Прикладная статистика и анализ данных

Съезд IX

Комбинации критериев

Постановка задачи

$X = (X_1, ..., X_n), Y = (Y_1, ..., Y_n)$ — выборки.

Являются ли они нормальными или нет, однородными или нет?

H_1: выборка X нормальна

H_2: выборка Y нормальна

H_3: выборки однородны: $X_i \stackrel{d}{=} Y_i$

Условие: $FWER \leqslant \alpha$

Идея:

1. Применяем критерии нормальности.
2. *Если нормальность X и Y не отвергнута*
   $\longrightarrow$ Критерии однородности для нормальных;

   Если нормальность X или Y отвергнута
   $\longrightarrow$ Критерии однородности непараметрические.

Идея:

1. Применяем критерии нормальности.
2. *Если нормальность X и Y не отвергнута*
   $\longrightarrow$ Критерии однородности для нормальных;

   Если нормальность X или Y отвергнута
   $\longrightarrow$ Критерии однородности непараметрические.

Реализация:

1. МПГ(критерии нормальности для каждой выборки;
   критерии однородности, требующие нормальность).

   Метод Холма, ур. значимости α.
2. *Если нормальность X и Y не отвергнута* $\longrightarrow$ Выдаем ответ;

   Если нормальность X или Y отвергнута
   $\longrightarrow$ Не смотрим на те критерии однор., делаем другую МПГ:

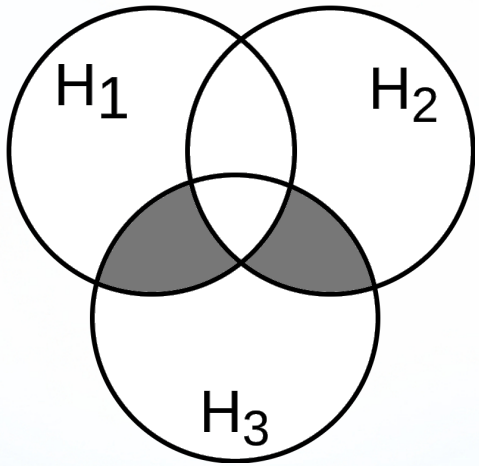
   МПГ(непараметрические критерии однородности)

   Метод Холма, ур. значимости α.

Анализ

Утверждение. Процедура обеспечивает $FWER \leqslant \alpha$ почти всегда.

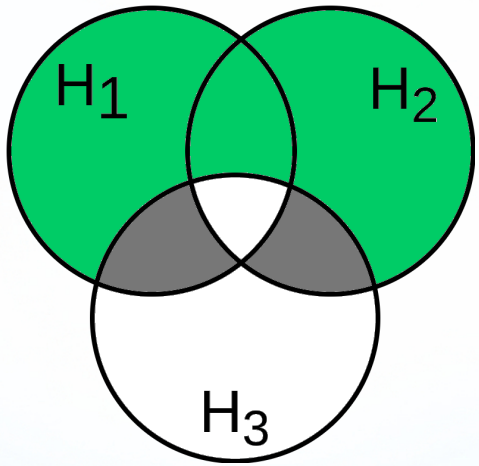
Разберем случаи. Серые области — невозможные.



Анализ

Утверждение. Процедура обеспечивает $FWER \leqslant \alpha$ почти всегда.

Разберем случаи. Серые области — невозможные.



Пусть верны H_1 **или** H_2 **или** $H_1 \,\&\, H_2$**.**

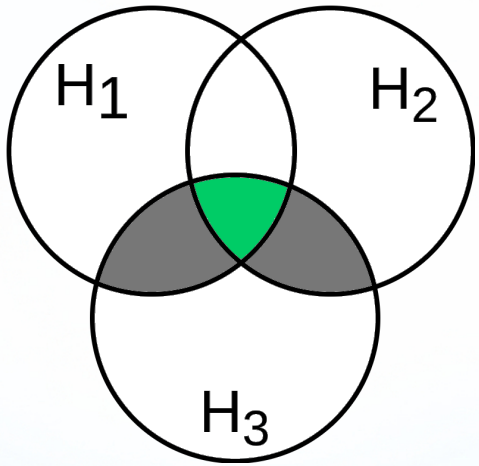
Соответствующие критерии в одной процедуре.

$\Longrightarrow$ отклоняет в $\leqslant \alpha$ случаев независимо от критериев однор..

Анализ

Утверждение. Процедура обеспечивает $FWER \leqslant \alpha$ почти всегда.

Разберем случаи. Серые области — невозможные.



Пусть верны H_1 & H_2 & H_3**.**

Ошибка происходит $\Leftrightarrow$ что-либо отклоняется в первой процедуре,

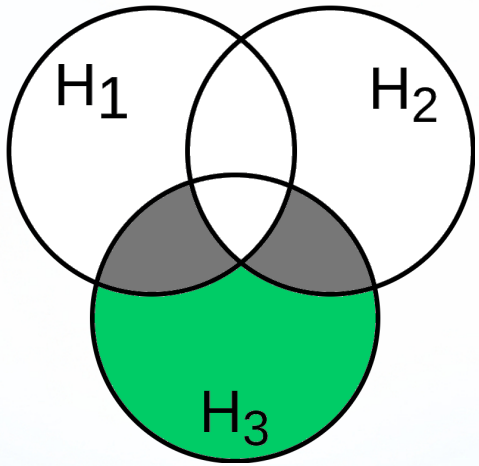
что происходит с вероятностью $\leqslant \alpha$.

Вторая процедура работает только внутри этого множества.

Анализ

Утверждение. Процедура обеспечивает $FWER \leqslant \alpha$ почти всегда.

Разберем случаи. Серые области — невозможные.

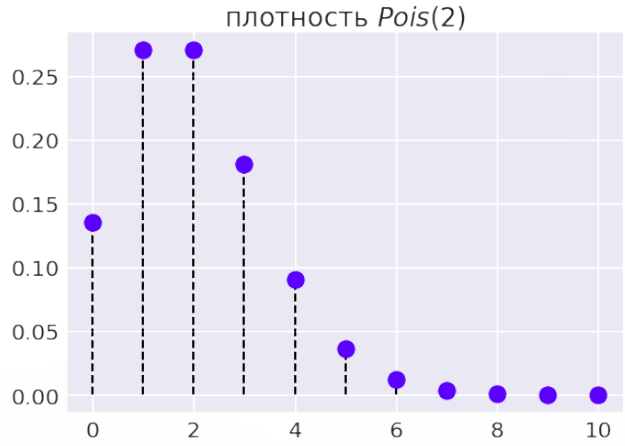


Пусть верна H_3**.**

Если H_1 или H_2 отвергаются (в чем нет ошибки), то переходим во вторую процедуру, где можем совершить ошибку с вер-тью $\leqslant \alpha$. Иначе смотрим на критерии однор. при неверных предположениях. Но если H_1 и H_2 не отвергаются, то данные обладают свойствами, похожими на нормальность. Обычно, именно это требуют критерии.

Сравнение интенсивностей событий

Пуассоновское распределение



$Pois(\lambda): \quad p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \ x \in \mathbb{Z}_+$

Смысл: число событий,

произошедших за единицу времени

Условия:

1. события происходят с фиксированной интенсивностью λ.
2. независимо друг от друга.

Утверждение: время между двумя событиями имеет распр. $Exp(\lambda)$ (*см. пуассоновские случайные процессы*)

Примеры:

1. число клиентов в час
2. число запросов на сервер за минуту

Интенсивность не постоянна, может зависеть от каких-то факторов.

Независимые экспоненциальные выборки

$X = (X_1, ..., X_n) \sim Exp(\lambda_1)$

$Y = (Y_1, ..., Y_m) \sim Exp(\lambda_2)$

$H_0 : \lambda_1 = \lambda_2$ *vs.* $H_1 : \lambda_1 \{<, \neq, >\} \lambda_2$

Статистика критерия: $F(X, Y) = \overline{X} \big/ \overline{Y} \sim F_{2n,2m}$ при H_0.

1. $H_1 : \lambda_1 \neq \lambda_2 \Longrightarrow S = \{F(x, y) < F_{2n,2m,\alpha/2}\} \cup \{F(x, y) > F_{2n,2m,1-\alpha/2}\}$
2. $H_1 : \lambda_1 > \lambda_2 \Longrightarrow S = \{F(x, y) < F_{2n,2m,\alpha}\}$
3. $H_1 : \lambda_1 < \lambda_2 \Longrightarrow S = \{F(x, y) > F_{2n,2m,1-\alpha}\}$

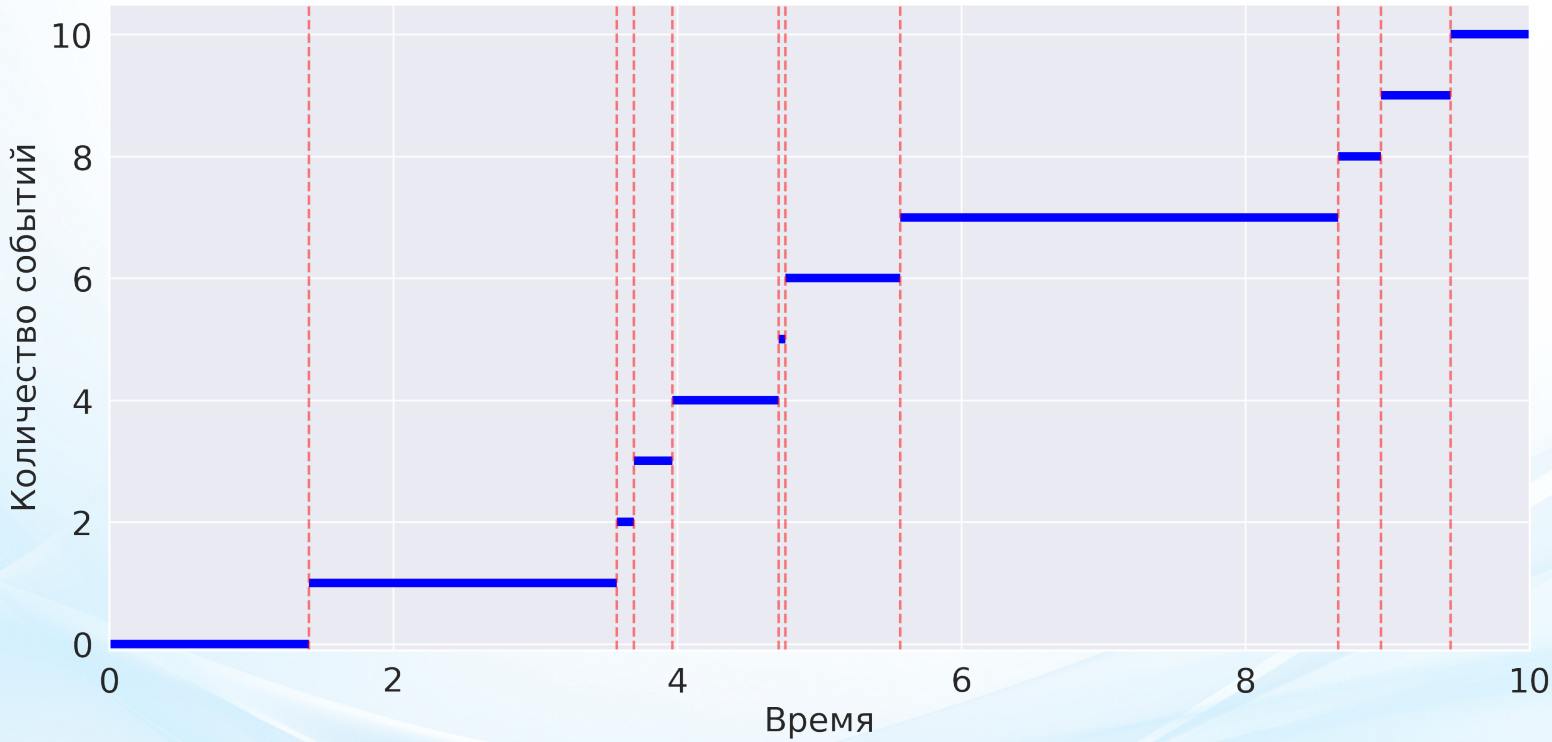
Доказательство:

$$\sum_{i=1}^{n} X_i \sim \Gamma(\lambda_1, n) \Longrightarrow 2\lambda_1 \sum_{i=1}^{n} X_i \sim \Gamma(1/2, n) = \chi^2_{2n}$$

$$\overline{X} \big/ \overline{Y} \overset{H_0}{=} \frac{2\lambda_1 \sum_{i=1}^{n} X_i \big/ 2n}{2\lambda_2 \sum_{i=1}^{m} Y_i \big/ 2m} \sim \frac{\chi^2_{2n} \big/ 2n}{\chi^2_{2m} \big/ 2m} = F_{2n,2m}$$

Заметим, что в общем случае $\frac{\lambda_1 \overline{X}}{\lambda_2 \overline{Y}} \sim F_{2n,2m}$

Пуассоновские процессы



Сравнение интенсивностей пуассоновских процессов

N_1 — кол-во событий процесса 1 интенсивности λ_1 за время t_1.

N_2 — кол-во событий процесса 2 интенсивности λ_1 за время t_2.

Тогда $N_1 \sim Pois(\lambda_1 t_1), N_2 \sim Pois(\lambda_2 t_2)$

$H_0 : \lambda_1 = \lambda_2$ *vs.* $H_1 : \lambda_1 \{<, \neq, >\} \lambda_2$

Статистика критерия

$$Z(X, Y) = \left(N_2 - \frac{t_2}{t_1} N_1 \right) \left[\frac{t_2}{t_1} (N_1 + 1) \right]^{-1/2}$$

Распределение приближается нормальным.

Для $H_0 : \lambda_1 = \lambda_2$ *vs.* $H_1 : \lambda_1 < \lambda_2$ критерий $\{Z(x, y) > z_\alpha\}$

Более точное приближение с помощью статистики

$$Z(X, Y) = 2 \left(\sqrt{N_2 + \frac{3}{8}} - \sqrt{\frac{t_2}{t_1} \left(N_1 + \frac{3}{8} \right)} \right) \left[1 + \frac{t_2}{t_1} \right]^{-1/2}$$

Перестановочные критерии

Перестановочные критерии

Как построить критерий?

$T(X)$ — статистика критерия

$\{T(X) \geqslant c_\alpha\}$ — критерий

Далее нужно либо найти c_α, либо научится считать p-value.

Для этого нужно знать распределение $T(X)$ при H_0. Всего лишь то...

Что делать если не выходит? Или просто лень...

Идея: найти группу "перестановок" G исходной выборки X, для которой распределение выборки при справедливости H_0 совпадает с распределением $gX, \quad \forall g \in G$.

Пример: гипотеза о среднем

$X = (X_1, ..., X_n)$ — выборка из неизвестного распределения $\mathsf{P} \in \mathscr{P}$,

где $\mathscr{P}$ — симметричные распределения относительно $\mathsf{E}X_1$.

$\mathsf{H}_0 : \mathsf{E}X_1 = 0 \quad vs. \quad \mathsf{H}_1 : \mathsf{E}X_1 < 0$ или $\mathsf{H}_1 : \mathsf{E}X_1 > 0$ или $\mathsf{H}_1 : \mathsf{E}X_1 \neq 0$

$T(X) = \sum\limits_{i=1}^{n} X_i$ — статистика критерия

$G = \{(s_1, ..., s_n) \mid s_i \in \{-1, 1\}\}$ — группа, для которой выполнено

$\forall g \in G : \; gX \overset{d_0}{=} X$, где $gX = (s_1 X_1, ..., s_n X_n)$

$x = (x_1, ..., x_n)$ — реализация выборки. Тогда pvalue:

$$p(x) = \frac{1}{2^n} \begin{cases} \sum\limits_{g \in G} I\{T(gx) \leqslant T(x)\}, \text{ если } \mathsf{H}_1 : \mathsf{E}X_1 < 0; \text{ (аналог. для } >) \\ \sum\limits_{g \in G} I\{|T(gx)| \geqslant |T(x)|\}, \text{ если } \mathsf{H}_1 : \mathsf{E}X_1 \neq 0. \end{cases}$$

Пример

Выборка $X = (-7, 1, 5)$

Гипотезы: $\mathrm{H}_0 : \mathrm{E}X_i = 0$ *vs.* $\mathrm{E}X_i < 0$

Статистика критерия: $T(x) = -1$

Группа:				Перестановки				$T(gx)$
1	1	1	$\Longrightarrow$	-7	1	5	$\Longrightarrow$	-1
1	1	-1	$\Longrightarrow$	-7	1	-5	$\Longrightarrow$	-11
1	-1	1	$\Longrightarrow$	-7	-1	5	$\Longrightarrow$	-3
1	-1	-1	$\Longrightarrow$	-7	-1	-5	$\Longrightarrow$	-13
-1	1	1	$\Longrightarrow$	7	1	5	$\Longrightarrow$	13
-1	1	-1	$\Longrightarrow$	7	1	-5	$\Longrightarrow$	3
-1	-1	1	$\Longrightarrow$	7	-1	5	$\Longrightarrow$	11
-1	-1	-1	$\Longrightarrow$	7	-1	-5	$\Longrightarrow$	1

p-value=$4/2^{-3} = 0.5$

Пример: гипотеза о равенстве средних

$X = (X_1, ..., X_n), Y = (Y_1, ..., Y_n)$ — связные выборки

$\mathrm{H}_0 : \mathrm{E}X_1 = \mathrm{E}Y_1$

$\mathrm{H}_1 : \mathrm{E}X_1 < \mathrm{E}Y_1$ или $\mathrm{H}_1 : \mathrm{E}X_1 > \mathrm{E}Y_1$ или $\mathrm{H}_1 : \mathrm{E}X_1 \neq \mathrm{E}Y_1$

$T(X, Y) = \sum\limits_{i=1}^{n} D_i$ — статистика критерия, где $D = (X_i - Y_i)_{i=1}^{n}$

$G = \{(s_1, ..., s_n) \mid s_i \in \{-1, 1\}\}$ — группа, для которой выполнено

$\forall g \in G : \ gD \overset{d_0}{=} D$, где $gD = (s_1 D_1, ..., s_n D_n)$

$d = (d_1, ..., d_n)$ — реализация разностей. Тогда pvalue:

$$p(x) = \frac{1}{2^n} \begin{cases} \sum\limits_{g \in G} I\{T(gd) \leqslant T(d)\}, \text{ если } \mathrm{H}_1 : \mathrm{E}X_1 < \mathrm{E}Y_1; \text{ (аналог. для } >) \\ \sum\limits_{g \in G} I\{|T(gd)| \geqslant |T(d)|\}, \text{ если } \mathrm{H}_1 : \mathrm{E}X_1 \neq \mathrm{E}Y_1. \end{cases}$$

Пример: гипотеза о равенстве средних

$X^1 = (X_1, ..., X_n), X^2 = (X_{n+1}, ..., X_{n+m})$ — независимые выборки

$H_0 : EX_i^1 = EX_i^2$

$H_1 : EX_i^1 < EX_i^2$ или $H_1 : EX_i^1 > EX_i^2$ или $H_1 : EX_i^1 \neq EX_i^2$

$T(X^1, X^2) = \overline{X^1} - \overline{X^2}$ — статистика критерия

$G = \{(s_1, ..., s_{n+m}) \mid \{s_1, ..., s_n\} \in C^n_{\{1,...,n+m\}}, (s_{n+1}, ..., s_{n+m})$– дополнение$\}$

— группа, для которой выполнено $\forall g \in G : \; g(X^1, X^2) \overset{d_0}{=} (X^1, X^2)$

$x^1 = (x_1, ..., x_n), x^2 = (x_{n+1}, ..., x_{n+m})$ — реализация. Тогда pvalue:

$$p(x) = \frac{1}{C^n_{n+m}} \begin{cases} \sum\limits_{g \in G} I\{T(g(x^1, x^2)) \leqslant T(x^1, x^2)\}, \text{ если } H_1 : EX_i^1 < EX_i^2; \\ \sum\limits_{g \in G} I\{|T(g(x^1, x^2))| \geqslant |T(x^1, x^2)|\}, \text{ если } H_1 : EX_i^1 \neq EX_i^2. \end{cases}$$

Пример

Независимые выборки $X^1 = (3, 2)$, $X^2 = (6, 4, 8)$

Гипотезы: $\mathrm{H}_0 : \mathrm{E}X_i^1 = \mathrm{E}X_i^2$ *vs.* $\mathrm{E}X_i^1 < \mathrm{E}X_i^2$

Статистика критерия: $T(x^1, x^2) = 2.5 - 6 = -3.5$

Перестановки						$T(gx)$
3	2	6	4	8	$\Longrightarrow$	-3,5
3	6	2	4	8	$\Longrightarrow$	-0,16
3	4	6	2	8	$\Longrightarrow$	-1,8
3	8	6	4	2	$\Longrightarrow$	1,5
2	6	3	4	8	$\Longrightarrow$	-1
2	4	6	3	8	$\Longrightarrow$	-2,67
2	8	6	4	2	$\Longrightarrow$	1
6	4	3	2	8	$\Longrightarrow$	0,67
6	8	3	2	2	$\Longrightarrow$	4,67
4	8	6	3	2	$\Longrightarrow$	2,33

p-value=1/10

2. Множественная проверка гипотез с помощью перестановок

Пакет на R:

http://web.mit.edu/r/current/arch/i386_linux26

/lib/R/library/multtest/html/mt.maxT.html

Задача МПГ

$X_j = (X_{j1}, ..., X_{jn_j})$ — j-ая выборка, $j = 1, ..., m$

H_j: $\mathrm{P}_j \in \mathscr{P}_j$, где P_j — распределение j-ой выборки.

$T_j(X_j)$ — статистика для проверки H_j.

$t_j = T_j(x_j)$ — реализация статистики

Предположения:

1. совместное распределение $T_{j_1}, ..., T_{j_s}$ при справедливости $\mathrm{H}_{j_1}, ..., \mathrm{H}_{j_s}$ не зависит от справедливости остальных гипотез;
2. для T_j определена группа перестановок G_j.

Временное предположение: все статистики T_j имеют одинаковое распределение, причем критерий правосторонний: $\{T_j \geqslant c\}$.

Б.о.о. считаем $t_1 \geqslant ... \geqslant t_m$

Мультигипотеза

Пусть $J \in \{1, ..., m\}$ — набор индексов.

H_J: $P \in \bigcap\limits_{j \in J} \mathscr{P}_j$, т.е. одновременно верны H_j $\forall j \in J$.

$T^{*1}, ..., T^{*B}$ — [набор значений статистик по методу перестановок]

или [бутстрепная выборка] в предположении справедливости H_J.

Тогда для H_J определим pvalue

$$p_J = \frac{1}{B} \sum_{b=1}^{B} I \left\{ \max_{j \in J} T_j^{*b} \geqslant \max_{j \in J} t_j \right\}$$

Процедура

1. Отвергнуть H_1, если $p_{\{1,\dots,m\}} \leqslant \alpha$, иначе остановиться;
2. Отвергнуть H_2, если $p_{\{2,\dots,m\}} \leqslant \alpha$, иначе остановиться;

   ...

j. Отвергнуть H_j, если $p_{\{j,\dots,m\}} \leqslant \alpha$, иначе остановиться;

   ...

Модифицированные pvalue: $\widetilde{p}_j = \max\limits_{s \in \{1,\dots,j\}} p_{\{s,\dots,m\}}$

Т.е. H_j отвергается $\Longleftrightarrow \widetilde{p}_j \leqslant \alpha$.

Избавление от временного предположения

1. Гипотезы упорядочены в соответствии со значениями pvalue:

   $p_1 \leqslant \ldots \leqslant p_m$;

2. $p^{*1}, \ldots, p^{*B}$ — [набор значений pvalue по методу перестановок]

   или [бутстрепная выборка pvalue]

   в предположении справедливости H_J.

   T.e. p_j^{*b} — pvalue для проверки гипотезы H_j,

   посчитанное для статистики T_j^{*b}.

   Тогда

$$p_J = \frac{1}{B} \sum_{b=1}^{B} I \left\{ \min_{j \in J} p_j^{*b} \leqslant \min_{j \in J} p_j \right\}$$