Методы современной прикладной статистики
4.2. Множественная проверка гипотез

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Проблема баяниста

5 раз Игорь Владимирович проходил одним подземным переходом, и 5 раз там сидел баянист и не играл. Можно ли отвергнуть гипотезу о том, что баянист играет большую часть времени?

Проблема баяниста

На шестой раз баянист играл. С тех пор он гораздо чаще играет, чем нет. Т.е. была совершена ошибка первого рода. Её можно было избежать, если бы Игорь Владимирович применил метод множественной проверки гипотез.

Постановка задачи

Пусть имеются данные $X = \{X_i^{(j)}\}$, $1 \le i \le n_j$, $1 \le j \le m$.

По ним проверяем гипотезы $H_j : P_j \in \mathcal{P}_j$ против альтернатив $H_j' : P_j \notin \mathcal{P}_j$ с помощью статистик $T_j = T_j(X_1^{(j)}, \ldots, X_{n_j}^{(j)})$.

Пусть $p_j = p_j(T_j)$ – p-значения критериев.

Постановка задачи

Обозначим $M = \{1, \ldots, m\}$, M_0 – индексы верных гипотез, $|M_0| = m_0$, R – число отвергнутых гипотез, V – число ошибок первого рода.

	$\sharp$ верных H_j	$\sharp$ ложных H_j	Всего
$\sharp$ принятых H_j	U	T	m-R
$\sharp$ отвергнутых H_j	V	S	R
Всего	m_0	$m-m_0$	m

Групповая вероятность ошибки I рода (family-wise error rate)

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

Контроль над FWER на уровне α означает, что

$$FWER = P(V > 0) \leq \alpha$$

для всех распределений из верных гипотез H_j, $j \in M_0$.

Пусть $\alpha_1, \ldots, \alpha_m$ – уровни значимости критериев проверки гипотез $H_1, \ldots, H_m$ соответственно. Хотим их выбрать таким образом, чтобы $FWER \leq \alpha$.

Метод Бонферрони

Метод Бонферрони: $\alpha_1 = \ldots = \alpha_m = \frac{\alpha}{m}$.
Действительно,

$$FWER = P(V > 0) = P(\exists j \in M_0 : \ p_j \leq \alpha/m) \leq$$

$$\sum_{j \in M_0} P(p_j \leq \alpha/m) \leq m_0 \cdot \frac{\alpha}{m} \leq \alpha.$$

Главный недостаток метода – резкое уменьшение мощности статистической процедуры при $m \to \infty$.

Метод Шидака

Метод Шидака: $\alpha_1 \dots \alpha_m = 1 - (1 - \alpha)^{1/m}$.

Метод дает $FWER \leq \alpha$ при условии, что статистики T_i независимы или выполнено свойство "положительной зависимости":

$$P(T_1 \leq t_1, \dots, T_m \leq t_m) \geq \prod_{i=1}^{m} P(T_i \leq t_i) \; \forall \vec{t} \in \mathbb{R}^m.$$

Положительную зависимость, в частности, можно установить с помощью FKG-неравенства: если $f(x)$ и $g(x)$ – возрастающие (убывающие) функции, то $Ef(X)g(X) \geq Ef(X)Eg(X)$.

Нисходящие процедуры

Составим вариационный ряд p-значений

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

где $H_{(1)}, \ldots, H_{(m)}$ – соответствующие гипотезы. Процедура выглядит так:

1. Если $p_{(1)} \geq \alpha_1$, то принимаем все гипотезы $H_{(1)}, \ldots, H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(1)}$ и продолжаем;
2. Если $p_{(2)} \geq \alpha_2$, то принимаем все гипотезы $H_{(2)}, \ldots, H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(2)}$ и продолжаем;
3. ...

Нисходящие процедуры

Метод Холма: нисходящая процедура с уровнями значимости

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

Свойства:

1. контролирует FWER на уровне значимости α;
2. равномерно мощнее метода Бонферрони;
3. если характер зависимости между статистиками $\{T_i\}$ неизвестен, то нельзя построить контролирующую FWER на уровне α процедуру мощнее, чем метод Холма.

Нисходящие процедуры

Метод Шидака-Холма: нисходящая процедура с уровнями значимости

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

Свойства:

1. контролирует FWER на уровне значимости α, если статистики $\{T_i\}$ **независимы в совокупности**;
2. если статистики $\{T_i\}$ независимы в совокупности, то нельзя построить контролирующую FWER на уровне α процедуру мощнее, чем метод Шидака-Холма;
3. при больших m мало отличается от метода Холма.

Ожидаемая доля ложных отклонений гипотез (false discovery rate)

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

Контроль над FDR на уровне значимости α означает, что $FDR \leq \alpha$ для всех распределений из верных гипотез H_j, $j \in M_0$.

Хотя $FDR = \mathrm{E}\left(\frac{V}{\max(R,1)}\right) \leq \mathrm{E}I(V > 0) = P(V > 0) = FWER$, но в рамках процедур, контролирующих FDR на уровне α, случается больше ошибок первого рода.

Восходящие процедуры

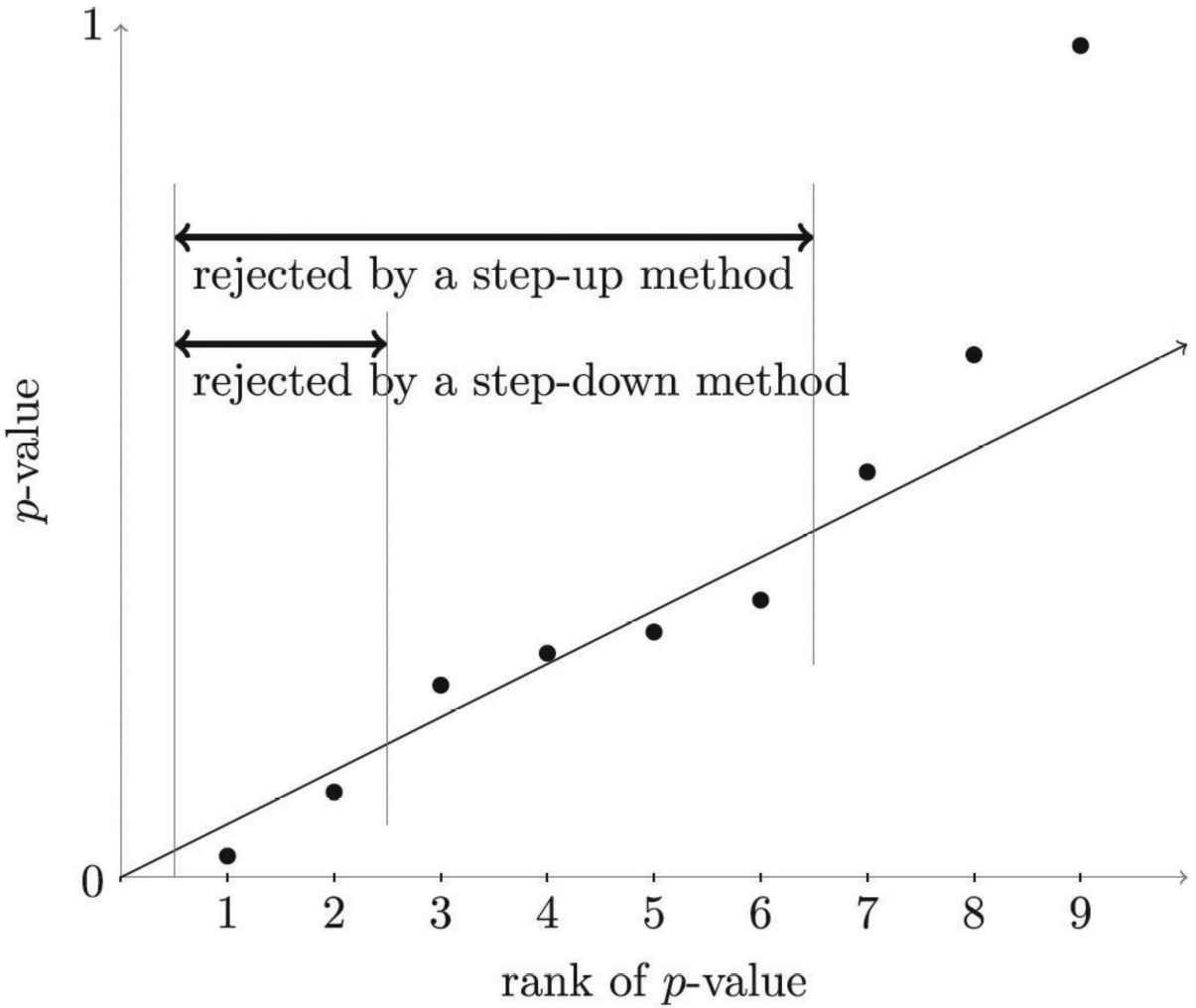
Пусть, как и ранее, $p_{(1)} \leq \ldots \leq p_{(m)}$ – вариационный ряд полученных p-значений, а $H_{(1)}, \ldots, H_{(m)}$ – соответствующие им гипотезы.

Процедура выглядит так:

1. Если $p_{(m)} < \alpha_m$, то отвергаем все гипотезы $H_{(1)}, \ldots, H_{(m)}$ и останавливаемся, иначе принимаем $H_{(m)}$ и продолжаем;
2. Если $p_{(m-1)} < \alpha_{m-1}$, то отвергаем все гипотезы $H_{(1)}, \ldots, H_{(m-1)}$ и останавливаемся, иначе принимаем $H_{(2)}$ и продолжаем;
3. ...

Восходящие процедуры

Очевидно, что восходящая процедура отвергает не меньше гипотез, чем нисходящая с теми же $\{p_i\}$ и $\{\alpha_i\}$.



Восходящие процедуры

Метод Бенджамини-Хохберга: восходящая процедура,

$$\text{для которой } \alpha_i = \alpha \cdot \frac{i}{m},\ i = 1, \ldots, m.$$

Метод контролирует FDR на уровне α, если $\{T_i\}$ независимы или выполнено свойство PDRS:

$$P(X \in D | T_i = x) \text{ не убывает по } x\ \forall i \in M_0,$$

где D – возрастающее множество, т.е. если $\vec{y} \in D$ и $\vec{z} \geq \vec{y}$, то $\vec{z} \in D$.

В частности, свойство PDRS выполнено, если $X \sim N(a, \Sigma)$, где все элементы ковариационной матрицы Σ неотрицательны.

Восходящие процедуры

Метод Бенджамини-Иекутиели: восходящая процедура с уровнями значимости

$$\alpha_i = \alpha \cdot \frac{i}{m} \left(\sum_{j=1}^{m} \frac{1}{j} \right)^{-1}, \ i = 1, \ldots, m.$$

Метод контролирует FDR на уровне $\frac{m_0}{m}\alpha \leq \alpha$ для любых T_i.

При отсутствии информации о зависимости между статистиками T_i метод не улучшаем.

Сравнение методов

Модельный эксперимент: пусть имеется $m = 200$ выборок размера $n = 20$ из нормального распределения $N(a, 1)$, причем первые $m_0 = 150$ выборок сделаны из $N(0, 1)$, а последние 50 – из $N(1, 1)$.

Проверим гипотезы $H_i : a = 0$ против альтернатив $H_i' : a \neq 0$, $i = 1, \ldots, m$, с помощью t-критерия Стьюдента, полагая $\alpha = 0.1$:

$$\text{если } \left| \sqrt{n} \frac{\overline{X} - a}{\sqrt{s^2}} \right| > t_{1-\alpha/2}, \text{ то отвергнуть } H_i,$$

где $t_{1-\alpha/2}$ – $(1 - \alpha/2)$-квантиль распределения Стьюдента $St(n - 1)$.

Сравнение методов

Матрицы ошибок модельного эксперимента:

Без поправки

	True	*False*
Accepted	142	0
Rejected	8	50

Бонферрони

	True	*False*
Accepted	150	27
Rejected	0	23

Шидак-Холм

	True	*False*
Accepted	150	24
Rejected	0	26

Бенджамини-Хохберг

	True	*False*
Accepted	148	4
Rejected	2	46

Выводы

1. Если мы проверяем цепочку гипотез о каком-то одном наборе данных, то при отклонении одной из гипотез в рамках процедуры множественной проверки стоит остановиться и отклонить все остальные. Например, в модельном эксперименте стоит отвергнуть гипотезу о том, что данные выбраны из стандартного нормального распределения.

2. Если мы последовательно проверяем гипотезы о различных наборах данных, то процедура множественной проверки гипотез также необходима, поскольку если поправки не делать, то вероятность того, что произойдет ошибка первого рода, будет расти с количеством проверяемых гипотез.

Finita!