Прикладная статистика

и анализ данных

Съезд VIII

Дисперсионный анализ II

Типы задач Д.А.

1. **Независимые выборки**

   Две группы пациентов. Одним дают одно лекарство, другим — другое. Верно ли, что первое лекарство эффективнее?

2. **Связные выборки**

   Пациент проходит испытание, принимает средство, затем снова проходит испытание. Отличается ли эффект?

- Методы для задач 2 типа можно использовать для задач 1 типа. При этом теряется важная информация.
- Методы для задач 1 типа *нельзя* использовать для задач 2 типа.

Напоминание: уже рассмотренные задачи

Бернуллиевские выборки

$X_1, ..., X_n \sim Bern(p)$

$Y_1, ..., Y_n \sim Bern(q)$.

$\mathrm{H}_0 : p = q$

$\mathrm{H}_1 : p \ \{<, \neq, >\} \ q$

Нормальные выборки

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

1. Равенство средних

$\mathrm{H}_0 : a_1 = a_2 \quad vs. \quad \mathrm{H}_1 : a_1 \ \{<, \neq, >\} \ a_2$

Способ зависит от доступной инф. о дисп.

2. Равенство дисперсий

$\mathrm{H}_0 : \sigma_1 = \sigma_2 \quad vs. \quad \mathrm{H}_1 : \sigma_1 \ \{<, \neq, >\} \ \sigma_2$

3. Однородность

$\mathrm{H}_0 : (a_1, \sigma_1^2) = (a_2, \sigma_2^2)$

Непараметрический случай

Непараметрический = свободный от семейства распределений

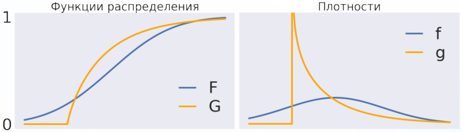
Альтернативы

$X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$ — две выборки из неизвестных **непрерывных** распределений с функциями распределений F и G.

$\mathrm{H}_0\colon F = G$ — гипотеза однородности

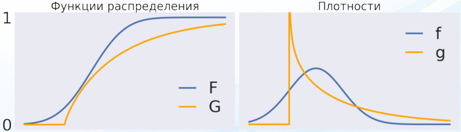
Гипотеза неоднородности:

$\mathrm{H}_1\colon F \neq G$



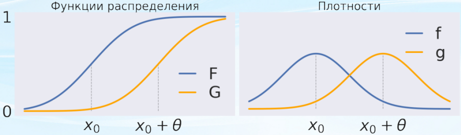
Гипотеза доминирования:

$\mathrm{H}_2\colon F \geqslant G$



Гипотеза сдвига:

$\mathrm{H}_3\colon F(x - \theta) = G(x)$



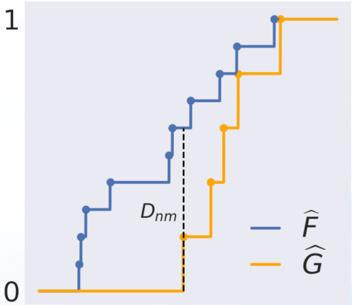
Непараметрический случай

Независимые выборки

Критерии на основе ЭФР: 1. Критерий Смирнова

$X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$ — две выборки из неизвестных непрерывных распределений с функциями распределений F и G.

$\mathrm{H}_0: F = G$ *vs.* $\mathrm{H}_1: F \neq G$



Статистика $D_{nm} = \sup\limits_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - \widehat{G}_m(x) \right|$

Вычисление:

$$D_{nm} = \max\left\{D_{nm}^+, D_{nm}^-\right\}$$
$$D_{nm}^+ = \max_{i=1..n} \left\{ i/n - \widehat{G}_m\left(X_{(i)}\right) \right\}$$
$$D_{nm}^- = \max_{j=1..m} \left\{ j/m - \widehat{F}_n\left(Y_{(j)}\right) \right\}$$

$$\sqrt{\frac{nm}{n+m}} D_{nm} \xrightarrow{d_0} Kolmogorov, \text{ при } n, m \to +\infty \text{ если } n/(n+m) \to \gamma \in (0,1)$$

Приближение точное при $n, m \geqslant 20$

Критерии на основе ЭФР: 2. Критерий Розенблатта

$X_1, ..., X_n$ и $Y_1, ..., Y_m$ — две выборки из неизвестных непрерывных распределений с функциями распределений F и G.

$\mathrm{H}_0: F = G$ *vs.* $\mathrm{H}_1: F \neq G$

Статистика $\omega^2_{nm} = \int\limits_{\mathbb{R}} \left(\widehat{F}_n(x) - \widehat{G}_m(x)\right)^2 d\widehat{H}_{n+m}(x)$,

$\widehat{H}_{n+m}(x) = \frac{n}{n+m}\widehat{F}_n(x) + \frac{m}{n+m}\widehat{G}_m(x)$ — ЭФР по объединенной выборке.

Вычисление:

R_i, S_j — ранги X_i, Y_j в вариац. ряду по выборке $(X_1, .., X_n, Y_1, .., Y_m)$

$$\omega^2_{nm} = \frac{1}{nm}\left(\frac{1}{6} + \frac{1}{m}\sum_{i=1}^{n}(R_i - i)^2 + \frac{1}{n}\sum_{j=1}^{m}(S_j - j)^2\right) - \frac{2}{3}$$

Критерий Уилкоксона-Манна-Уитни

Рассматриваем альтернативу $\mathrm{H}_2\colon F \geqslant G$.

S_j — ранг Y_j в вариационном ряду по выборке $(X_1, .., X_n, Y_1, .., Y_m)$.

$V = S_1 + \ldots + S_m$ — статистика критерия.

$$\frac{V - \mathrm{E}V}{\sqrt{\mathrm{D}V}} \xrightarrow{d_0} \mathcal{N}(0,1),$$

где $\mathrm{E}V = \frac{m(n+m+1)}{2}, \quad \mathrm{D}V = \frac{nm(n+m+1)}{12}$ при H_0.

Идея: если H_0 верна, то значения $Y_{(j)}$ равномерно разбросаны по вар. ряду.
Большие значения V указывают на преобладание Y_j над X_i.
Критерий имеет вид $S = \{V > c\}$.

- приближение при $n, m \geqslant 50$;
- если $n, m \geqslant 25$, используется поправка Имана;
- при малых n и m используются таблицы.

Критерий Уилкоксона-Манна-Уитни

Совпадения

- Рассматриваются средние ранги
- Дисперсия

$$\mathrm{D}V = \frac{nm}{12}\left(n+m+1-\frac{1}{(n+m)(n+m-1)}\sum_{k=1}^{g} l_k(l_k-1)\right),$$

g — число групп совпадений

l_k — количество элементов в k-ой группе.

Критерий Уилкоксона-Манна-Уитни

Оценка параметра сдвига

В случае альтернативы H_3: $F(x-\theta) = G(x)$ оценка

$$\widehat{\theta} = med\{W_{ij} = Y_j - X_i, i = 1..n, j = 1..m\}$$

Свойство: $\sqrt{\frac{nm}{n+m}}\left(\widehat{\theta} - \theta\right) \xrightarrow{d_0} \mathcal{N}\left(0, \sigma^2\right), \quad \left[n, m \to +\infty, \frac{n}{n+m} \to \gamma \in (0,1)\right]$
где $\sigma^{-1} = \sqrt{12}\int\limits_{\mathbb{R}} p^2(x)dx$,

$p(x)$ — плотность ф.р. F.

Доверительный интервал параметра сдвига

$(W_{(k_\alpha+1)}, W_{(nm-k_\alpha)})$,
где $k_\alpha = \left\lfloor nm/2 - 1/2 - z_{1-\alpha}\sqrt{nm(n+m+1)/12}\right\rfloor$

Связь оценки и критерия

Статистика Манна-Уитни:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I\{X_i \leqslant Y_j\}$$

При отсутствии совпадений $U = V - \frac{m(m+1)}{2}$.

Пусть θ — неизвестный сдвиг.

Тогда $(X_1, \ldots, X_n)$ и $(Y_1 - \theta, \ldots, Y_m - \theta)$ однородны.

$\Longrightarrow$ для них распределение U симметрично относительно $\frac{nm}{2}$.

Получаем уравнение

$$\sum_{i=1}^{n} \sum_{j=1}^{m} I\{X_i \leqslant Y_j - \theta\} = \sum_{i=1}^{n} \sum_{j=1}^{m} I\{Y_i - X_i \geqslant \theta\} = \frac{nm}{2}$$

Откуда $\widehat{\theta} = med\{W_{ij} = Y_j - X_i, i = 1..n, j = 1..m\}$

Пример: рост кошек и собак

Выборка кошек:

Выборка собак:

1

2

3

4

5

6

7

8

9

10

11

Пример: рост кошек и собак

$X_1, ..., X_5$ — рост $n = 5$ кошек

$Y_1, ..., Y_6$ — рост $m = 6$ собак

$\mathrm{H}_0:$ рост собак и кошек в среднем одинаковый

$\mathrm{H}_2:$ в среднем рост собак больше роста кошек

$V = 4 + 7 + 8 + 9 + 10 + 11 = 49$

$pvalue = 0.009$

Непараметрический случай

Связные выборки

Связные выборки: модель

$X_1, \dots, X_n$ и $Y_1, \dots, Y_n$ — связные выборки

Перейдем к **выборке разностей**:

$Z_i = Y_i - X_i = \theta + \varepsilon_i$,

- $\theta > 0$ — интересующий систематический эффект воздействия;
- $\varepsilon_1, \dots, \varepsilon_n$ — случайные ошибки.

Предположения об ошибках:

- независимы;
- имеют непрерывные распределения (м.б. разные);
- медиана $= 0$.

Гипотезы: $H'_0 \colon \theta = 0$ *vs.* $H'_3 \colon \theta > 0$

Пример, когда $\varepsilon_1, ..., \varepsilon_n$ имеют разные распред.

Пусть $X_1, X_2 \sim \mathcal{N}(0,1)$ независимы.

Тогда $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2 \sim \mathcal{N}(0,2)$ независимы.

Но $Z_1 = Y_1 - X_1 = X_2 \sim \mathcal{N}(0,1)$

$Z_2 = Y_2 - X_2 = X_1 - 2X_2 \sim \mathcal{N}(0,5)$

Получаем

$\varepsilon_1 = Z_1 - \theta \sim \mathcal{N}(-\theta, 1)$

$\varepsilon_2 = Z_2 - \theta \sim \mathcal{N}(-\theta, 5)$

Кроме того, $cov(\varepsilon_1, \varepsilon_2) = -2$, т.е. они зависимы.

Критерий знаков

Рассмотрим знаки $U_i = I\{Z_i > 0\} \sim Bern(p)$

$\mathrm{H}_0'\colon\ p = 1/2 \quad vs. \quad \mathrm{H}_3'\colon\ p > 1/2$

Статистика критерия $S = U_1 + \ldots + U_n \overset{\mathrm{H}_0'}{\sim} Bin(n, 1/2)$

Критерий $\{S > c\}$.

Аппроксимация при $n > 15$: $\quad \frac{S - n/2 - 1/2}{\sqrt{n/4}} \xrightarrow{d_0} \mathcal{N}(0,1),$

Совпадения: выбрасываем соответствующие наблюдения.

Оценка параметра: $\widehat{\theta} = med\{Z_i, i = 1..n\}$

Доверительный интервал для параметра

$(Z_{(k_\alpha+1)}, Z_{(n-k_\alpha)})$ — д.и. уровня доверия $1 - 2\alpha$,

где $k_\alpha = \left\lfloor n/2 - 1/2 - z_{1-\alpha}\sqrt{n/4} \right\rfloor$

Связь оценки и критерия

Знаки $U_i = I\{Z_i > 0\} \sim Bern(p)$

Статистика критерия $S = U_1 + ... + U_n$

Пусть θ — неизвестный сдвиг.

Тогда для $(Z_1 - \theta, ..., Z_n - \theta)$ медиана равна нулю.

Получаем уравнение

$$\sum_{i=1}^{n} I\{Z_i - \theta > 0\} = \sum_{i=1}^{n} I\{Z_i > \theta\} = \frac{n}{2}$$

Откуда $\widehat{\theta} = med\{Z_i, i = 1..n\}$.

Пример: времена реакции (Лагутин)

X_i — время реакции i-го испытуемого на световой сигнал

Y_i — время реакции i-го испытуемого на звуковой сигнал

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	176	163	152	155	156	178	160	164	169	155	122	144
y_i	168	215	172	200	191	197	183	174	176	155	115	163
z_i	-8	+52	+20	+45	+35	+19	+23	+10	+7	0	-7	+19

$S(x) = 9$ — значение статистики критерия

$pvalue = (1 + 11 + 55)/2048 \approx 0.033$

$\widehat{\theta}(x) = 19$

$(7, 35)$ — 90% доверительный интервал

Критерий ранговых сумм Уилкоксона

Дополнительное предположение: $\varepsilon_1, ..., \varepsilon_n$ имеют (одно) симметричное распределение относительно нуля.

R_i — ранг величины $|Z_i|$ в вариационном ряду $|Z_1|, ..., |Z_n|$

$U_i = I\{Z_i > 0\}$ — знак

$T = R_1 U_1 + ... + R_n U_n$ — статистика критерия

$$\frac{T - \mathsf{E}T}{\sqrt{\mathsf{D}T}} \xrightarrow{d_0} \mathcal{N}(0,1), \qquad n > 15$$

где $\mathsf{E}T = \frac{n(n+1)}{4}$, $\mathsf{D}T = \frac{n(n+1)(2n+1)}{4}$ при H_0.

Идея: если H_0 верна, то в силу симметричности ошибок распределения Z_i тоже симметричны, а значит и ранги не зависят от знаков. Критерий имеет вид $S = \{T > c\}$.

Критерий ранговых сумм Уилкоксона

Совпадения

- Если $Z_i = 0$, то его отбрасываем;
- Если среди оставшихся есть совпадения,

  то рассматриваются средние ранги;
- Дисперсия

$$\mathrm{D}V = \frac{1}{24}\left(n(n+1)(2n+1) - \frac{1}{2}\sum_{k=1}^{g} l_k(l_k^2 - 1)\right),$$

g — число групп совпадений;

l_k — количество элементов в k-ой группе.

Критерий ранговых сумм Уилкоксона

Оценка параметра сдвига — медиана средних Уолша

$$\widehat{\theta} = med\{V_{ij} = (Z_i + Z_j)/2, 1 \leqslant i \leqslant j \leqslant n\}$$

Свойство: $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d_0} \mathcal{N}\left(0, \sigma^2\right)$,

где $\sigma^{-1} = \sqrt{12} \int\limits_{\mathbb{R}} p^2(x)dx$,

$p(x)$ — плотность ф.р. F.

Доверительный интервал параметра сдвига

$\left(V_{(k_\alpha+1)}, V_{(n(n+1)/2-k_\alpha)}\right)$ — д.и. уровня доверия $1 - 2\alpha$,

где $k_\alpha = \left\lfloor n(n+1)/4 - 1/2 - z_{1-\alpha}\sqrt{n(n+1)(2n+1)/24}\right\rfloor$

Связь оценки и критерия

$T = R_1 U_1 + \ldots + R_n U_n$ — статистика критерия.

Отсутствии нулей и совпадений среди $|Z_i|$ выполнено

$$T = \sum_{i \leqslant j} I\left\{\frac{Z_i + Z_j}{2} > 0\right\}$$

Пусть θ — неизвестный сдвиг.

Тогда для $(Z_1 - \theta, \ldots, Z_n - \theta)$ распределение статистики T

симметрично относительно среднего $\frac{n(n+1)}{4}$

Получаем уравнение

$$\sum_{i \leqslant j} I\left\{\frac{Z_i - \theta + Z_j - \theta}{2} > 0\right\} = \sum_{i \leqslant j} I\left\{\frac{Z_i + Z_j}{2} > \theta\right\} = \frac{n(n+1)}{2}$$

Откуда $\widehat{\theta} = med\{V_{ij} = (Z_i + Z_j)/2, 1 \leqslant i \leqslant j \leqslant n\}$.

Пример: времена реакции (Лагутин)

X_i — время реакции i-го испытуемого на световой сигнал

Y_i — время реакции i-го испытуемого на звуковой сигнал

z_i	-7	7	-8	10	19	19	20	23	35	45	52
$\lvert z_i \rvert$	7	7	8	10	19	19	20	23	35	45	52
R_i	1.5	1.5	3	4	5.5	5.5	7	8	9	10	11
U_i	0	1	0	1	1	1	1	1	1	1	1

$T(x) = 61.5$ — значение статистики критерия

$T^*(x) = 2.54$ — нормированное значение статистики

$pvalue = 0.006$

$\widehat{\theta}(x) = 19.25$

$(7.5, 31)$ — 90% доверительный интервал

Проверка симметрии

Визуальная проверка симметрии

Пусть u_p — p-квантиль симметричного распределения.

Тогда $u_{1/2} - u_p = u_{1-p} - u_{1/2}$.

Для порядковых статистик стоит ожидать

$\xi_i = \widehat{med} - Z_{(i)} \approx Z_{(n-i+1)} - \widehat{med} = \eta_i, \quad i = 1, ..., \lfloor n/2 \rfloor$

$\Longrightarrow$ точки (ξ_i, η_i) должны располагаться вблизи $y = x$.

Строгая проверка симметрии

См., например, критерий Гупты.

Сравнение критериев

Мощность критериев связана с асимптотической эффективностью соответствующих оценок параметра сдвига.

$ARE_{\widehat{\mu},W}$ — относительная ас. эффективность $\widehat{\mu}$ по отношению к W.

- $ARE_{\widehat{\mu},W} \approx 0.42$ — для нормального распределения;
- $ARE_{\widehat{\mu},W} = 4/3$ — для распределения Лапласса;
- Чем легче хвосты, тем предпочтительнее W по сравнению с $\widehat{\mu}$.


ВСЁ!