7 ANOVA

Jednoczynnikowa ANOVA

Na test jednoczynnikowej analizy wariancji możemy patrzeć jak na uogólnienie testu t Studenta dla dwóch prób niezależnych, na przypadek k, $(k > 2)$ prób niezależnych.

Model

$$X_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

gdzie

μ_i - “prawdziwa” wartość badanej cechy i i-tej grupie,

ε_{ij} - błędy (niezależne zmienne losowe o jednakowym rozkładzie $N(0, \sigma^2)$).

Hipoteza zerowa: wartości oczekiwane (średnie) badanej cechy w k grupach nie różnią się istotnie:

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_k.$$

Hipoteza alternatywna: co najmniej dla jednej pary grup, wartości oczekiwane (średnie) badanej cechy różnią się istotnie:

$$H_1:\ \neg H_0.$$

Statystyka testowa:

$$F = \frac{SSA}{k-1} \Big/ \frac{SSE}{n-k},$$

gdzie

$$SSA = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2, \quad SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2,$$

$$\bar{X}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},\ \bar{X}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij},\ n = \sum_{i=1}^{k} n_i.$$

Rozkład statystyki testowej:

$$F|_{H_0} \sim F(k-1, n-k)$$

Tabela analizy wariancji

Tradycyjnie wyniki analizy wariancji przedstawiamy w postaci tabeli.

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średnie kwadraty	Statystyka testowa
Pomiędzy grupami	SSA	$k-1$	MSA	F
Wewnątrz grup	SSE	$n-k$	MSE	
Całość	SST	$n-1$	MST	

$$MSA = SSA/(k-1),\ MSE = SSE/(n-k),\ MST = SST/(n-1).$$

Założenia jednoczynnikowej analizy wariancji

1. Niezależność obserwacji dla poszczególnych jednostek eksperymentalnych.
2. Błędy mają rozkłady normalne z zerową wartością oczekiwaną (brak błędu systematycznego) i jednorodną wariancją.

Uwaga: Założenie jednorodności wariancji błędów możemy zweryfikować testem Bartletta.

Test Bartletta

Założenia: Model normalny, wiele prób niezależnych.

Hipoteza zerowa:

$$H_0:\ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

Hipoteza alternatywna:

$$H_1:\ \neg H_0$$

Statystyka testowa:

$$B = \frac{1}{C}(n-k)\ln MSE - \sum_{i=1}^{k}(n_i-1)\ln S_i^2,$$

gdzie

$$C = 1 + \frac{1}{3(k-1)} \sum_{i-1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\ .$$

Rozkład statystyki testowej:

$$B|_{H_0} \sim \chi^2(k-1),\ (\text{graniczny})$$

Jednoczynnikowa ANOVA - układ doświadczalny

Na test jednoczynnikowej analizy wariancji możemy patrzeć jak na badanie istotności wpływu **czynnika** A na mającą charakter ilościowy i ciągły cechę X. Czynnik występuje na k **poziomach** które oznaczamy $A_1, A_2, \ldots, A_k$. Poziomy czynnika A nazywamy **obiektami** doświadczalnymi.

Obiekty doświadczalne są kontrolowane przez eksperymentatora, przy czym każdy z nich jest związany z pewną liczbą **jednostek doświadczalnych**. Liczba jednostek doświadczalnych związana z określonym obiektem nazywana jest **liczbą replikacji** tego obiektu.

Kojarząc różne obiekty z jednostkami doświadczalnymi, eksperymentator kreuje różne populacje, które pragnie porównać na podstawie obserwacji badanej w doświadczeniu cechy X.

Jednoczynnikowa ANOVA – układ doświadczalny

Alternatywna postać modelu

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

gdzie

$$ - średnia ogólna,

α_i - efekt i–tego obiektu, $\sum_{i=1}^{k} \alpha_i = 0$,

ε_{ij} - błędy (niezależne zmienne losowe o jednakowym rozkładzie $N(0, \sigma^2)$).

Hipoteza zerowa: czynnik A nie ma istotnego wpływu na cechę X:

$$H_0:\ \alpha_1 = \alpha_2 = \cdots = \alpha_k.$$

Hipoteza alternatywna: czynnik A ma istotny wpływ na cechę X:

$$H_1:\ \neg H_0.$$

Tabela analizy wariancji

Tradycyjnie wyniki analizy wariancji przedstawiamy w postaci tabeli.

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średnie kwadraty	Statystyka testowa
Obiekty	SSA	$k-1$	MSA	F
Błąd	SSE	$n-k$	MSE	
Całość	SST	$n-1$	MST	

$$MSA = SSA/(k-1), \quad MSE = SSE/(n-k), \quad MST = SST/(n-1).$$

Porównania wielokrotne (post hoc)

Procedury porównań wielokrotnych stosujemy wtedy, gdy zostanie odrzucona hipoteza zerowa w analizie wariancji!!!

Procedura NIR – Fishera.

Polega na testowaniu, dla każdej pary (i, j), $i, j = 1, 2, \dots, k$, $i \neq j$, oddzielnie hipotezy zerowej:

$$H_0: \ \mu_i = \mu_j,$$

przeciwko hipotezie alternatywnej

$$H_1: \ \mu_i \neq \mu_j.$$

Wartość statystyki testowej obliczamy ze wzoru:

$$t = \frac{\bar{x}_{i.} - \bar{x}_{j.}}{\sqrt{MSE}} \sqrt{\frac{n_i n_j}{n_i + n_j}}.$$

Przy braku istotnych różnic statystyka ta ma rozkład t Studenta z $n-k$ stopniami swobody.

Procedura HSD – Tukey'a.

Niech $n_1 = n_2 = \cdots = n_k = m$.

Polega na testowaniu, jednocześnie dla wszystkich par (i, j), $i, j = 1, 2, \dots, k$, $i \neq j$, hipotez zerowych:

$$H_0: \ \mu_i = \mu_j,$$

przeciwko hipotezom alternatywnym

$$H_1: \ \mu_i \neq \mu_j.$$

Wartość statystyki testowej obliczamy ze wzoru:

$$q = \frac{\bar{x}_{i.} - \bar{x}_{j.}}{\sqrt{MSE}}\sqrt{m}.$$

Przy braku istotnych różnic statystyka ta ma rozkład q (rozkład studentyzowanego rozstępu) z k i $k(m-1)$ stopniami swobody.

Funkcje związane z ANOVA:

aov - procedura główna,

LSD.test(agricolae) - procedura NIR Fishera,

HSD.test(agricolae) - procedura HSD Tukeya,

bartlett.test - test Bartletta.