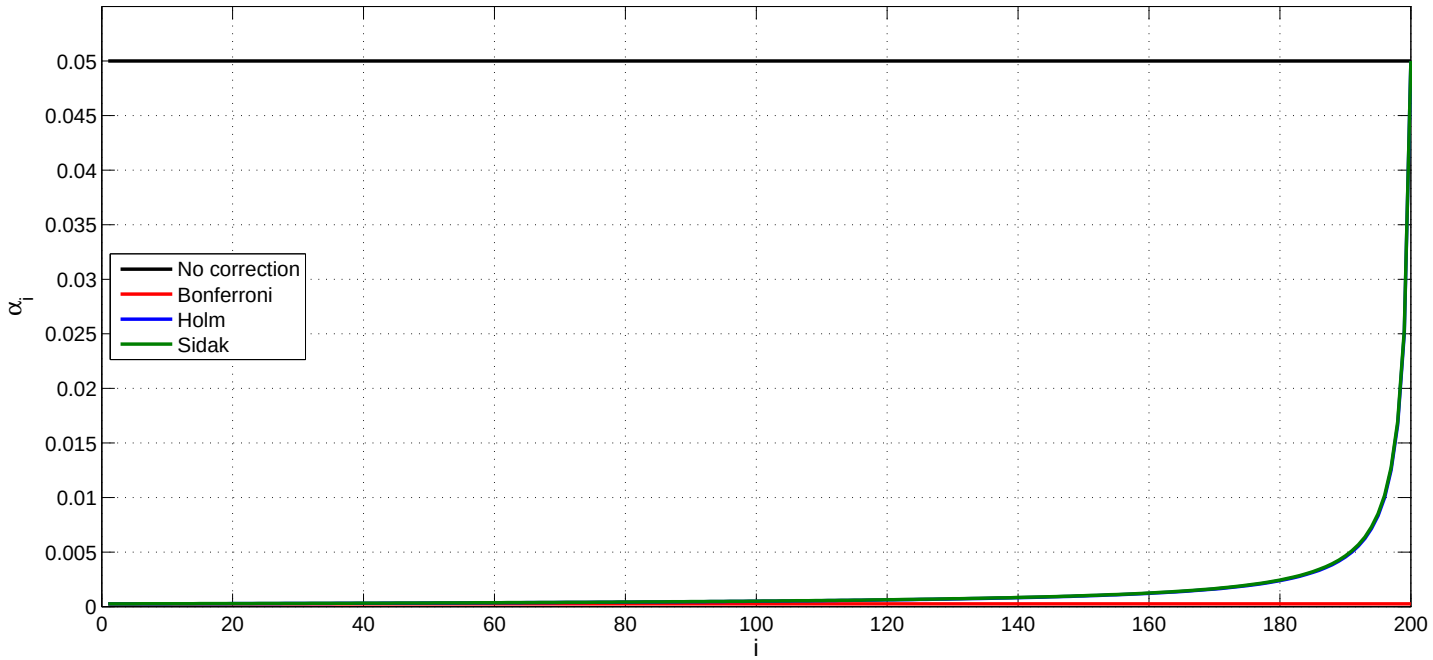
Метод обеспечивает контроль над $FWER$ на уровне α при условии, что статистики T_i **независимы**.

Альтернативный вид — переход к модифицированным достигаемым уровням значимости:

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

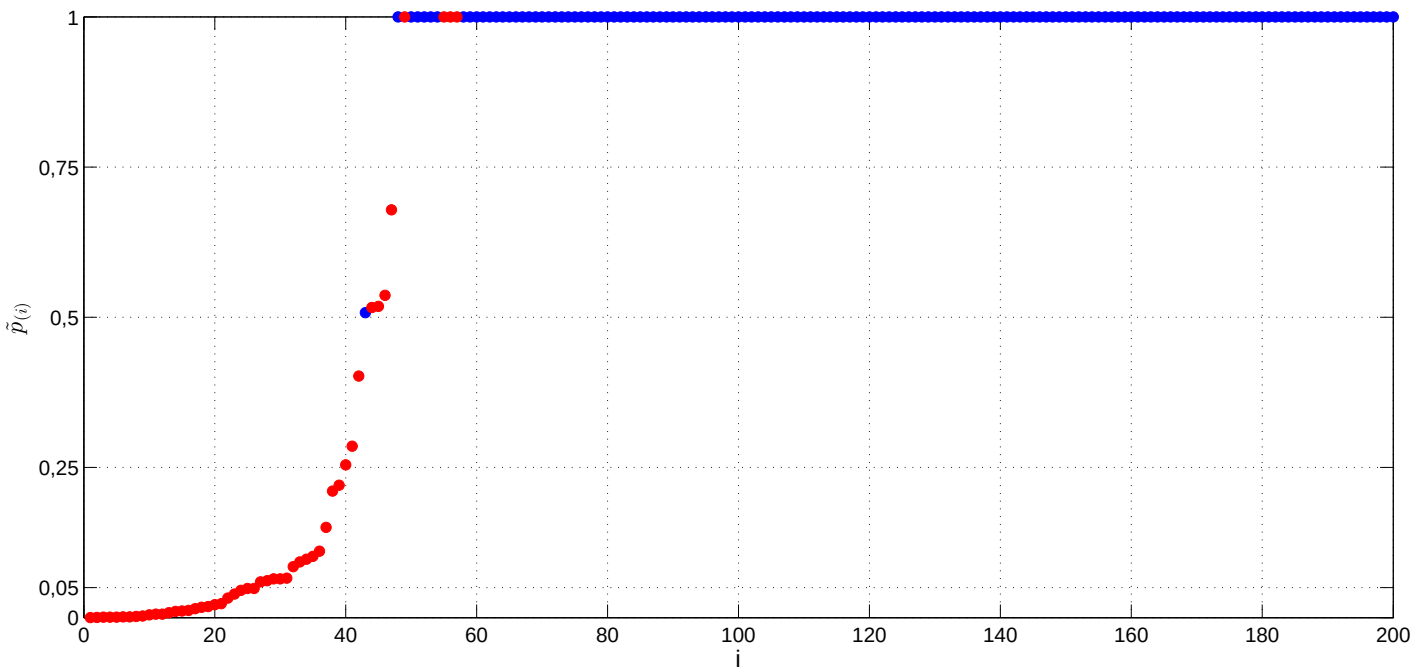
Нисходящая модификация

На практике при достаточно больших m не слишком отличается от метода Холма:



Модельный эксперимент

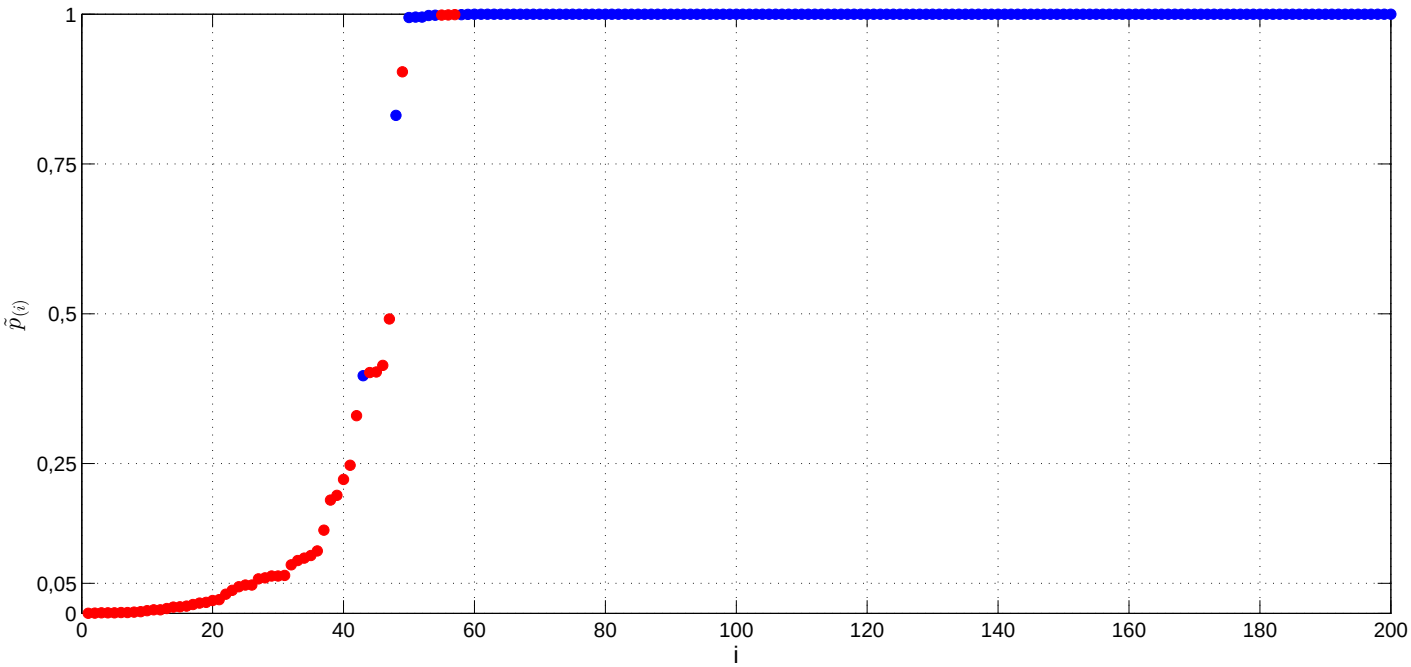
Модифицированные достигаемые уровни значимости, метод Холма:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	24	174
Отвергнутых H_i	0	26	26
Всего	150	50	200

Модельный эксперимент

Модифицированные достигаемые уровни значимости, нисходящий метод Шидака:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	24	174
Отвергнутых H_i	0	26	26
Всего	150	50	200

Зависимость между статистиками

- Не учитывая характер зависимости между статистиками, нельзя построить контролирующую $FWER$ процедуру мощнее, чем метод Холма.
- Если статистики независимы, нельзя построить контролирующую $FWER$ процедуру мощнее, чем метод Шидака-Холма.
- Чем сильнее связь между статистиками, тем меньше нужно модицифировать уровни значимости.

Для построения мощной процедуры множественной проверки гипотез необходимо учесть структуру зависимости статистик.

Параметрические методы

Если совместное нулевое распределение статистик $T_1, \ldots, T_m$ известно, константы α_i могут быть так, что контроль над $FWER$ будет точным ($FWER = \alpha$).

Примеры: HSD Тьюки для попарных сравнений нормально распределённых выборок друг с другом; критерий Даннета для сравнения средних m нормально распределённых выборок со средним контрольной выборки.

Перестановочные методы

Неявно учесть зависимости между статистиками можно при помощи перестановочных методов. Подробнее:

- Westfall, P., Troendle, J. (2008). Multiple testing with minimal assumptions. Biometrical Journal, 50(5), 745–755 и другие работы Westfall, P.
- Bretz, F., Hothorn, T., Westfall, P. (2010). Multiple Comparisons Using R. Boca Raton: Chapman and Hall/CRC. (Раздел 5.1)

Методы обеспечивает контроль над $FWER$ на уровне α при условии выполнения свойства **subset pivotality**, когда нулевое распределение любого подмножества статистик T_i не зависит от того, верну или неверны соответствующие оставшимся статистикам гипотезы:

$$P\left(\bigcap_{i\in\mathbf{M}^*}\{T_i \geq t^*\}\middle|\bigcap_{i\in\mathbf{M}^*} H_i\right) = P\left(\bigcap_{i\in\mathbf{M}^*}\{T_i \geq t^*\}\middle|\bigcap_{i\in\mathbf{M}} H_i\right) \quad \forall\mathbf{M}^*.$$

Многомерные обобщения ошибки первого рода

Ожидаемая доля ложных отклонений гипотез (false discovery rate):

$$FWER = \mathbb{E}\left(\frac{V}{R}\left[R > 0\right]\right).$$

Контроль над ожидаемой долей ложных отклонений на уровне α означает

$$FDR = \mathbb{E}\left(\frac{V}{R}\left[R > 0\right]\right) \leq \alpha \quad \forall P.$$

Для любой процедуры множественной проверки гипотез $FDR \leq FWER$.

Метод Бенджамини-Хохберга

Метод Бенджамини-Хохберга — нисходящая процедура со следующими уровнями значимости:

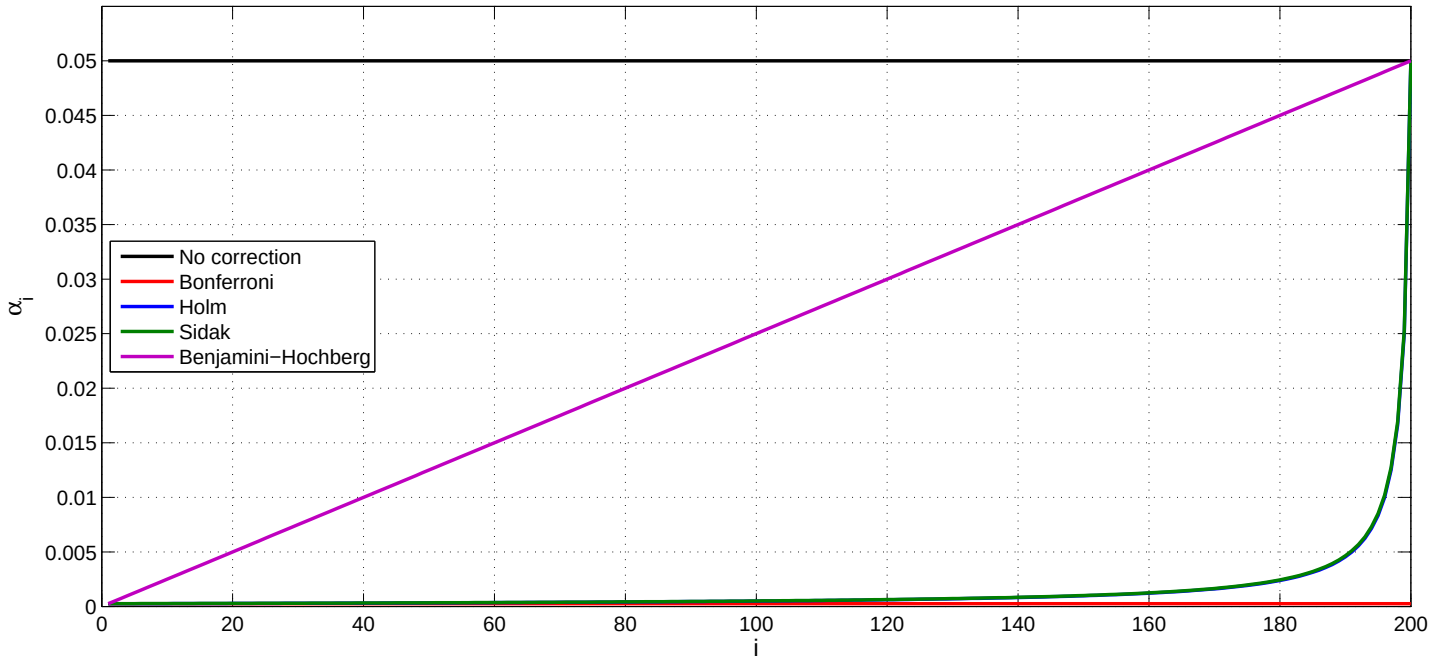
$$\alpha_1 = \frac{\alpha}{m}, \ \ldots, \ \alpha_i = \frac{\alpha i}{m}, \ \ldots, \ \alpha_m = \alpha.$$

Метод обеспечивает контроль над FDR на уровне α при условии, что статистики T_i **независимы** или выполняется свойство PRDS (Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 29(4), 1165–1188.).

Альтернативный вид — переход к модифицированным достигаемым уровням значимости:

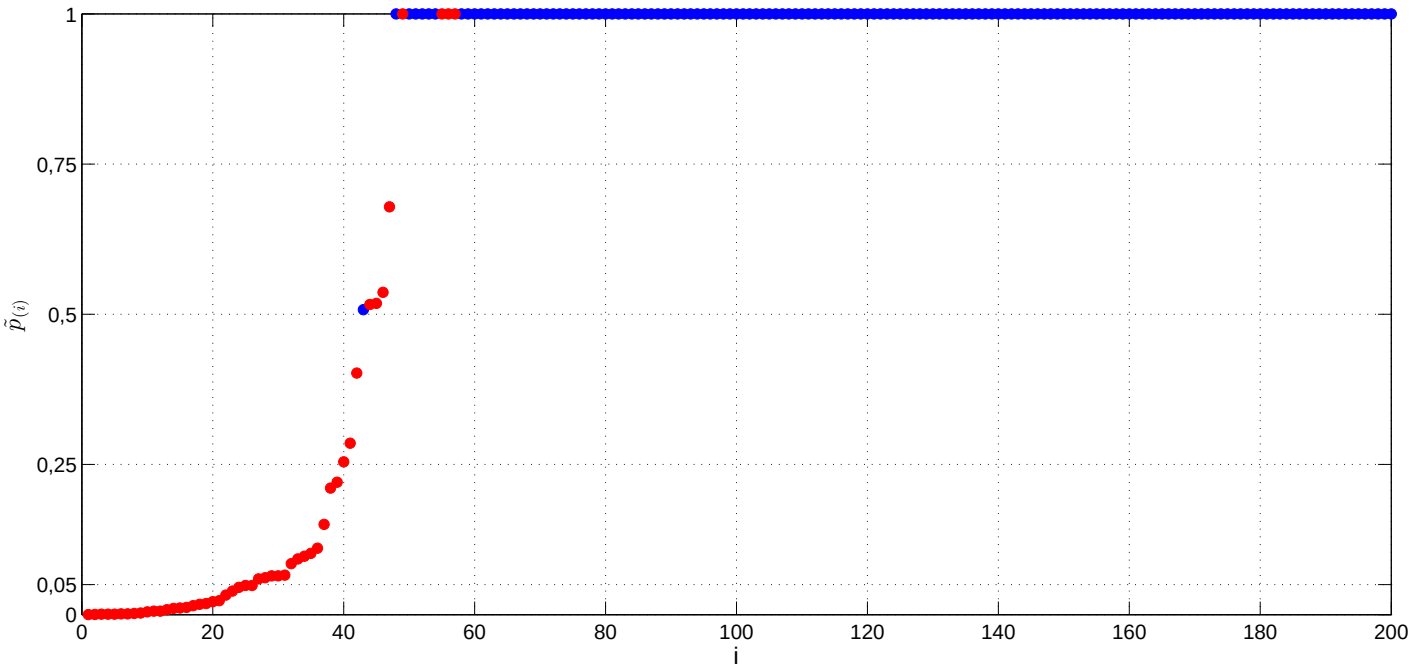
$$\tilde{p}_{(i)} = \min\left(1, \max\left(\frac{mp_{(i)}}{i}, \tilde{p}_{(i-1)}\right)\right).$$

Метод Бенджамини-Хохберга



Модельный эксперимент

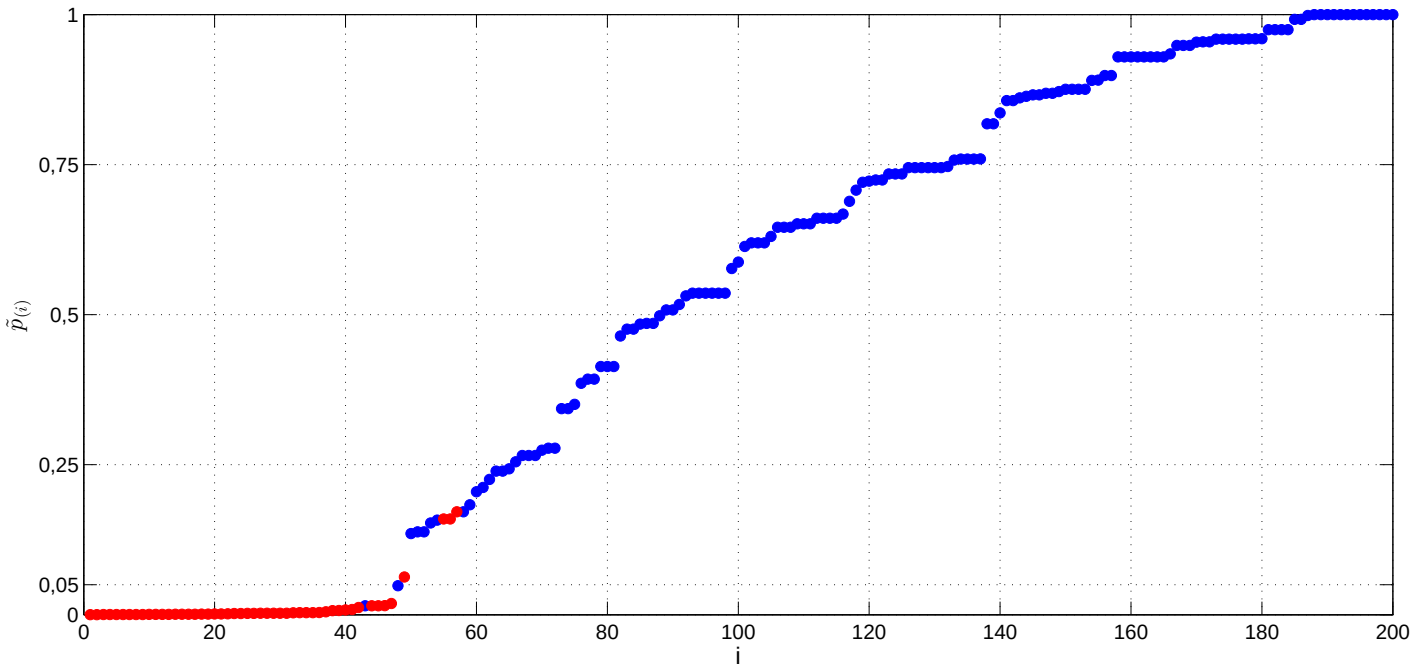
Модифицированные достигаемые уровни значимости, метод Холма:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	24	174
Отвергнутых H_i	0	26	26
Всего	150	50	200

Модельный эксперимент

Модифицированные достигаемые уровни значимости, нисходящий метод Бенджамини-Хохберга:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	148	4	152
Отвергнутых H_i	2	46	48
Всего	150	50	200

Метод Бенджамини-Иекутиели

Метод Бенджамини-Иекутиели — нисходящая процедура со следующими уровнями значимости:

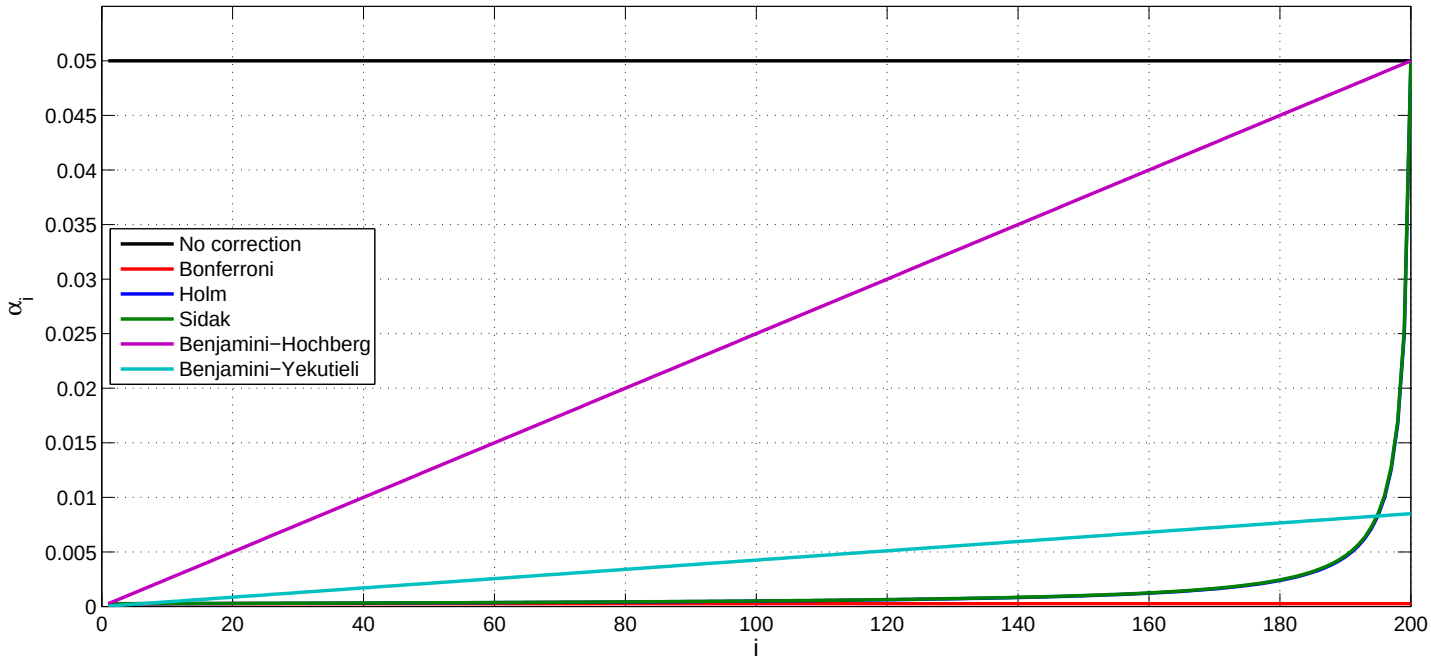
$$\alpha_1 = \frac{\alpha}{m\sum\limits_{i=1}^{m}\frac{1}{i}}, \ \ldots, \ \alpha_i = \frac{\alpha i}{m\sum\limits_{i=1}^{m}\frac{1}{i}}, \ \ldots, \ \alpha_m = \frac{\alpha}{\sum\limits_{i=1}^{m}\frac{1}{i}}.$$

Метод обеспечивает контроль над FDR на уровне $\frac{m_0}{m}\alpha \leq \alpha$ при любых p_i и T_i.

Альтернативный вид — переход к модифицированным достигаемым уровням значимости:

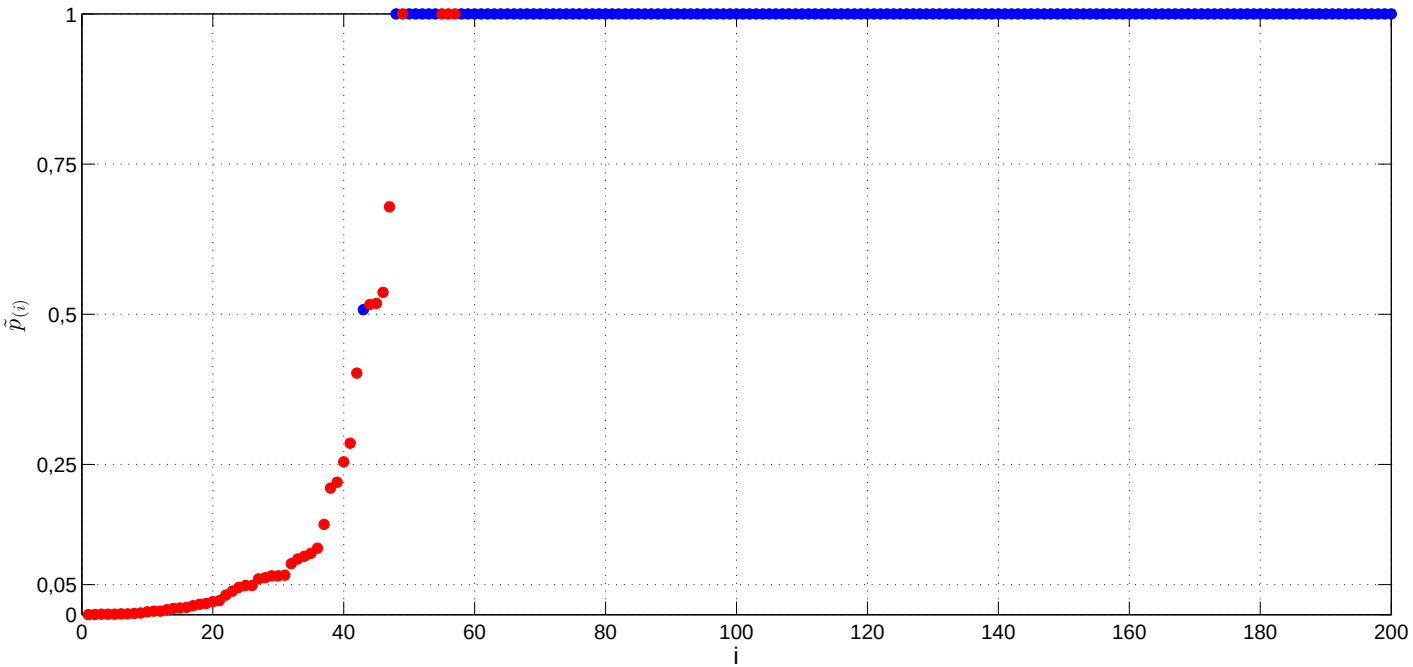
$$\tilde{p}_{(i)} = \min\left(1, \max\left(\frac{mp_{(i)}\sum\limits_{i=1}^{m}\frac{1}{i}}{i}, \tilde{p}_{(i-1)}\right)\right).$$

Метод Бенджамини-Иекутиели



Модельный эксперимент

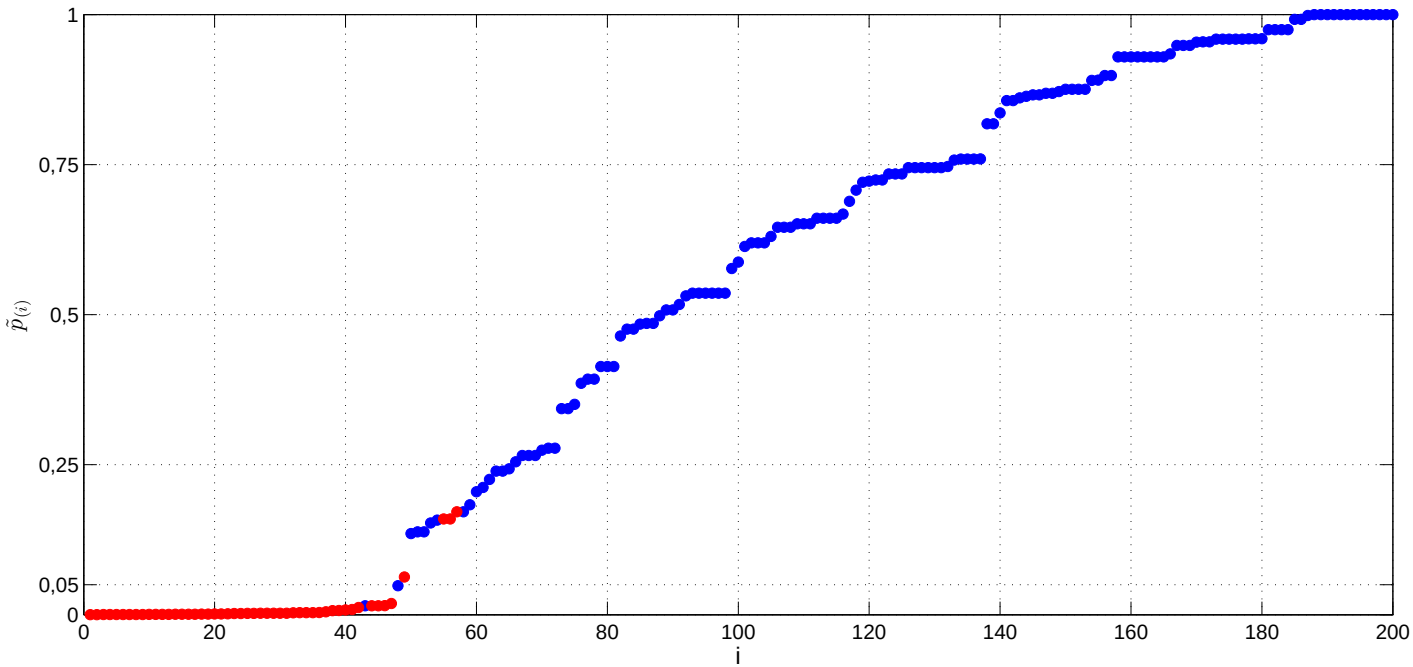
Модифицированные достигаемые уровни значимости, метод Холма:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	24	174
Отвергнутых H_i	0	26	26
Всего	150	50	200

Модельный эксперимент

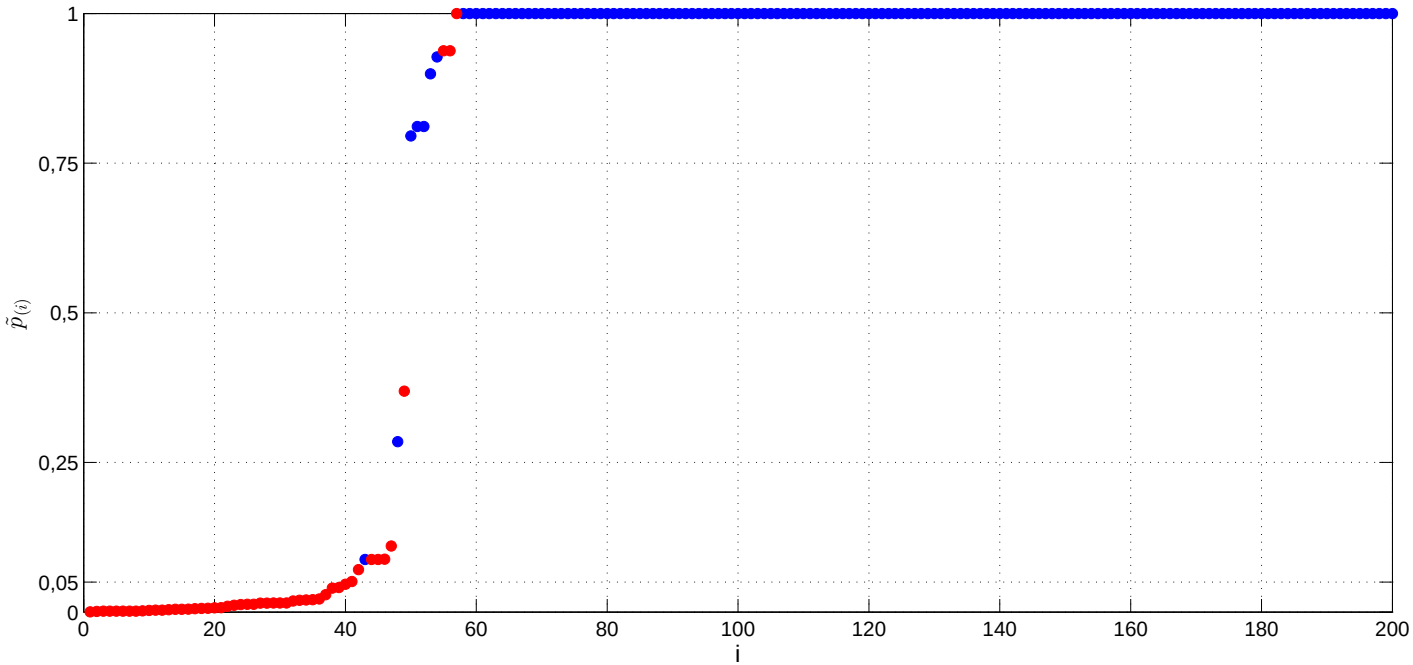
Модифицированные достигаемые уровни значимости, нисходящий метод Бенджамини-Хохберга:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	148	4	152
Отвергнутых H_i	2	46	48
Всего	150	50	200

Модельный эксперимент

Модифицированные достигаемые уровни значимости, нисходящий метод Бенджамини-Иекутиели:



	Верных H_i	Неверных H_i	Всего
Принятых H_i	150	10	160
Отвергнутых H_i	0	40	40
Всего	150	50	200

Мутации

	Контроль (100)	Больные (100)	p
Мутация	1 из 100	8 из 100	0.0349
Фамилия начинается с гласной	36 из 100	40 из 100	0.6622

Бонферрони, Холм: p_1 сравнивается с $\frac{0.05}{2} = 0.025$

Шидак: p_1 сравнивается с $1 - (1 - 0.05)^{\frac{1}{2}} \approx 0.02532$

Сравнение алгоритмов

AUC базового алгоритма C4.5 и трёх его модификаций:

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

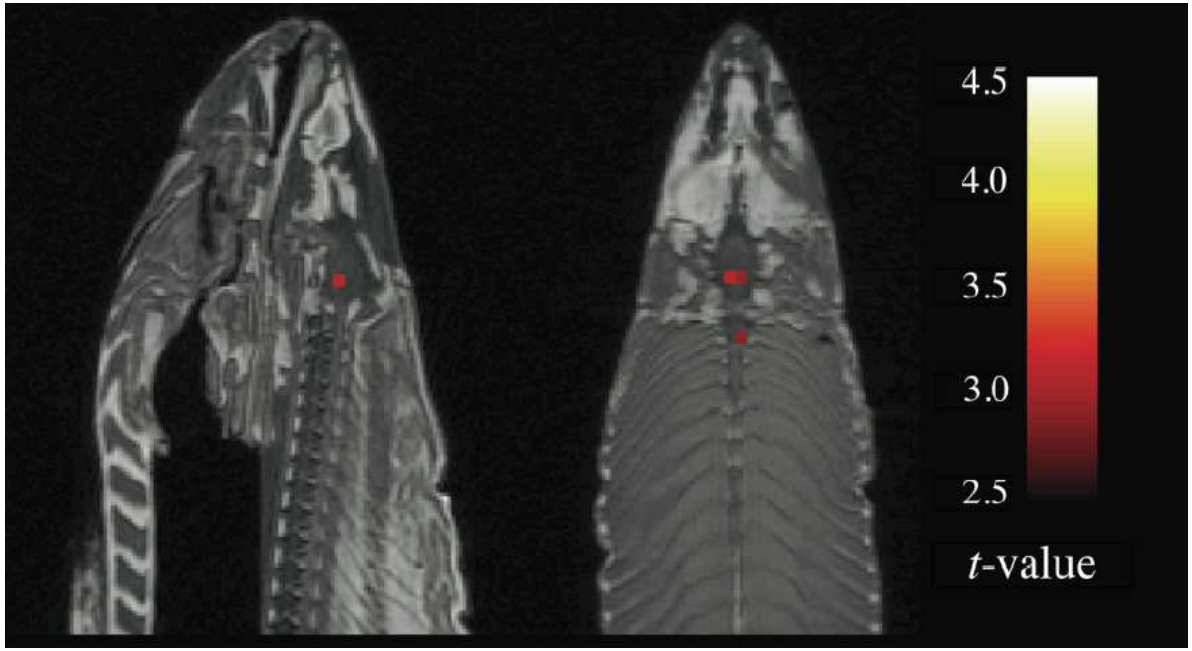
Сравнение алгоритмов

Гипотеза: модификации алгоритмов не влияют на AUC.
Критерий Фридмана: $p = 0.0197$

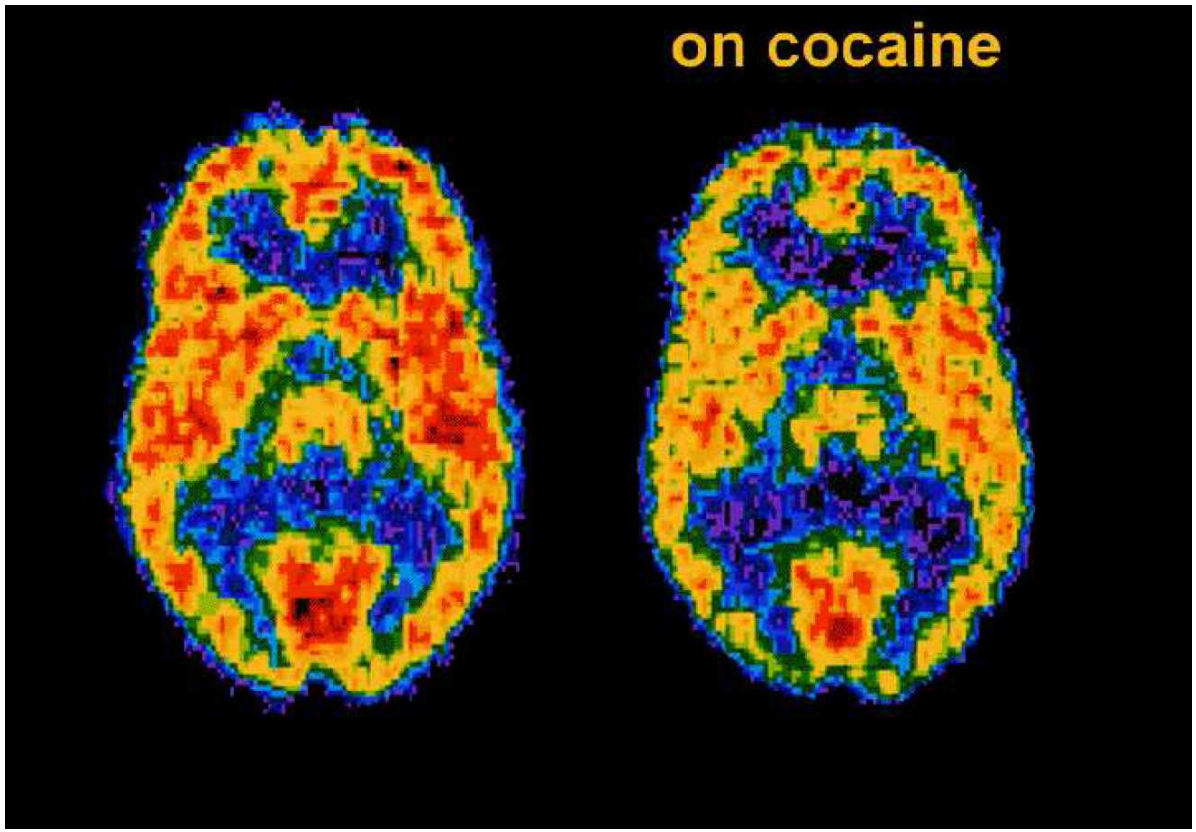
Между какими версиями алгоритмов есть отличия?
Критерий Неменьи для попарного сравнения всех средних: ни одного значимого отличия.

Какие модификации отличаются от базового алгоритма?
Критерий знаковых рангов Уилкоксона + контроль $FWER$ на уровне 0.05 методом Холма: C4.5+m ($p = 0.048$) и C4.5+m+cf ($p = 0.048$).

SPM



SPM



Прикладная статистика
5. Множественная проверка гипотез.

Рябенко Евгений
riabenko.e@gmail.com