МСПС ANOVA I

Будем рассматривать теперь проверку гипотезы однородности (ну или более слабые гипотезы: отсутствие эффекта, $EX=0$ и тд.) для нескольких (больше 2) выборок.

Сначала рассмотрим случай нормальных выборок, и проверим сначала гипотезу о равенстве их дисперсий (как мы это делали в t-критерии Стьюдента)

$N = \sum_{j=1}^{k} n_j$

Пусть $X_{ij} \sim N(\mu_j, \sigma_j^2)$ $(i=1..n_j, j=1..k)$ – независ., μ_j и σ_j неизвестны

Обозначим $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$ – несмещ. оц. для μ_j;

$S_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2$ – несмещ. оц. для σ_j^2; $\overline{X} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}$.

Для проверки гипотезы $H': \sigma_1^2 = \dots = \sigma_k^2$ используют крит. Бартлетта:

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

Теорема Пусть H' верна и $n_j \geq 3$ $\forall j$. Тогда

$$B^* = \gamma^{-1}(N-k)\ln B \xrightarrow{d} \chi^2_{k-1}, \text{ где } \gamma = 1 + \frac{1}{3(k-1)}\left[\left(\sum_{j=1}^{k}\frac{1}{n_j}\right) - \frac{1}{N}\right].$$

Замечание Крит. Бартлетта крайне чувств. к отклонениям выборок от нормального распред.. Например, если при $k=10$ заменить $N(0,1)$ на t_7 (расп. Стьюдента), то фактич. уровень значимости критерия $\{B^* > u_{0,95}\}$ возрастет с 0.05 до 0.49.

Теперь проверим гипотезу $H'': \mu_1 = \mu_2 = \dots = \mu_k$ (при условии, что H' не отвергли)

Для этого используем критерий Фишера (однофактор. дисперс. анализа)

Введём $R = \dfrac{\frac{1}{k-1}\sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2}{\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2}$.

Если H_0: {все выборки о.р.} верна, то $R \sim F_{k-1, N-k}$, если $\forall n_j > 1$.

Замечание 1) Оказывается, критерий работает не только для нормальных выборок.

Если $N-k-1 \geq 20$ и отношение $\dfrac{\max_j S_j^2}{\min_j S_j^2} < 10$, то при верной H_0 $R \xrightarrow{d} F_{k-1, N-k}$ (т.е. можно приближением пользоваться)

2) При $n_1 = \dots = n_k$ метод устойчив к нарушению двух этих 2 предположений, если же объёмы выборок неравны, то то вер. ошибки I рода может возрасти при невыполнении условия про выборочные дисперсии.

3) Не устойчив к выбросам.

Ещё критерии проверки одинаковости дисперсий: тест Левена (Levene's test), тест Брауна-Форсайта (Brown-Forsythe test)

Фактич., мы рассматривали сейчас т.н. однофакторную модель $X_{ij} = \mu + \beta_j + \varepsilon_{ij}$, где μ – общее сред., β_j – эффект воздействия, ε_{ij} – сл. ошибка. Пусть теперь ε_{ij} независ. и одинаково распред.

Проверим гипотезу сдвига $H_0: \beta_1 = \beta_2 = \dots = \beta_k$ с помощью крит. Краскела – Уоллиса:

Пусть $R_{ij} = R(X_{ij})$ – ранг наблюд. X_{ij} в общей совокупности.

$\overline{R}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ij}$, $\overline{R} = \frac{\sum R_{ij}}{N} = \frac{N+1}{2}$.

Обозначим $W := (N-1)\,\frac{\sum_{j=1}^{k} n_j(\overline{R}_j - \overline{R})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(R_{ij} - \overline{R})^2} = \frac{12}{N(N+1)}\sum_{j=1}^{k} n_j R_j^2 - 3(N+1)$.

(Оказывается, знаменатель равен $(N^3-N)/12$)

При $n_j > 5$ можно получ. аппроксимацией $W \sim \chi^2_{k-1}$, но при малых n_j типа 4-6 лучше брать поправку Имана

$\widetilde{W} = \frac{1}{2} W\left[\frac{N-k}{N-1-W}+1\right]$, приближ. критич. значения $\widetilde{w}_{1-\alpha} = \frac{1}{2}\left[x_{1-\alpha} + y_{1-\alpha}\right]$, где x – χ^2_{k-1}, y – $F_{k-1,N-k}$.

Гипотезу сдвига проверяет и критерий Джонкхира (против $H_A: \beta_1 \le \beta_2 \le \dots \le \beta_k$)

$S = \sum_{j=1}^{k}\sum_{i=1}^{n_j} a_{ij}$, где a_{ij} – число наблюд. из первых $j-1$ выборок, $< X_{ij}$.

S имеет табл. распред., при $n_j > 10$ можно получ. нормальным приближением: $S \sim N(\mu, \sigma^2)$, где $\mu = \frac{1}{4}\left(N^2 - \sum_{j=1}^{k} n_j^2\right)$,

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

Оба этих теста устойчивы к выбросам.

Джонкхир мощнее крит. Краскела-Уол. на альтернативе H_A.

Если у нас гипотеза однородности отвергнута, то появляется естеств. желание узнать, какие пары (группы) выборок неоднородны. Это делается (в основном) с помощью двухвыборочных критериев, пройденных на пред. лекции. А также см. стр. 3.

Двухфактор. дисперсионный анализ

Данные представлены в виде $X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, $i = 1..n$, $j = 1..k$.

Эта пар-зация избыточна и не позволяет однозн. восст. μ, α и β.

Пример

Блоки	Обработки 1	2	...	k
1	X_{11}	X_{12}		X_{1k}
2	X_{21}	X_{22}		X_{2k}
⋮				
n	X_{n1}	X_{n2}		X_{nk}

α_i – эффект блока

β_j – эффект обработки (он нас и интересует)

ε_{ij} – н.о.р. с непрер. распред., $E\varepsilon_{ij} = 0$.

Проверяем гипотезу $H_0: \beta_1 = \dots = \beta_k$ против альтерн. H_1: не все равны.

Критерий Фридмана

Его статистика $F = \frac{12}{nk(k+1)} \sum_{j=1}^{k} T_j^2 - 3n(k+1) \left(= \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(\overline{R}_{\cdot j} - \overline{R} \right)^2 \right)$

где R_{ij} = ранг X_{ij} в строке i, $\overline{R}_j = \frac{1}{n} \sum_{i=1}^{n} R_{ij}$; $\overline{R} = \frac{k+1}{2}$ – сред. ранг по матрице; $T_j = n \overline{R}_j$.

1) При $n > 15$, $k > 4$ $F \sim \chi^2_{k-1}$.

2) Более точная: $\frac{(n-1)F}{n(k-1) - F} \sim F\left(n-1, (n-1)(k-1)\right)$.

Стоит ещё обратить внимание на крит. Пейджа, он проверяет гипотезу $H_0: \beta_1 = \dots = \beta_k$ против альтерн. $H_1: \beta_1 \le \dots \le \beta_k$ (т.е. для такой альтернативы он является состоятельным)

Статистика: $L = \sum_{j=1}^{k} j R_j$.

Аппроксимация: $L^* = \frac{L - nk(k+1)^2/4}{\sqrt{\frac{1}{144(k-1)} n (k^3-k)^2}} \sim N(0,1)$, приближением можно польз. при $n > 15$, $k > 10$.

Замечание Критерии для проверки неоднородности 2 выборок в задаче однофакт. дисперсионного анализа: критерий Шеффе (для норман. выборок), критерии Даннета, LSD Фишера, MSD Тьюки, Неменьи (см. презентацию Рябенко).

Наконец, мы ещё не говорили о ситуации, когда выборки у нас зависимые. Моделируется эта ситуация так:

$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, где μ – "глобальное" среднее признака X,

α_i – отклонение от μ, вызванное влиянием i-того объекта

β_j – отклонение от $\mu + \alpha_i$, вызван. влиянием фактора f_j

ε_{ij} – н.о.р. сл. ошибки

Объект \ Фактор	1	j	k
1	X_{11}		X_{1k}
2			
i			
n	X_{n1}		X_{nk}

Хотим понять, влияют ли факторы $f_1 - f_k$ на средние объектов $X_{\cdot j}$.

Как легко видеть, это очень похоже на двухфакторный дисперсионный анализ. Считать, что выборки связаны как-то по-другому, довольно сложно. Т.е. можно пользоваться крит. Фридмана и Пейджа для проверки гипотезы H_0: $\beta_1 = \dots = \beta_k$, а также критерием Фишера: (против альтерн. H_1: H_0 неверна)

$$F = \frac{n \cdot \sum_{j=1}^{k} \left(\overline{X}_{\cdot j} - \overline{X} \right)^2 \cdot (n-1)}{\sum_{i=1}^{n} \sum_{j=1}^{k} \left(X_{ij} - \overline{X} \right)^2 - k \sum_{i=1}^{n} \left(\overline{X}_i - \overline{X} \right)^2}$$

, где $\overline{X}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$; $\overline{X}_i = \frac{1}{k} \sum_{j=1}^{k} X_{ij}$; $\overline{X} = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij}$

При верной H_0 $F \sim F\left(k-1, (n-1)(k-1)\right)$

(Лучше всего работает в предположении о нормальности данных и "сферичности" или "скрытой симметрии": $\forall j$ $X_{ij} - X_{kj}$ имеют одинак. дисперсию $\forall i \neq k$.