

Nieparametryczna ANOVA

Jeżeli z pewnych względów **założenie normalności błędów** w modelu ANOVA efektów stałych **jest nie do przyjęcia**, to można zbudować ogólniejszy model nie korzystający z tych krępujących założeń. Rozważmy pewien nieparametryczny odpowiednik jednoczynnikowej ANOVA.

Oznaczmy przez $\mathcal{F}_{ac}$ nieparametryczną rodzinę absolutnie ciągłych dystrybuant na prostej R

Mamy więc model

$X_{11}, \dots, X_{1n_1}$ próba prosta z rozkładu z nieznaną absolutnie ciągłą dystrybuantą F_1

$\vdots$

$X_{k1}, \dots, X_{kn_k}$ próba prosta z rozkładu z nieznaną absolutnie ciągłą dystrybuantą F_k .

Chcemy testować hipotezę

$H_0 : F_1 = \dots = F_k$ przeciwko alternatywie $H_1 : \exists i, j : F_i \neq F_j$.

Szczególnym przypadkiem powyższego modelu jest **następujący model nieparametryczny**

$$X_{ij} = m_i + \varepsilon_{ij} \quad i=1, \dots, k; j=1, \dots, n_i$$

$$\varepsilon_{ij} \text{ iid } F(x) \in \mathcal{F}_{ac} \quad (E(\varepsilon_{ij}) = 0, V(\varepsilon_{ij}) = \sigma^2)$$

Oczywiście $X_{i1}, \dots, X_{in_i}$ jest próbą prostą z rozkładu o dystrybuancie $F(x - m_i)$

W tym modelu testujemy hipotezę

$H_0 : m_1 = \dots = m_k$ przeciwko alternatywie $H_1 : \exists i, j : m_i \neq m_j$.

Powyższy problem przypomina problem rozważany wcześniej w 1 czynnikowej ANOVA efektów stałych. **Inna jest jednak rodzina rozkładów na stosownej przestrzeni prób.** W tym przypadku mamy do czynienia z **nieparametryczną rodziną rozkładów** parametryzowaną parametrem $(m_1, \dots, m_k, F) \in R^k \times \mathcal{F}_{ac}$.

Jak testować H_0 ?

Pomysł: Zmodyfikować rozwiązanie uzyskane w odpowiednim modelu parametrycznym tak, aby przynajmniej przy prawdziwości H_0 rozkład zmodyfikowanej statystyki testowej nie zależał od nieznaney dystrybuanty $F(x) \in \mathcal{F}_{ac}$.

Fakty

- Jeżeli $X = (X_1, \dots, X_n)$ jest próbą prostą z rozkładu o absolutnie ciągłej dystrybuancie $F \in \mathcal{F}_{ac}$, to **wektor statystyk pozycyjnych $X^* = (X_{(1)}, \dots, X_{(n)})$ jest statystyką dostateczną zupełną** (więc także minimalną dostateczną).
Przejście od próby X do X^* redukuje próbę bez straty informacji lecz rozkład X^* w dalszym ciągu zależy od nieznaney dystrybuanty F .

- Za pomocą ściśle rosnącej transformacji można dowolny rozkład ciągły na prostej przekształcić na dowolny inny ciągły rozkład na prostej. Jeżeli zmienna losowa X ma pewien rozkład $F \in \mathcal{F}_{ac}$, to zmienna $F(X)$ ma rozkład jednostajny $U[0,1]$. Jeżeli $G \in \mathcal{F}_{ac}$ jest zadaną dystrybucją na prostej, to $G^{-1}(F(X))$ ma rozkład o dystrybucji G a funkcja złożona $G^{-1}F$ jest ściśle monotoniczna
- Aby procedura testowa nie zależała do nieznannej dystrybucji $F \in \mathcal{F}_{ac}$ "powinna" być ona niezmiennicza względem grupy rosnących bijekcji $f: R \rightarrow R$
- Maksymalnym niezmiennikiem względem grupy rosnących bijekcji jest wektor rang $(R_1, \dots, R_n)$ gdzie R_i = miejsce (ranga) obserwacji X_i w uporządkowanym rosnąco wektorze $X^* = (X_{(1)}, \dots, X_{(n)})$

Wniosek. Test H_0 należy oprzeć nie na surowych obserwacjach lecz na ich rangach

Test ANOVA Kruskala Wallisa (1952)

Uwaga. Dla rozkładów ciągłych z prawdopodobieństwem 1 wszystkie obserwacje są różne więc i rangi są różne.

Niech R_{ij} będzie **rangą obserwacji X_{ij} w połączonej próbie** $X_{11}, \dots, X_{1n_1}, \dots, X_{k1}, \dots, X_{kn_k}$.

Oznaczenia:

$$n = \sum_{i=1}^k n_i$$

$$R_{i+} = \sum_{j=1}^{n_i} R_{ij} \quad - \text{suma rang w } i\text{-tej grupie}$$

$$\bar{R}_{i\bullet} = \frac{1}{n_i} R_{i+} \quad - \text{średnia ranga w } i\text{-tej grupie}$$

$$R_{++} = \sum_{i=1}^k R_{i+} = 1 + \dots + n = \frac{n(n+1)}{2} \quad - \text{suma wszystkich rang}$$

$$\bar{R}_{\bullet\bullet} = \frac{1}{n} R_{++} = \frac{n+1}{2} \quad \text{średnia ranga w próbie } n \text{ elementowej}$$

Statystyka

$$K = \frac{12}{n(n+1)} \sum_{i=1}^k n_i (\bar{R}_{i\bullet} - \bar{R}_{\bullet\bullet})^2 = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{(R_{i+} - n_i \bar{R}_{\bullet\bullet})^2}{n_i} = \frac{6}{n} \sum_{i=1}^k \frac{(R_{i+} - n_i \bar{R}_{\bullet\bullet})^2}{n_i \bar{R}_{\bullet\bullet}}$$

jest miarą typu χ^2 zróżnicowania średnich rang w grupach. **Duże wartości statystyki K świadczą przeciwko hipotezie H_0 .**

Przy prawdziwości hipotezy H_0 każdy układ rang ma takie samo prawdopodobieństwo $\frac{1}{n!}$. Każdy z

$\frac{n!}{n_1! \dots n_k!}$ układów tych rang prowadzi do tego samego układu sum rang $R_{1+}, \dots, R_{k+}$. Metodami

kombinatorycznymi można więc wyznaczyć k_{kryt} , aby $P_{H_0}(K > k_{kryt}) \leq \alpha$ (patrz Zieliński - Tablice...)

Dla $n_i \geq 5$ rozkład statystyki K można przybliżać rozkładem χ^2_{k-1} .

Uwaga. Jeżeli powodu np. małej precyzji pomiaru otrzymujemy obserwacje o tych samych wartościach, to przypisujemy im średnią rangę- mówimy wówczas o rangach związanych. W takiej sytuacji statystyka K zostaje zastąpiona statystyką K_T z poprawką na rangi związane

$$K_T = \frac{K}{1 - \frac{\sum_{i=1}^r (S_i^3 - S_i)}{n^3 - n}}, \text{ gdzie } S_i - \text{ilość obserwacji z } i\text{-tą rangą związaną (} i\text{-te miejsce ex aequo)}$$

Można pokazać, że statystyka K jest związana ze statystyką

$$A = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{R}_{i\bullet} - \bar{R}_{\bullet\bullet})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (R_{ij} - \bar{R}_{i\bullet})^2} \quad (\text{ANOVA wykonana na rangach}) \text{ wzorami}$$

$$A = \frac{n-k}{n-1} \frac{K}{(n-1) - K} \quad ; \quad K = \frac{(n-1)(k-1)A}{(n-k) + (k-1)A}.$$

Test Kruskala-Wallisa jest równoważny testowi ANOVA wykonanemu na rangach

Uwaga. O.J. Dunn zaproponowała asymptotyczną procedurę **porównań jednoczesnych** typu Bonferroni, która kontroluje błąd dla wszystkich porównań parami i procedurę porównań jednoczesnych z wyróżnioną grupą kontrolną (szczegóły np., Rohatgi i Woolson)

Dla porównań parami k grup (czyli jednocześnie $\binom{k}{2}$ porównań) proponowany przez O.J. Dunn test każe uznać grupy i -tą oraz j -tą za istotnie różne, gdy

$$|\bar{R}_{i\bullet} - \bar{R}_{j\bullet}| > z_{1-\frac{\alpha}{k(k-1)}} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}; \quad 1 \leq i < j \leq k,$$

gdzie $z_{1-\frac{\alpha}{k(k-1)}}$ jest kwantylem rzędu $1 - \frac{\alpha}{k(k-1)}$ rozkładu $N(0,1)$.

Przy prawdziwości H_0 prawdopodobieństwo zaobserwowania fałszywie istotnej różnicy jest równe α .

Przy porównaniach $k-1$ grup z wyróżnioną grupą kontrolną (o numerze k) mamy regułę

Grupa i -ta istotnie różni się od grupy kontrolnej gdy

$$|\bar{R}_{i\bullet} - \bar{R}_{k\bullet}| > z_{1-\frac{\alpha}{2(k-1)}} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_k} \right)}; \quad 1 \leq i < k-1,$$

Przy prawdziwości H_0 prawdopodobieństwo zaobserwowania fałszywie istotnej różnicy jest równe α .

Nieparametryczna 2 czynnikowa ANOVA Friedmana.

Model

- $X_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij}$, $i=1, \dots, r, j=1, \dots, c$
- $\sum_{i=1}^r \alpha_i = 0$, $\sum_{j=1}^c \beta_j = 0$
- ε_{ij} iid o pewnej ciągłej dystrybucji $F \in \mathcal{F}_{ac}$ i $E(\varepsilon_{ij}) = 0$, $V(\varepsilon_{ij}) = \sigma^2$

Jest to oczywiście model nieparametryczny, gdyż rodzina rozkładów na przestrzeni prób jest rodziną nieparametryczną. Jest to nieparametryczny odpowiednik 2 czynnikowej ANOVA efektów głównych (bez interakcji- model addytywny) z 1 obserwacją w każdej klatce (pomijamy więc indeks numeru obserwacji w klatce). Do takiego modelu prowadzi plan zrandomizowany w układzie blokowym lub jednoczynnikowa ANOVA w układzie z powtarzalnymi pomiarami. Zestawiając obserwacje X_{ij} w macierz, wiersze traktujemy jako **jednorodne bloki** a kolumny jako **zabiegi**

Testujemy hipotezę

$H_0: \beta_1 = \dots = \beta_c$ (zabiegi nie różnią się między sobą) przeciwko alternatywie $H_1: \sim H_0$

Procedura testowania powinna być niezmiennicza ze względu na transformacje zachowujące porządek (w jednorodnych blokach) Jednorodność bloku zapewniamy poprzez zastosowanie dla każdego bloku porządkowej skali o c kategoriach.

Rangujemy więc obserwacje w każdym z bloków (wierszy) z osobna i oznaczamy

$R_{ij} = R(X_{ij})$ - ranga j -tej obserwacji w i -tym bloku (stad $1 \leq R_{ij} \leq c$)

Oznaczmy (dla $j=1, \dots, c$) $R_{+j} = \sum_{i=1}^r R_{ij}$ - suma rang dla j (j -tego zabiegu, j -tej kolumny)

Oczywiście $\sum_j R_{+j} = \frac{c(c+1)}{2}$. Jeżeli prawdziwa jest hipoteza H_0 , to $E(R_{+j}) = \frac{1}{c} \frac{rc(c+1)}{2} = \frac{r(c+1)}{2}$

Statystyka $S = \sum_{j=1}^c (R_{+j} - E(R_{+j}))^2$ mierzy różnicę pomiędzy kolumnami (zabiegami)

Statystyka F Friedmana oparta na S ma postać

$$F = \frac{12S}{rc(c+1)} = \frac{12}{rc(c+1)} \sum_{j=1}^c R_{+j}^2 - 3r(c+1)$$

i jej rozkład przy prawdziwości H_0 nie zależy od nieznanej dystrybucji $F \in \mathcal{F}_{ac}$ jest wyznaczony metodami kombinatorycznymi. Punktem wyjścia jest fakt, że przy prawdziwości H_0 **każdy z $(c!)^r$ układów rang w blokach jest jednakowo prawdopodobny**. Ponadto znany jest rozkład

asymptotyczny: dla $r \rightarrow \infty$ F ma rozkład χ_{c-1}^2 . Dla $c < 6$ są tablice rozkładu F dla małych r . Zobacz (

Zieliński R., Zieliński W. Tablice statystyczne str. 406 -Uwaga u Zielińskich obserwacje są rangowane w kolumnach -macierz obserwacji jest więc transpozycją macierzy tu rozważanej))
Jeżeli **zabiegi** istotnie różnią się między sobą , to dlatego że rangowanie w poszczególnych blokach

było podobne. Miara tego podobieństwa jest **Kendalla współczynnik zgodności**. $W = \frac{F}{r(c-1)}$,

który przyjmuje wartości z przedziału $[0, 1]$ (sprawdzić). Jeżeli rangowanie w każdym bloku było identyczne (pełna zgodność), to sumy rang $R_{+1}, \dots, R_{+c}$ są permutacjami liczb $r, 2r, \dots, cr$ i $W=1$.

Podobnie jeżeli rangowania w blokach są niezależne to sumy rang $R_{+1}, \dots, R_{+c}$ są sobie bliskie i mogą być nawet identyczne. Wówczas współczynnik zgodności $W=0$.

Podobnie jak w przypadku ANOVA Kruskala-Wallisa znane są testy **porównań wielokrotnych** (zobacz Woolson R.F. Statistical methods for the analysis of biomedical data , Wiley, str. 378 i 379).

Grupy i -tą i j -tą (i -ty i j -ty zabieg) uznajemy za istotnie różne jeżeli

$$|\bar{R}_{\bullet i} - \bar{R}_{\bullet j}| > q_{1-\alpha}(c) \sqrt{\frac{c(c+1)}{12r}} ; \quad 1 \leq i < j \leq c ,$$

gdzie $q_{1-\alpha}(c)$ jest kwantylem odpowiedniego rozkładu (zobacz Woolson - tablica 15).

Prawdopodobieństwo przy H_0 zaobserwowania fałszywie istotnej różnicy łącznie w $\binom{c}{2}$ porównaniach jest równe α .

Dla porównań z wyróżnioną grupą kontrolną (o numerze c) reguła jest podobna

$$|\bar{R}_{\bullet i} - \bar{R}_{\bullet c}| > q_{1-\alpha}(c) \sqrt{\frac{c(c+1)}{6r}} ; \quad 1 \leq i \leq c-1 ,$$

przy czym $q_{1-\alpha}(c)$ jest kwantylem stosownego (innego niż poprzednio) tablicowanego rozkładu (zobacz Woolson - tablica 16).

Uwaga. Test Friedmana może być także użyty do testowania jednorodności rozkładów w próbach zależnych pochodzących z $c > 2$ wymiarowego rozkładu ciągłego. Mianowicie niech $\mathbf{X} = (X_1, \dots, X_c)$ będzie wektorem (wierszowym) z c -wymiarowego rozkładu ciągłego , przy czym składowe $X_1, \dots, X_c$ mają odpowiednio rozkłady brzegowe o dystrybuantach $F_1, \dots, F_c$.

Na podstawie r elementowej próby prostej

$$\mathbf{X}_1 = (X_{11}, \dots, X_{1c})$$

.....

$$\mathbf{X}_r = (X_{r1}, \dots, X_{rc})$$

z rozpatrywanego rozkładu zweryfikować hipotezę

$$H_0: F_1 = \dots = F_c \text{ wobec alternatywy } H_1: \exists i, j \ X_i \leq^{\text{st}} X_j \text{ (tzn. } F_i \geq F_j \text{)}.$$