Прикладная статистика и анализ данных

Съезд X

Дисперсионный анализ III

Типы задач Д.А. — несколько выборок

1. **Независимые выборки**

   Несколько групп пациентов, которым дают лекарство **с различной дозировкой** активного вещества.

   Есть ли различия в эффективности лечения?

2. **Связные выборки**

   Измеряется активность работников предприятия в течении недели. Отличается ли эффективность в зависимости от **дня недели**?

- 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.

Схема анализа

Имеется несколько выборок в соответствии со значением **фактора** — категориальная переменная.

Этапы анализа:

1. Оказывает ли фактор влияние на исследуемую величину?
2. Если да, то post hoc анализ:
   - Какие значения фактора оказывают различное влияние?
   - Насколько отличаются степени влияния различных значений фактора?

Независимые выборки

Независимые выборки

1	2	...	k
X_{11}	X_{12}	...	X_{1d}
X_{21}	X_{22}	...	X_{2d}
...	...	...	...
$X_{n_1 1}$	$X_{n_2 2}$	...	$X_{n_k k}$

$$X_{ij} = \overbrace{\mu + \beta_j}^{\mu_j} + \varepsilon_{ij},$$

$i = 1, ..., n_j$ — номер наблюдения в выборке

$j = 1, ..., k$ — номер выборки

μ — неизвестное общее среднее

β_j — неизвестный эффект воздействия фактора для j-й выборки

ε_{ij} — случайная ошибка

Предположение:

ε_{ij} независимы и имеют одинаковое непрерывное распределение.

$H_0: \mu_1 = ... = \mu_k$ *vs.* $H_1: \exists j_1, j_2$ т.ч. $\mu_{j_1} \neq \mu_{j_2}$

Независимые выборки (пример)

Исследуется эффективность трех жаропонижающих.

Лекарство 1	Лекарство 2	Лекарство 3
X_{11}	X_{12}	X_{13}
X_{21}	X_{22}	X_{23}
X_{31}		X_{33}
X_{41}		

$j \in \{1, 2, 3\}$ — номер группы пациентов.

X_{ij} — изменения температуры i-го пациента в j-й группе после введения жаропонижающего.

μ_j — среднее изменение температуры после j-го лекарства.

$H_0 : \mu_1 = ... = \mu_k$ — эффект лекарств не отличается

$H_1 : \exists j_1, j_2$ т.ч. $\mu_{j_1} \neq \mu_{j_2}$ — для некоторых лекарств эффект отлич..

Что делаем в первую очередь?

Ящики с усами!!!

F-критерий [требование: $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$]

$$X_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}, \qquad X_{\bullet\bullet} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}, \qquad N = \sum_{j=1}^{k} n_j$$

$$V_{tot} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - X_{\bullet\bullet})^2 = V_{in} + V_{out} \quad - \text{ общая изменчивость}$$

$$V_{in} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - X_{\bullet j})^2 \sim \sigma^2 \cdot \chi^2_{N-k}$$

мера общей изменчивости внутри выборок

$$V_{out} = \sum_{j=1}^{k} n_j (X_{\bullet j} - X_{\bullet\bullet})^2 \sim \sigma^2 \cdot \chi^2_{k-1}$$

мера разброса между выборками

$$F(X) = \frac{V_{out}/(k-1)}{V_{in}/(N-k)} \overset{d_0}{\sim} F_{k-1,N-k}$$

Идея: если разброс между выборками V_{out} сильно больше разброса внутри выборок, то H_0 надо отклонить.
Критерий имеет вид $S = \{F(x) > F_{k-1,N-k,1-\alpha}\}$.

Как проверить равенство дисперсий? $[\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)]$

Критерий Бартлетта

$\mathrm{H}_0 : \sigma_1 = ... = \sigma_k$ при любых $\mu_1, ..., \mu_k$

Статистика критерия

$$B = \left(\frac{1}{N} \sum_{j=1}^{k} n_j S_j^2 \right) \Bigg/ \sqrt[N]{\prod_{j=1}^{k} \left(S_j^2 \right)^{n_j}},$$

где $S_j^2 = \frac{1}{n_j - 1} \sum\limits_{i=1}^{n_j} (X_{ij} - X_{\bullet j})^2$ — несмещ. оценка дисперсии выборки j

$N = n_1 + ... + n_j$ — общее число наблюдений.

Если $n_j > 3$, то при H_0 выполнено $\gamma^{-1} N \ln B \overset{d_0}{\approx} \chi^2_{k-1}$,

где $\gamma = 1 + \frac{1}{3(k-1)} \left(\sum\limits_{j=1}^{k} \frac{1}{n_j} - \frac{1}{N} \right)$

Применимость критериев Бартлетта и F-критерия

Критерий Бартлетта чувствителен к отклонениям от нормальности!

Например, заменим $\mathcal{N}(0,1)$ на T_7.

Значения *реального* уровня значимости при $\alpha = 0.05$:

$k = 2$	$k = 5$	$k = 10$
0.17	0.32	0.49

Критерий Фишера:

1. Если $n_1 = \ldots = n_k$, то критерий устойчив к отклонениям от нормальности и равенства дисперсий;
2. Вместо нормальности достаточно $n_1 \approx \ldots \approx n_k$ и $N - k > 20$;
3. Вместо равенство дисперсий достаточно $\max S_j^2 / \min S_j^2 < 10$;
4. Чувствителен к выбросам.

Критерий Краскела-Уоллиса [непараметрический]

R_{ij} — ранг наблюдения X_{ij} в наборе $(X_{ij}, i = 1, ..., n_j, j = 1, ..., k)$

$R_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} R_{ij}$ — средний ранг в выборке j.

$R_{\bullet\bullet} = \frac{N+1}{2}$ — общий средний ранг

$$W = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R_{\bullet j} - R_{\bullet\bullet}\right)^2 \xrightarrow{d_0} \chi^2_{k-1}$$

Идея: если H_0 верна, то средние ранги по выборкам не сильно отклоняются от общего среднего ранга. Критерий имеет вид $S = \{W(x) > \chi^2_{k-1,1-\alpha}\}$.

При *совпадениях* надо рассмотреть средние ранги и поделить W на

$\gamma = 1 - \frac{1}{N(N^2-1)} \sum_{m=1}^{g} l_m(l_m^2 - 1)$,

где g — число групп совпадений, а l_k — количество элементов в k-ой группе.

Пример

Исследуется эффективность трех жаропонижающих.

Лекарство 1	Лекарство 2	Лекарство 3
$-7.4 \quad (R_{11} = 1)$	$-1.5 \quad (R_{21} = 7)$	$+2.5 \quad (R_{31} = 9)$
$-6.2 \quad (R_{12} = 2)$	$+1.7 \quad (R_{22} = 8)$	$-1.9 \quad (R_{32} = 6)$
$-4.6 \quad (R_{13} = 4)$		$-2.1 \quad (R_{33} = 5)$
$-5.3 \quad (R_{14} = 3)$		
$n_1 = 4$	$n_2 = 2$	$n_3 = 3$
$R_{\bullet 1} = 2.5$	$R_{\bullet 2} = 7.5$	$R_{\bullet 3} = 6.66$

$R_{\bullet\bullet} = \frac{N+1}{2} = 5$ — общий средний ранг.

$W = \frac{12}{9 \cdot 10}\left[4 \cdot (2.5-5)^2 + 2 \cdot (7.5-5)^2 + 3 \cdot (6.66-5)^2\right] \approx 6.1.$

$pvalue \approx 0.0471 \implies$ отвергаем гипотезу об одинаковом эффекте.

Критерий Джонкхиера [непараметрический]

$\mathrm{H}_0 : \mu_1 = \ldots = \mu_k$ *vs.* $\mathrm{H}_1 : \mu_1 \leqslant \ldots \leqslant \mu_k$.

Например, лекарства отличаются дозировкой

$\Longrightarrow$ стоит ожидать упорядоченный эффект

$U_{rs} = \sum\limits_{i=1}^{n_r} \sum\limits_{l=1}^{n_s} I\{X_{ir} < X_{ls}\}$ — статистики критерия Манна-Уитни

$J = \sum\limits_{r<s} U_{rs}$ — статистика критерия Джонкхиера

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

$$\mathrm{E}J = \sum_{r<s} \mathrm{E}U_{rs} = \frac{1}{2} \sum_{r<s} n_r n_s = \frac{1}{4} \left(N^2 - \sum_{j=1}^{k} n_j^2 \right)$$

$$\mathrm{D}J = \frac{1}{72} \left(N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3) \right)$$

Независимые выборки

Post hoc анализ

Общая суть:

если $|X_{\bullet r} - X_{\bullet s}| > c$ или $|R_{\bullet r} - R_{\bullet s}| > c$, то отвергаем $H_{rs}: \mu_r = \mu_s$.

LSD Фишера [требование: $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$]

Пусть H_0 **отвергается**.

Далее проверяем гипотезы H_{rs}: $\mu_r = \mu_s$ с помощью T-критерия:

$$T(X) = \frac{X_{\bullet r} - X_{\bullet s}}{S\sqrt{1/n_r + 1/n_s}} \overset{H_{rs}}{\sim} T_{n_r+n_s-2} \qquad +\text{МПГ}$$

$S^2 = \frac{(n_r-1)S_r^2+(n_s-1)S_s^2}{n_r+n_s-2}$ — несмещенная оценка σ.

S_j^2 — *несмещенная* оценка дисперсии выборки j.

$LSD_{rs} = t_{n_r+n_s-2,1-\alpha/2} \cdot S\sqrt{1/n_r + 1/n_s}$

Если $|X_{\bullet r} - X_{\bullet s}| > LSD_{rs}$, то H_{rs} отвергается

LSD_{rs} — наименьшая значимая разность (least significant difference)

HSD Тьюки

[требование: $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$]

Не зависимо от справедливости H_0: $\mu_1 = ... = \mu_k$:

Гипотеза H_{rs}: $\mu_r = \mu_s$ отвергается если

$|X_{\bullet r} - X_{\bullet s}| > HSD,$ +МПГ

где $HSD = q_{n-k,\alpha} S/M$ — критерий подлинной значимости (honest significant difference),

$M = k \Big/ \sum_{j=1}^{k} 1/n_j$,

S_j^2 — *несмещенная* оценка дисперсии выборки j,

$S^2 = \frac{1}{n-k} \sum_{j=1}^{k} (n_k - 1) S_k^2$,

$q_{n-k,\alpha}$ — α-квантиль распределения стьюдентизированного размаха с $n - k$ степенями свободы (см. далее).

Критерий Неменья [непараметрический]

Пусть H_0 **отвергается**.

Далее проверяем гипотезы $H_{rs}: \mu_r = \mu_s$ +МПГ

Если $|R_{\bullet r} - R_{\bullet s}| > CD = q_{k,1-\alpha/2}\sqrt{\frac{k(k+1)}{6N}}$, то H_{rs} отвергается,

где $q_{k,p}$ — p-квантиль распр. стьюд. размаха с k ст. свободы

Пусть $\xi_{ij} \sim \mathcal{N}(0, \sigma^2), i = 1..n, j = 1..k$. **Распр. стюдентизированного размаха** с k степенями свободы — распр. величины

$$\frac{\max\limits_j \xi_{\bullet j} - \min\limits_j \xi_{\bullet j}}{S/\sqrt{n}},$$

S — выборочная дисперсия по всей совокупности

Критерий Данна [непараметрический]

Пусть H_0 **отвергается**.

Далее проверяем гипотезы $H_{rs}: \mu_r = \mu_s$ +МПГ

приближенным критерием: если

$$|R_{\bullet r} - R_{\bullet s}| > C_{rs} = z_{1-\alpha/2}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_r} + \frac{1}{n_s}\right)},$$

то H_{rs} отвергается,

где $N = n_1 + ... + n_k$ — общее число наблюдений.

Оценка контраста

Пусть H_{rs} отвергается $\Longrightarrow$ оцениваем контраст $\Delta_{rs} = \mu_r - \mu_s$.

$V_{rs} = med\{X_{ir} - X_{ls}, i = 1..n_r, j = 1..n_s\}$ — первичная оценка

$W_r = \frac{1}{N} \sum\limits_{s=1}^{k} n_s V_{rs}$, где $V_{rr} = 0$

$\widehat{\Delta} = W_r - W_s$ — уточненная оценка контраста

Свойства:

1. Первичные оценки могут быть несогласованными: $V_{12} \neq V_{13} + V_{32}$
2. Уточненные оценки согласованы и состоятельны.
3. Уточненные оценки зависят от всех выборок.

Связные выборки

Связные выборки

	1	2	...	k
1	X_{11}	X_{12}	...	X_{1d}
2	X_{21}	X_{22}	...	X_{2d}
	...	...	...	...
n	X_{n1}	X_{n2}	...	X_{nk}

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

$i = 1, ..., n$ — номер наблюдения в выборке

$j = 1, ..., k$ — номер выборки

μ — неизвестное общее среднее

α_i — неизвестный эффект блока i (мешающий параметр)

β_j — неизвестный эффект выборки j (интересующий нас параметр)

ε_{ij} — случайная ошибка

Предположение:

ε_{ij} независимы и имеют одинаковое непрерывное распределение.

$H_0 : \beta_1 = ... = \beta_k$ *vs.* $H_1 : \exists j_1, j_2$ т.ч. $\beta_{j_1} \neq \beta_{j_2}$.

Связные выборки

Исследуется эффективность работы в зависимости от дня недели.

Человек	Понедельник	Вторник	Среда	Четверг	Пятница
Иван	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
Ольга	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
Артем	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}
Ирина	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}

$j \in \{1, 2, 3, 4, 5\}$ — номер дня недели; $i \in \{1, 2, 3, 4\}$ — номер человека.

X_{ij} — время работы i-го человека в день j.

μ — время работы сотрудника по приказу.

α_i — среднее отклонение i-го человека от приказа.

β_j — влияние j-го дня.

$H_0: \mu_1 = ... = \mu_k$ — дни недели не влияют на эффективность

$H_1: \exists j_1, j_2$ т.ч. $\mu_{j_1} \neq \mu_{j_2}$ — для некоторых дней эффективность отлич..

F-критерий [требование: $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$]

$$X_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \qquad X_{i\bullet} = \frac{1}{k}\sum_{j=1}^{k} X_{ij}, \qquad X_{\bullet\bullet} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}$$

$$V_{tot} = \sum_{i=1}^{n}\sum_{j=1}^{k} (X_{ij} - X_{\bullet\bullet})^2 = V_\alpha + V_\beta + V_{in} \quad — \text{ общая изменчивость}$$

$$V_\alpha = k\sum_{i=1}^{n} (X_{i\bullet} - X_{\bullet\bullet})^2 \sim \sigma^2 \cdot \chi^2_{n-1} \quad — \text{изменчивость блоков}$$

$$V_\beta = n\sum_{j=1}^{k} (X_{\bullet j} - X_{\bullet\bullet})^2 \sim \sigma^2 \cdot \chi^2_{k-1} \quad — \text{изменчивость выборок}$$

$$V_{in} = \sum_{i=1}^{n}\sum_{j=1}^{k} (X_{ij} - X_{i\bullet} - X_{\bullet j} + X_{\bullet\bullet})^2 \sim \sigma^2 \cdot \chi^2_{(n-1)(k-1)} \quad — \text{изменч. данных}$$

$$F(X) = \frac{V_\beta/(k-1)}{V_{in}/(n-1)(k-1)} \sim F_{k-1,(n-1)(k-1)}$$

Идея: если изменчивость выборок V_β
сильно больше общей изменчивости данных, то H_0 надо отклонить.
Критерий имеет вид $S = \{F(x) > F_{k-1,(n-1)(k-1),1-\alpha}\}$.

Критерий Фридмана [непараметрический]

Q_{ij} — ранг наблюдения X_{ij} в наборе $(X_{i1}, ..., X_{ik})$

$T_j = \sum_{i=1}^{n_j} Q_{ij}$ — сумма рангов выборки j

$Q_{\bullet j} = T_j / n$ — средний ранг в выборке j

$Q_{\bullet\bullet} = \frac{k+1}{2}$ — общий средний ранг

$$F = \frac{12n}{k(k+1)} \sum_{j=1}^{k} (Q_{\bullet j} - Q_{\bullet\bullet})^2 \xrightarrow{d_0} \chi^2_{k-1}$$

Идея: если H_0 верна, то средние ранги по выборкам не сильно отклоняются от общего среднего ранга. Критерий имеет вид $S = \{F(x) > \chi^2_{k-1,1-\alpha}\}$.

При *совпадениях* надо рассмотреть средние ранги и взять

$F = \frac{12 \sum_{j=1}^{k} (T_j - nQ_{\bullet\bullet})^2}{nk(k+1) - \frac{1}{k-1} \sum_{i=1}^{n} \left(\sum_{m=1}^{g_i} l_{im}^3 - k \right)}$, где g_i — число групп совпадений в блоке i, а l_{im} — количество элементов в m-ой группе блока i.

Пример

Исследуется эффективность работы в зависимости от дня недели.

В скобках указаны ранги по блокам

Ч.	Понедельник	Вторник	Среда	Четверг	Пятница
1	6.7 ($Q_{11}=1$)	7.6 ($Q_{12}=3$)	8.2 (5)	7.9 (4)	7.1 (2)
2	4.9 ($Q_{21}=2$)	6.2 ($Q_{22}=5$)	5.1 (3)	5.9 (4)	4.7 (1)
3	7.8 ($Q_{31}=1$)	8.3 ($Q_{32}=4$)	8.2 (3)	8.9 (5)	8.1 (2)
4	9.3 ($Q_{41}=3$)	11.1 ($Q_{42}=4$)	11.7 (5)	9.2 (2)	9.0 (1)
$Q_{\bullet j}$	1.75	4	4	3.75	1.5

$Q_{\bullet\bullet} = \frac{k+1}{2} = 3$ — общий средний ранг.

$$F = \frac{12\cdot 4}{5\cdot 6}\left[(1.75-3)^2+(4-3)^2+(4-3)^2+(3.75-3)^2+(1.5-3)^2\right] \approx 10.2.$$

$pvalue \approx 0.0372 \implies$ отвергаем гипотезу об одинаковом эффекте.

Критерий Пейджа [непараметрический]

$\mathrm{H}_0 : \beta_1 = ... = \beta_k$ *vs.* $\mathrm{H}_1 : \beta_1 \leqslant ... \leqslant \beta_k$.

Например, выборки отличаются интенсивностью стимулов

$\Longrightarrow$ стоит ожидать упорядоченный эффект

Q_{ij} — ранг наблюдения X_{ij} в наборе $(X_{i1}, ..., X_{ik})$

$T_j = \sum\limits_{i=1}^{n_j} Q_{ij}$ — сумма рангов выборки j

$L = \sum\limits_{j=1}^{k} jT_j$ — статистика критерия Пейджа

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

$\mathrm{E}L = \frac{nk(k+1)^2}{4}$, $\mathrm{D}L = \frac{n(k-1)k^2(k+1)^2}{144}$

Связные выборки

Post hoc анализ

Сравнение выборок

Пусть H_0 **отвергается**.

Далее проверяем гипотезы $\mathrm{H}_{rs}\colon \beta_r = \beta_s$ +МПГ

Если $|T_r - T_s| > q_{k,1-\alpha}\sqrt{nk(k+1)/12}$, то H_{rs} отвергается,

где $q_{k,p}$ — p-квантиль распределения нормального размаха,

т.е. величины $\xi_{(k)} - \xi_{(1)}$, где $\xi_1, ..., \xi_k \sim \mathcal{N}(0,1)$.

Оценка контраста

Пусть H_{rs} отвергается $\Longrightarrow$ оцениваем контраст $\Delta_{rs} = \beta_r - \beta_s$.

$Z_{rs} = med\{X_{ir} - X_{is}, i = 1..n\}$ — первичная оценка

$Z_{r\bullet} = \frac{1}{k} \sum\limits_{l=1}^{k} Z_{jl}$, где $Z_{jj} = 0$

$\widehat{\Delta} = Z_{r\bullet} - Z_{s\bullet}$ — уточненная оценка контраста

Свойства:

1. Первичные оценки могут быть несогласованными: $Z_{12} \neq Z_{13} + Z_{32}$
2. Уточненные оценки согласованы.
3. Уточненные оценки зависят от всех выборок.

Двухфакторный дисперсионный анализ

Два фактора

Два типа разбиения:

фактор $F_1 \in \{1, ..., k_1\}$;

фактор $F_2 \in \{1, ..., k_2\}$;

	$F_2 = 1$	$\dots$	$F_2 = j$	$\dots$	$F_2 = k_2$
$\dots$	$\dots$		$\dots$		$\dots$
$F_1 = i$	X_{i11} $\dots$ $X_{i1n_{i1}}$	$\dots$	X_{ij1} $\dots$ $X_{ijn_{ij}}$	$\dots$	X_{ik_21} $\dots$ $X_{ik_2n_{ik_2}}$
$\dots$	$\dots$		$\dots$		$\dots$
$F_1 = k_1$	X_{k_111} $\dots$ $X_{k_11n_{k_11}}$	$\dots$	X_{k_1j1} $\dots$ $X_{k_1jn_{k_1j}}$	$\dots$	$X_{k_1k_21}$ $\dots$ $X_{k_1k_2n_{k_1k_2}}$

Задача: влияют ли факторы на среднее значение?

Два фактора

$$X_{ijm} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijm},$$

$i = 1, ..., k_1$ — значения фактора F_1

$j = 1, ..., k_2$ — значения фактора F_2

$m = 1, ..., n_{k_1 k_2}$ — номер наблюдения в выборке

μ — неизвестное общее среднее

α_i — неизвестный эффект воздействия фактора $F_1 = i$

β_j — неизвестный эффект воздействия фактора $F_2 = j$

γ_{ij} — неизвестный эффект воздействия комбинации $F_1 = i, F_2 = j$

ε_{ijm} — случайная ошибка

Решение задачи дисперсионного анализа: функция `aov` в R.

Два фактора

	Доза 10 мг	Доза 20 мг
Препарат A	X_{111}, X_{112}, X_{113}, X_{114}, X_{115}, X_{116}	X_{121}, X_{122}, X_{123}
Препарат B	X_{211}, X_{212}, X_{213}, X_{214}	X_{221}, X_{222}, X_{223}, X_{224}, X_{225}

Что делать если есть еще контрольная группа?

- Двухфакторный без контрольной группы;
- Однофакторный по 5 группам;
- Два однофакторных по каждому препарату с тремя дозами (для контрольной 0 мг);


ВСЁ!