RPiS

Analiza wariancji – ANalysis Of VAriance – ANOVA

Załóżmy, że dane są niezależne obserwacje zmiennej losowej $X \sim N(\mu, \sigma^2)$. Wyniki obserwacji grupujemy względem pewnej cechy jakościowej, wyróżniamy I grup. Dla każdej grupy mamy J obserwacji. Symbolem x_{ij} oznaczamy j-tą obserwację w i-tej grupie ($i = 1, \ldots, I$; $j = 1, \ldots, J$).

Grupa	Obserwacje				Średnie
1	x_{11}	x_{12}	$\ldots$	x_{1J}	$x_{1\bullet}$
2	x_{11}	x_{12}	$\ldots$	x_{1J}	$x_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
I	x_{I1}	x_{I2}	$\ldots$	x_{IJ}	$x_{I\bullet}$

Ostatnia kolumna powyższej tabeli zawiera średnie grup (wierszy). tzn. $x_{k\bullet} = \frac{1}{J}\sum_{j=1}^{J} x_{kj}$. Symbolem $\bar{x}$ oznaczamy średnią wszystkich obserwacji: $\bar{x} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}$. Przykłady grupowania danych: 3 grupy opon (zimowe, letnie, uniwersalne) i notujemy stopień zużycia po określonym przebiegu; skuteczność pewnego leku w grupach: początkowe stadium choroby, choroba w pełnym objawie, stan ciężki; porównujemy podobne leki od 3 producentów itp.

Zakładamy – zgodnie z założeniem początkowym, – że każda ze zmiennych losowych $X_{i\bullet}$ ma rozkład $N\left(\mu, {}^{\sigma^2}\!/_n\right)$. W istocie chcemy zaprzeczyć temu założeniu, to znaczy wywnioskować z danych iż jedna z grup (niektóre, wiele) jest różna od pozostałych. Kolejne obliczenia powinny wskazać, która grupa odróżnia się "na plus", ale to na razie odkładamy na później. To co nas interesuje, to odpowiedź w postaci: opony A są lepsze od pozostałych; pewne lekarstwo najbardziej nadaje się do któregoś stadium choroby; producent C ma najlepszy produkt.

Interesującym faktem jest iż możemy powiedzieć coś o średnich grup ($x_{i\bullet}$) na podstawie wariancji wszystkich obserwacji oraz wariancji wewnątrz grup (wierszy). Rozpocznijmy od wariancji wszystkich obserwacji. Wielokrotnie stosowaliśmy wzór

$$
\begin{array}{ccccc|c}
\sum_{k=1}^{n}(X_k-\mu)^2 & = & \sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2 & + & n\cdot\left(\bar{X}-\mu\right)^2 & \text{wzór} \\
\frac{nS_\mu^2}{\sigma^2} & = & \frac{nS^2}{\sigma^2} & + & \left(\frac{\bar{X}-\mu}{\sigma\sqrt{n}}\right)^2 & \text{skróty} \\
\chi^2(n) & = & \chi^2(n-1) & + & \chi^2(1) & \text{rozkłady}
\end{array}
\tag{1}
$$

Podstawowy fakt jest taki: dla obserwacji x_{ij} zmienna losowa $\frac{nS^2}{\sigma^2} = \sum_{ij}\left(X_{ij}-\bar{X}\right)^2$ ma rozkład $\chi^2(IJ-1)$, ponieważ mamy I grup po J obserwacji oraz "znika" jeden stopień swobody, jak wynika ze wzoru (1).
Zróżnicowanie obserwacji przedstawiamy jako $\mathrm{SS_{Tot}} = \sum_{i,j}(x_{ij}-\bar{x})^2$. Z dokładnością do stałej jest to wariancja wszystkich obserwacji.[1] Jest zatem: $\frac{\mathrm{SS_{Tot}}}{\sigma^2} \sim \chi^2(IJ-1)$.
Zróżnicowanie grup (**zmienność międzygrupową**) można wyrazić poprzez średnie grup $x_{i\bullet}$. Ponieważ zmienne X_{ij} są niezależne, więc niezależne są też zmienne $X_{i\bullet}$. Mamy zatem I niezależnych

[1] SS ≡ sum of squares.

zmiennych losowych $X_{1\bullet},\dots,X_{I\bullet}$. Średnia $\bar{X}$ wszystkich obserwacji jest równocześnie średnią wziętą ze średnich poszczególnych grup.[2] Traktując grupę jako "uogólnioną obserwację" stwierdzamy iż wyrażenie $\mathrm{SSA} = J\cdot\sum_{i=1}^{I}(x_{i\bullet}-\bar{x})^2$ - z dokładnością do stałej – ma rozkład $\chi^2(I-1)$. Do rozpatrzenia pozostaje drugi składnik zmienności, wielkość $\mathrm{SSE} = \sum_{i,j}(x_{ij}-x_{i\bullet})^2$, nazywana **zmiennością wewnątrzgrupową**.

Twierdzenie 1.

$$\mathrm{SS_{Tot}} = \mathrm{SSA} + \mathrm{SSE}. \tag{2}$$

Komentarze:

- Teza twierdzenia to: wariancja całkowita dzieli się na sumę wariancji pomiędzy grupami i wariancji wewnątrz grup.
- Jeżeli większość wariancji znajduje się wewnątrz grup, to skłonni jesteśmy uznać, że średnie grup są takie same (albo zbliżone do siebie).
- Na odwrót: jeżeli wariancja między grupami przeważa nad wariancją wewnątrz grup to można sądzić, że średnie grup różnią się.
- W podsumowaniu: na podstawie wariancji (a raczej jej podziału na dwa składniki) można wyciągnąć wnioski o średnich w obrębie grup.

Dowód.

$$\begin{aligned}\mathrm{SS_{Tot}} &= \sum_{i,j}(x_{ij}-\bar{x})^2 = \sum_{i,j}(x_{ij}-x_{i.\bullet}+x_{i\bullet}-\bar{x})^2 = \\ &= J\cdot\sum_{i}(x_{i\bullet}-\bar{x})^2 + \sum_{i,j}(x_{ij}-x_{i.\bullet})^2 + 2\cdot\sum_{i,j}(x_{ij}-x_{i.\bullet})\cdot(x_{i\bullet}-\bar{x}).\end{aligned}$$

Trzeci składnik w ostatniej równości można przekształcić do postaci

$$\begin{aligned}&\sum_{i,j}(x_{ij}-x_{i.\bullet})\cdot(x_{i\bullet}-\bar{x}) = \sum_{i}(x_{i\bullet}-\bar{x})\sum_{j}\cdot(x_{ij}-x_{i.\bullet}) = \\ &= (x_{i\bullet}-\bar{x})\cdot(n\cdot x_{i\bullet}-n\cdot x_{i\bullet}) = 0.\end{aligned}$$

Stąd wynika już, że

$$\mathrm{SS_{Tot}} = \sum_{i,j}(x_{ij}-\bar{x})^2 = J\cdot(x_{i\bullet j}-\bar{x})^2\sum_{i,j}+(x_{ij}-x_{i.\bullet})^2 = \mathrm{SSA}\ +\ \mathrm{SSE}. \tag{3}$$

□

2-czynnikowa ANOVA

Załóżmy, że dane są niezależne obserwacje zmiennej losowej $X\sim N(\mu,\sigma^2)$. Wyniki obserwacji grupujemy względem cechy jakościowej (czynnika) A oraz cechy jakościowej B, wyróżniamy odpowiednio I oraz J grup. Dla każdej kombinacji grup mamy jedną obserwację.[3] Symbolem x_{ij} oznaczamy j-tą obserwację dla której cecha A przyjęła i-tą wartość, natomiast cecha B wartość j-tą ($i=1,\dots,I;\ \ j=1,\dots,J$).

[2] Wszystkie grupy mają tę samą liczbę obserwacji.

[3] Mówimy wówczas o 2-czynnikowej analizie ANOVA bez powtórzeń.

Grupa	1	2	...	J	Średnie
1	x_{11}	x_{12}	...	x_{1J}	$x_{1\bullet}$
2	x_{11}	x_{12}	...	x_{1J}	$x_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
I	x_{I1}	x_{I2}	...	x_{IJ}	$x_{I\bullet}$
Średnie	$x_{\bullet 1}$	$x_{\bullet 2}$	...	$x_{\bullet J}$	

Symbole $x_{i\bullet}, x_{\bullet j}$ oznaczają – odpowiednio – średnią wartość i-tej grupy czynnika A oraz średnią wartość j-tej grupy czynnika B. Symbol $\bar{x}$ oznacza średnią wszystkich obserwacji. Niech ponadto

$$\begin{aligned} \text{SSTot} &= \sum_{ij} (x_{ij} - \bar{x})^2, \quad & \text{SSA} &= J \cdot \sum_i (x_{i\bullet} - \bar{x})^2 \\ \text{SSB} &= I \cdot \sum_j (x_{\bullet j} - \bar{x})^2, \quad & \text{SSE} &= \sum_{ij} (x_{ij} - x_{i\bullet} - x_{\bullet j} + \bar{x})^2. \end{aligned} \tag{4}$$

Twierdzenie 2.

$$\text{SSTot} = \text{SSA} + \text{SSB} + \text{SSE}. \tag{5}$$

Dowód. Zauważmy, że $\text{SSTot} = \sum_{ij} (x_{ij} - \bar{x})^2 = \sum_{ij} \left(\underbrace{x_{ij} - x_{i\bullet} - x_{\bullet j} + \bar{x}}_{(a)} + \overbrace{x_{i\bullet} - \bar{x}}^{(b)} + \underbrace{x_{\bullet j} - \bar{x}}_{(c)} \right)^2.$

Zauważmy, że sumowanie kwadratów wyrażeń oznaczonych jako $(a), (b), (c)$ daje składniki SSA, SSB, SSE prawej strony równania (5). Pozostaje zatem do wykazania, że sumowanie iloczynów $(a) \cdot (b), (a) \cdot (c), (b) \cdot (c)$ daje w wyniku 0.
$(b) \cdot (c) = \sum_{ij} (x_{i\bullet} - \bar{x}) \cdot (x_{\bullet j} - \bar{x}) = \sum_i (x_{i\bullet} - \bar{x}) \cdot \sum_j (x_{\bullet j} - \bar{x}) = 0$, bo $\sum_i (x_{i\bullet} - \bar{x}) = I \cdot \bar{x} - I \cdot \bar{x} = 0$.
$(a) \cdot (b) = \sum_{ij} (x_{i\bullet} - \bar{x}) \cdot (x_{ij} - x_{i\bullet} - x_{\bullet j} + \bar{x}) = \sum_i (x_{i\bullet} - \bar{x}) \cdot \sum_j (x_{ij} - x_{i\bullet} - x_{\bullet j} + \bar{x}) = (*)$
Dla ustalonego i rozpatrzmy wewnętrzne sumowanie po j daje $\sum_j (x_{ij} - x_{i\bullet}) = 0$. Stąd

$$(*) = \sum_i (x_{i\bullet} - \bar{x}) \cdot \sum_j (\bar{x} - x_{\bullet j}) = 0.$$

Dowód dla sumowania iloczynów postaci $(a) \cdot (c)$ jest praktycznie taki sam, wystarczy zamienić miejscami indeksy i, j. □

[Popularne|Ulubione] wzory i rozkłady – powtórzenie

1. Załóżmy, że zmienne losowe X, Y są niezależne i podlegają rozkładom $X \sim \chi^2(n)$, $Y \sim \chi^2(k)$. Wówczas zmienna losowa $Z = X + Y$ podlega rozkładowi $Z \sim \chi^2(n+k)$.
2. Załóżmy, że zmienna X podlega rozkładowi $\mathrm{N}(\mu, \sigma^2)$. Niech dodatkowo $Y = \dfrac{X - \mu}{\sigma}$. Zachodzi Fakt: $X \sim \mathrm{N}(\mu, \sigma^2) \iff Y \sim \mathrm{N}(0,1)$.
3. $\text{Gamma}\,(1/2, n/2) \equiv \chi^2(n)$.
4. Załóżmy, że zmienne $X_1, \ldots, X_n$ są niezależne i podlegają rozkładowi $\mathrm{N}(\mu, \sigma^2)$ każda. Wówczas zmienna $Z = \sum_{k=1}^{n} \left(\dfrac{X_k - \mu}{\sigma} \right)^2$ ma rozkład $\chi^2(n)$.
5. Niezależne zmienne X, Y mają rozkłady $X \sim \chi^2(k)$, $Y \sim \chi^2(l)$ odpowiednio. Mówimy, że zmienna $F(k,l) = \dfrac{X}{Y} \cdot \dfrac{l}{k}$ ma rozkład F-Fishera z (k,l) stopniami swobody.
6. Niezależne zmienne X, Y mają rozkłady $X \sim \mathrm{N}(0,1)$, $Y \sim \chi^2(k)$ odpowiednio. Mówimy, że zmienna $t(k) = \dfrac{X}{\sqrt{Y/k}}$ ma rozkład t-Studenta z k stopniami swobody.

7. Intuicja: iloraz niezależnych i normalizowanych rozkładów χ^2 to rozkład F-Fishera zaś iloraz standardowego rozkładu normalnego i pierwiastka normalizowanego rozkładu χ^2 to rozkład t-Studenta.
8. Załóżmy, że zmienne $X_1, \ldots, X_n$ są niezależne i podlegają rozkładowi $\mathrm{N}(\mu, \sigma^2)$ każda. Niech dodatkowo $S_\mu^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$. Wówczas $\frac{nS_\mu^2}{\sigma^2} \sim \chi^2(n)$.
9. Załóżmy, że zmienne $X_1, \ldots, X_n$ są niezależne i podlegają rozkładowi $\mathrm{N}(\mu, \sigma^2)$ każda. Niech dodatkowo $S^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$. Wówczas $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$.
10. . . . inne, jeszcze ciekawsze.

Witold Karczewski