МСПС (Множественная проверка гипотез)

В исследованиях, связанных с экспериментами и реальными данными, никогда не делается вывод на основании одной проверки гипотез, наоборот, практикуется множественная проверка гипотез.

Итак, пусть имеются данные $X = \{X_i^{(j)}\}_{j=1..m}$, j — номер выборки, i — номер сл. в. в выборке.

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

Пусть $p_j = p_j(X_1^{(j)} \dots X_{n_j}^{(j)})$ — p-значения критериев.

	Число верных H_i	# неверных H_i	Всего
# принятых H_i	U	T	$m - R$
# отверг. H_i	V	S	R
Всего	m_0	$m - m_0$	

ошибки II рода

ошибки I рода

Как и ранее, считаем, что ошибка I рода — плохая, и её надо избегать.

<u>I подход</u> $FWER = P(V>0)$ — групповая ошибка I рода

Контроль над FWER на уровне α означает: $FWER < \alpha \;\; \forall P$

Пусть $\alpha_1 \dots \alpha_m$ — уровни значимости, на кот. проверяем $H_1 \dots H_m$, хотим их выбрать так, чтобы $FWER \le \alpha$.

<u>Метод Бонферрони</u> $\alpha_1 = \dots = \alpha_m = \alpha/m$ — очевидно, что работает.

Проблема метода — резкое уменьшение мощности стат. процедуры при $m \to \infty$.

<u>Метод Шидака</u> $\alpha_1 = \dots = \alpha_m = 1-(1-\alpha)^{\frac{1}{m}}$. Метод дает $FWER \le \alpha$ при условии, что статистики T_i независимы или вып. св-во „положительной ~~коррелированности~~ зависимости": $P(T_1 \le t_1 \dots T_m \le t_m) \ge \prod_{i=1}^{m} P(T_i \le t_i) \;\; \forall \vec{t}$.

<u>Нисходящие процедуры</u>: улучшают FWER, но может стать много ошибок второго рода.

Составим вар. ряд уровней значимости: $p_{(1)} \le p_{(2)} \le \dots \le p_{(m)}$, где $H_{(1)} \dots H_{(m)}$ — соответствующие гипотезы. Процедура устроена так:

1) Если $p_{(1)} \ge \alpha_1$, принимаем все гипотезы $H_{(1)}, \dots H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(1)}$ и продолжаем.
2) Если $p_{(2)} \ge \alpha_2$, принимаем $H_{(2)} \dots H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(2)}$ и продолжаем.

И т.д.

<u>Метод Холма</u> $\alpha_1 = \frac{\alpha}{m}$, $\alpha_2 = \frac{\alpha}{m-1} \dots \alpha_m = \alpha$, тогда $FWER \le \alpha$.
(нисходящая проц. с такими $\alpha_i \nearrow$). Если хар-р зависимости между статистиками не учитывать, то нельзя построить контролирующую FWER процедуру мощнее, чем метод Холма.

<u>Нисход. метод Шидака (метод Шидака-Холма)</u> $\alpha_i = 1-(1-\alpha)^{\frac{1}{m-i+1}}$, $i=1 \dots m$.
$FWER \le \alpha$, если статистики T_i независимые. При дост. больших m мало отлич. от метода Холма. Если статистики T_i независимые, то мощнее метода Шидака-Холма процедуру с $FWER \le \alpha$ построить нельзя.

Множественная проверка гипотез (МСПС)

II Метод: Ожидаемая доля ложных отклон. гипотез

$$FDR = E\left(\frac{V}{\max(R,1)}\right) \quad (\text{false discovery rate})$$

$R = V + S$ ← S отвергли, если неверна

Соотв., хотим, чтобы $FDR \le \alpha$ (контроль над FDR на ур. α)

Легко заметить, что всегда $FDR \le FWER$.

Восходящ. метод мн-венной проверки гипотез:

Пусть, как и ранее, $p_{(1)} \le \ldots \le p_{(m)}$ – вариац. ряд p-значений, $H_{(1)}, \ldots, H_{(m)}$ – соответ. гипотезы.

Процедура: 1) если $p_{(m)} \le \alpha_m$, отвергнуть все $H_{(1)} - H_{(m)}$ и остановиться, иначе принять $H_{(m)}$ и продолжить

2) если $p_{(m-1)} \le \alpha_{m-1}$, отвергнуть $H_{(1)} \ldots H_{(m-1)}$ и остановиться, иначе принять $H_{(m-1)}$ и продолжить. И т.д.

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

Метод Бенджамини – Хохберга – восход. процедура с $\alpha_i = \frac{i}{m} \cdot \alpha$.

Обеспеч. контроль над $FDR \le \alpha$, если T_i независимы или если $P(\vec{X} \ge \vec{y} \mid T_i = x)$ не убывает по x, если H_i верна (условие PDRS)

Важный случай: если есть многомер. норм. распред. и корреляции между T_i и T_j, если H_i и H_j верны, неотрицательные.

Метод Бенджамини – Иекутиели – восход. процедура с $\alpha_i = \frac{i}{m} \cdot \alpha \cdot \frac{1}{\sum_{j=1}^{m} \frac{1}{j}}$

Дает $FDR \le \frac{m_0}{m} \alpha \le \alpha \quad \forall p_i$ и T_i (где $m_0 = U + V$ – кол-во верных гипотез)

При отсутствии инф. о зависимости между статистиками метод наилучший.

Если доля невер. гипотез мала, метод Бенджамини – Иекутиели отвергает меньше гипотез, чем метод Холма (но может больше ошибок I рода сделать)

Ср. на один. данных

	H_i+	H_i-
H_i v	150	24
H_i x	0	26

– Холм

148	4
2	46

– Хохберг

~~Замечание. Если проверяем всё время только одну гипотезу различными критериями, то понятно, что пользоваться можно только FWER.~~

Замечание Если проверяем всё время только одну гипотезу (если, к примеру, несколько различных выборок проверяем на нормальность, то это считается за несколько гипотез), то если мы её отвергли хоть бы раз, то в рамках статистич. процедур, контролирующих FWER или FDR на уровне α, следует её отвергнуть.

МСПС | Критерии случайности |

Ранее объектом нашего изучения всегда становилась выборка, т.е. последов. незав. о.р. сл.в. Проблема в том, что не всегда данные образуют выборку. Про проверку одинаковой распред. или отсутствия таковой поговорим позже, когда будем изучать регрессии и врем. ряды. В ~~а~~ любом случае, надо сначала ~~выч~~есть тренд, очевидный если он есть, а потом уже смотреть на случайность. Итак, рассм. H_0: данные образуют выборку.

Сериальный крит. Вальда-Волфовитца

Пусть $X_1 \ldots X_n$ – данные, а $\hat{\mu}$ – их выбороч. медиана.

Обозначим $\delta_i = I(X_i \geq \hat{\mu}) \Rightarrow$ получаем ряд двоич. $\underbrace{0}_{1\,серия}\overset{вторая}{1}\underbrace{000}_{третья}\overset{4}{11}\underbrace{0}_{5}\overset{6}{111}$ и смотрим, сколько там „серий“, т.е. последовательностей из одинак. чисел.

В нашей выборке их 6. Ну так вот, пусть число серий $= N$, а число 0 и 1 – n_0 и n_1 соотв. Тогда при верной гипотезе о ~~независимости эл.~~ случайности данных, т.е. что данные обр. выборку ~~выборки из данных~~

$$N^* = \frac{N - \left(\frac{2n_0 n_1}{n_0+n_1} + 1\right)}{\sqrt{\frac{2n_0 n_1 (2n_0 n_1 - n_0 - n_1)}{(n_0+n_1)^2 (n_0+n_1-1)}}} \xrightarrow{d} N(0,1).$$

Приближением к $N(0,1)$ можно пользоваться при $n_0, n_1 > 20$.

Критерий популярный, но не самый мощный.

Критерий инверсий Пусть есть данные $X_1 \ldots X_n$. Кол-во инверсий

$I = \sum_{i<j} I(X_i > X_j)$ (см. статистику Манна-Уитни!)

При $n \geq 20$ I приблизит. распред. как $N\Big(\underbrace{\frac{n(n-1)}{4}}_{E(I)}, \underbrace{\frac{2n^3+3n^2-5n}{72}}_{D(I)}\Big)$, т.е.

$\frac{I - E(I)}{\sqrt{DI}} \xrightarrow{d} N(0,1)$. Критерий по эффектив. превосходит большинство непарам. критериев для тренда.

Знаково-ранговый критерий Холлина

Его статистика: $r = \frac{1}{k\cdot(n-1)} \sum_{i=2}^{n} \delta[(X_i - \hat{\mu})(X_{i-1} - \hat{\mu})] R_i R_{i-1}$, $k = k(n)$ – коэфф., зависящий только от объема выборки, R_i – ранг величины $|X_i - \hat{\mu}|$ в ~~ряду~~ вариац. ряду $\{|X_i - \hat{\mu}|\}_{i=1}^{n}$, $\delta(y) = \begin{cases} 1, & y>0 \\ 0, & y=0 \\ -1, & y<0 \end{cases}$, $\hat{\mu}$ – выбороч. медиана

Значение $k(n)$ и критич. знач. критерия можно найти в Кобзаре или в статье Холлина 1990г. Зато довольно мощный.

__Замечание.__ Все перечисленные методы подходят и для улавливания тренда.