Analiză dispersională

În literatura anglo-saxonă analiza dispersională este denumită ANOVA (de la ANalysis of VAriance).

Ne vom ocupa în continuare de testarea ipotezelor referitoare la mai mult de două medii, de exemplu

$$H_0 : m_1 = m_2 = m_3 = m_4 = m_5.$$

Utilizând tehnicile întâlnite până acum am putea testa pe rând ipotezele

$$\begin{array}{llll} H_{01} : m_1 = m_2 & H_{02} : m_1 = m_3 & H_{03} : m_1 = m_4 & H_{04} : m_1 = m_5 \\ H_{05} : m_2 = m_3 & H_{06} : m_2 = m_4 & H_{07} : m_2 = m_5 & H_{08} : m_3 = m_4 \\ H_{09} : m_3 = m_5 & H_{010} : m_4 = m_5. & & \end{array}$$

Acceptarea lui H_0 înseamnă acceptarea tuturor celor 10 ipoteze, iar respingerea lui H_0 respingerea a cel puţin una din ele. Această metodă este foarte laborioasă, iar eroarea totală este posibil să fie mai mare decât eroarea de genul I, α, asociată cu un singur test. Tehnicile ANOVA ne permit să testăm ipoteza nulă (toate mediile egale) în raport cu ipoteza alternativă (cel puţin o pereche de medii diferă), la un prag de semnificaţie α.

Introducere în tehnicile analizei dispersionale

Exemplul *Se crede că temperatura dintr-o întrepridere poate afecta productivitatea. Datele din tabelul 11.1 sunt numerele x de unităţi produse*

pe oră, pentru perioade de o oră selectate aleator, cu procesul de producţie desfăşurat la 3 ***niveluri*** *de temperatură. Datele din selecţiile repetate se numesc* ***replici****. Pentru două din temperaturi au fost obţinute 4 replici sau valori pentru date, iar pentru cea de-a treia temperatură 5 valori. Sugerează aceste date faptul că temperatura are un efect semnificativ asupra productivităţii la un nivel de 0.05?*

	Selecţie 13°	Selecţie 15°	Selecţie 16°
	10	7	3
	12	6	3
	10	7	5
	9	8	4
		7	
Total pe coloană	$C_1 = 41$ $\bar{x}_1 = 10.25$	$C_2 = 35$ $\bar{x}_2 = 7.0$	$C_3 = 15$ $\bar{x}_3 = 3.75$

Tabela 11.1: Nivelul şi influenţa temperaturii

Nivelul producţiei este măsurat prin valoarea medie; $\bar{x}_i$ indică producţia medie observată la nivelul i, unde $i = \overline{1,3}$ corespunde temperaturilor de 13, 15 şi respectiv 16°. Există o anumită variaţie între aceste medii. Deoarece mediile de selecţie nu se repetă neapărat când se iau selecţii repetate dintr-o populaţie, sunt de aşteptat anumite variaţii. Vom urmări în continuare problema: „este variaţia între valorile $\bar{x}$ datorată şansei sau se datorează efectului temperaturii asupra productivităţii?"

Soluţie. Ipoteza nulă pe care o vom testa este

$$H_0 : m_{13} = m_{15} = m_{16},$$

adică producţia medie este aceeaşi pentru fiecare nivel de temperatură testat. Cu alte cuvinte, temperatura nu are un efect semnificativ asupra productivităţii. Ipoteza alternativă este

$$H_a : m_{13} \neq m_{15} \vee m_{13} \neq m_{16} \vee m_{15} \neq m_{16},$$

adică nu toate mediile sunt egale. Vom respinge ipoteza nulă dacă datele ne arată că una sau mai multe medii diferă semnificativ de celelate. Decizia de acceptare sau respingere a lui H_0 se ia utilizând distribuţia şi statistica F.

$$F = \frac{s_1'^2}{s_2'^2}$$

Reamintim că valoarea lui F este raportul a două dispersii. Procedura de analiză dispersională va separa variaţiile pentru întreaga mulţime a datelor în două categorii. Pentru a realiza separarea vom lucra cu numărătorul expresiei

$$s'^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Numărătorul acestei fracţii se numeşte **suma pătratelor** (sum of squares). Vom calcula suma pătratelor, dar fără a utiliza pe $\bar{x}$

$$SS(total) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2. \qquad \boxed{1}$$

În cazul nostru avem

$$\sum_{i=1}^{n} x_i^2 = 10^2 + 12^2 + 10^2 + 9^2 + 7^2 + 6^2 + 7^2 + 8^2 + 7^2 + 3^2 + 3^2 + 5^2 + 4^2$$
$$= 731,$$
$$\sum_{i=1}^{n} x_i = 10 + 12 + 10 + 9 + 7 + 6 + 7 + 8 + 7 + 3 + 3 + 5 + 4 = 91$$
$$SS(total) = 731 - \frac{91^2}{13} = 94.$$

În continuare valoarea $SS(total)$=94.0 va fi separată în două părţi, $SS(temp)$ datorată nivelurilor de temperatură şi $SS(eroare)$ datorată erorilor de replicare. Această separare se numeşte **partiţionare**, deoarece $SS(temp) + SS(eroare) = SS(total)$. Suma pătratelor $SS(factor)$ (în cazul nostru $SS(temp)$) care măsoară variaţia între nivelurile factorilor (temperaturi) se obţine cu formula

$$SS(factor) = \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2, \qquad (2)$$

în care C_i reprezintă totalul coloanei i, k_i reprezintă numărul de replici pentru fiecare nivel al factorului, iar n reprezintă volumul selecţiei ($n = \sum k_i$).

Datele au fost aranjate astfel încât fiecare coloană reprezintă un nivel diferit al factorului care urmează a fi testat. Putem găsi $SS(temp)$ *pentru exemplul nostru cu ajutorul formulei* (3)

$$SS(temp) = \left(\frac{41^2}{4} + \frac{35^2}{5} + \frac{15^2}{4}\right) - \frac{91^2}{13} = 84.5.$$

Suma pătratelor $SS(eroare)$ care măsoară **variaţia în interiorul liniilor** se determină cu formula

$$SS(eroare) = \sum_{i=1}^{n} x_i^2 - \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \cdots\right). \qquad (4)$$

Din (1), (2) si (3) rezultă $SS(temp) + SS(eroare) = SS(total)$.

Este convenabil să utilizăm o **tabelă ANOVA** pentru a înregistra sumele de pătrate şi a organiza restul calculelor. Formatul unei tabele ANOVA este următorul

Sursa	SS	gl	MS
factor			
eroare			
Total			

Numărul de grade de libertate gl, asociat cu fiecare din cele trei surse se determină după cum urmează:

1. $$gl(factor) = c - 1,$$ 4

   unde c este numărul de niveluri pentru care factorul este testat (în cazul nostru numărul de coloane);

2. $$gl(total) = n - 1,$$ 5

   unde $n = k_1 + k_2 + k_3 + \cdots$ (k_i este numărul de replici pentru fiecare nivel), n este volumul selecţiei;

3. $gl(eroare)$ este suma gradelor de libertate a tuturor nivelurilor testate (coloane în tabelele de date); fiecare coloană are $k_i - 1$ grade de libertate, deci

$$gl(eroare) = (k_1 - 1) + (k_2 - 1) + (k_3 - 1) + \cdots$$

sau

$$gl(eroare) = n - c.$$

6

În cazul nostru avem

$$gl(temp) = c - 1 = 3 - 1 = 2$$
$$gl(total) = n - 1 = 13 - 1 = 12$$
$$gl(eroare) = n - c = 13 - 3 = 10.$$

Întotdeauna se verifică următoarele condiţii

$$SS(factor) + SS(eroare) = SS(total),$$
$$gl(factor) + gl(eroare) = gl(total).$$

7
8

Mediile pătraticedin ultima coloană a tabelului, $MS(factor)$, pentru factorul de testat şi respectiv $MS(eroare)$, pentru eroarea de replicare, se obţin împărţind suma pătratelor la numărul corespunzător de grade de libertate:

$$MS(factor) = \frac{SS(factor)}{gl(factor)},$$
$$MS(eroare) = \frac{SS(eroare)}{gl(eroare)}.$$

9
10

Pentru exemplul nostru avem

$$MS(temp) = \frac{84.5}{2} = 42.25$$
$$MS(eroare) = \frac{9.5}{10} = 0.95.$$

Tabelul anova complet este

Sursa	SS	gl	MS
temperatură	84.5	2	42.25
eroare	9.5	10	0.95
Total	94.0	12	

Testul utilizează cele două medii pătratice ca măsură a dispersiilor. Statistica testului este

$$F = \frac{MS(factor)}{MS(eroare)}. \qquad (11)$$

Pentru exemplul nostru se obţine

$$F = \frac{42.25}{0.95} = 44.47 = F^*.$$

Decizia de a respinge H_0 sau de a o accepta se ia comparând valoare calculată F^* cu valoarea critică unilaterală dreapta a distribuţiei F (cuantila $f_{gl(factor),gl(eroare),1-\alpha}$). În cazul nostru $F^* = 44.7 > f_{2,10,0.95} = 4.10$, deci vom respinge ipoteza nulă (vezi figura 11.1). De aceea concluzionăm că temperatura încăperii are un efect semnificativ asupra productivităţii.

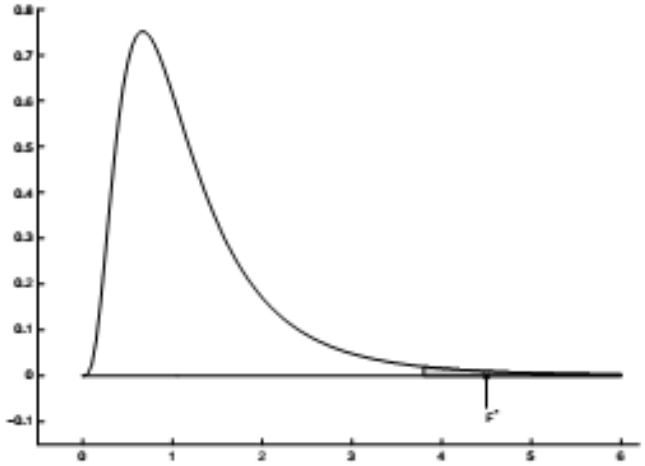


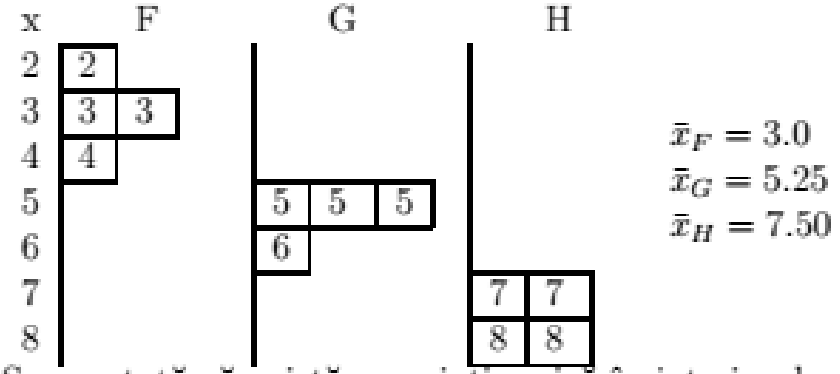
Figura 11.1: Regiune critică ANOVA

Logica ANOVA

Multe experimente se realizează pentru a determina efectul pe care un factor de test îl are asupra unei variabile de răspuns. Factorul de test ar putea fi temperatura (ca în exemplul anterior), fabricantul unui produs, ziua din săptămână şi aşa mai departe. În esenţă proiectarea unui test ANOVA cu un singur factor constă în a obţine selecţii independente pentru fiecare nivel al factorului care trebuie testat. Apoi vom lua o decizie statistică asupra efectului nivelurilor factorului de test asupra variabilei de răspuns (observate). Pe scurt, raţiunea acestei tehnici este următoarea: pentru a compara nivelurile factorilor de test, măsura **variaţiei dintre niveluri** (între coloanele sau liniile tabelei de date în funcţie de modul de organizare), $MS(factor)$ va fi comparată cu măsura **variaţiei în interiorul factorilor,** $MS(eroare)$. Dacă $MS(factor) > MS(eroare)$ în mod semnificativ, vom trage concluzia că mediile nivelurilor factorului care urmează a fi testat nu sunt identice. De aici rezultă că factorul care urmează a fi testat are un efect semnificativ asupra variabilei de răspuns. Dacă $MS(factor)$ nu este semnificativ mai mare decât $MS(eroare)$, nu vom putea respinge ipoteza nulă conform căreia toate mediile sunt egale.

Exemplu *Ne permit datele din tabelul 11.2 să afirmăm că există o diferenţă între mediile a trei populaţii m_F, m_G, m_H?*

	Nivelurile factorului		
	Selecţia pentru nivelul F	Selecţia pentru nivelul G	Selecţia pentru nivelul H
	3	5	8
	2	6	7
	3	5	7
	4	5	8
Total coloane	$C_F = 12$	$C_G = 21$	$C_H = 30$
	$\bar{x}_F = 3.0$	$\bar{x}_G = 5.25$	$\bar{x}_H = 7.50$



Se constată că există o variaţie mică în interiorul selecţiilor şi o variaţie mare între selecţii.

Aplicaţii ale ANOVA cu un singur factor

Înainte de a continua discuţia să sistematizăm notaţiile utilizate. Toate datele sunt dublu indexate: primul indice indică nivelul factorului de test (în cazul nostru coloana), iar al doilea numărul replicii (linia). C_i semnifică totalul pe coloana i, iar T totalul general.

	Nivelurile factorilor					
Replica	Selecţia din nivelul 1	Selecţia din nivelul 2	Selecţia din nivelul 3	...	Selecţia din nivelul c	
$k=1$	$x_{1,1}$	$x_{2,1}$	$x_{3,1}$	...	$x_{c,1}$	
$k=2$	$x_{1,2}$	$x_{2,2}$	$x_{3,2}$	...	$x_{c,2}$	
$k=3$	$x_{1,3}$	$x_{2,3}$	$x_{3,3}$	...	$x_{c,3}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
Totaluri coloane	C_1 $T=$	C_2 Total gen.$=$	C_3 $\sum\sum x_{i,j}$	... $=$	C_c $\sum C_i$	T

Pentru ANOVA cu un singur factor modelul matematic este dat de ecuaţia

$$x_{l,k} = m + F_l + \varepsilon_{k(l)},$$

pe care o putem interpreta astfel:

1. m este media tuturor datelor, fără a ţine cont de factorul de test;
2. F_l este efectul care factorul de testat îl are asupra variabilei de raspuns la fiecare nivel diferit l;
3. $\varepsilon_{k(l)}$ este eroarea experimentală ce apare printre cele k replici din fiecare c coloane.

Exemplu *Un club de tir realizează un experiment pe un grup selectat aleator de trăgători începători. Scopul experimentului este de a determina dacă precizia tragerii este influenţată de metoda de ochire utilizată: cu ochiul drept, cu ochiul stâng sau cu ambii ochi. Au fost selectaţi aleator 15 trăgători începători şi împărţiţi în 3 grupuri. Fiecare grup are aceişi pregătire şi experienţă, diferind doar metoda de ochire . După antrenarea completă fiecare trăgător a primit acelaşi număr de cartuşe şi i s-a spus să tragă la ţintă. Punctajul apare în tabela 11.3. Se poate afirma la nivelul de semnificaţie 0.05 că aceste metode de ochire sunt echivalente?*

Soluţie. În acest experiment factorul este metoda de ochire, iar nivelurile sunt cele trei metode de ochire. Replicile vor fi scorurile obţinute de trăgători. Ipoteza nulă este „cele trei metode au acelaşi efect (mediile obţinute pentru cele trei metode sunt aceleaşi)".

OD	OS	AO
12	10	16
10	17	14
18	16	16
12	13	11
14		20
		11

P1. $H_0 : m_{OD} = m_{OS} = m_{AO}$.

P2. $H_1 : m_{OD} \neq m_{OS} \vee m_{OD} \neq m_{AO} \vee m_{OS} \neq m_{AO}$ (nu toate mediile sunt egale).

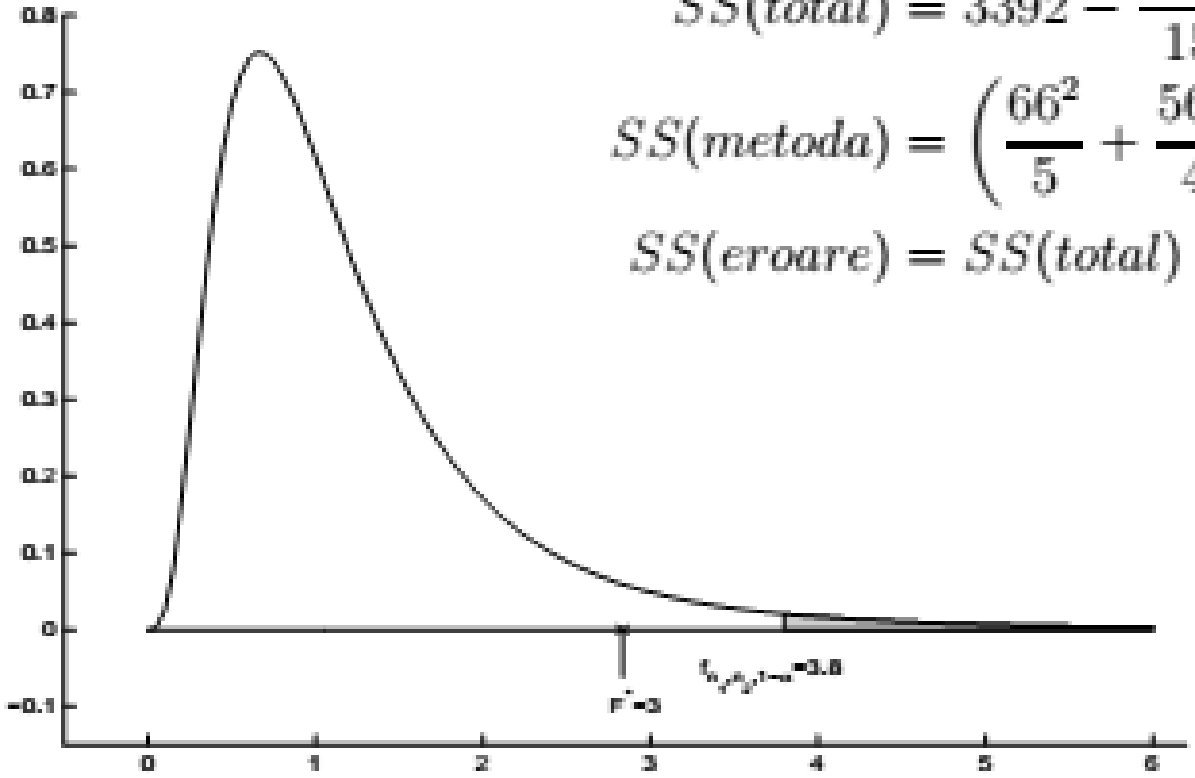
P3. Statistica testului este dată de (11), $\alpha = 0.05$, $gl(metoda) = 3 - 1 = 2$, $gl(eroare) = 15 - 3 = 12$, $f_{2,12,0.95} = 3.89$, iar regiunea critică apare în figura 11.2.

P4. Totalurile pe coloane sunt $C_{OD} = 66$, $C_{OS} = 56$, $C_{AO} = 98$, $\sum\sum x_{ej} = 220$ şi $\sum\sum x_{ej}^2 = 3392$.
avem:

$$SS(total) = 3392 - \frac{220^2}{15} = 165.33,$$

$$SS(metoda) = \left(\frac{66^2}{5} + \frac{56^2}{4} + \frac{98^2}{6}\right) - 3226.67 = 29.20,$$

$$SS(eroare) = SS(total) - SS(metoda) = 136.13.$$



Mediile pătratice sunt

$$MS(metoda) = \frac{29.20}{2} = 14.60$$
$$MS(eroare) = \frac{136.3}{12} = 11.35.$$

Rezultatele calculelor apar în tabela ANOVA de mai jos:

Sursa	SS	gl	MS
metoda	29.20	2	14.60
eroare	136.13	12	11.35
Total	165. 33	14	

Statistica testului este

$$F = \frac{14.60}{11.35} = 1.\,286 = F^*.$$

P5. Deoarece $F^* = 1.286 < f_{2,12,0.95} = 3.89$ (vezi figura), nu putem respinge H_0.

Examples

expand all

One-Way ANOVA

Create X with columns that are constants plus random normal disturbances with mean zero and standard deviation one.

```
X = meshgrid(1:5);
rng('default') % For reproducibility
X = X + normrnd(0,1,5,5)
```

```
X =

    1.5377    0.6923    1.6501    3.7950    5.6715
    2.8339    1.5664    6.0349    3.8759    3.7925
   -1.2588    2.3426    3.7254    5.4897    5.7172
    1.8622    5.5784    2.9369    5.4090    6.6302
    1.3188    4.7694    3.7147    5.4172    5.4889
```

Perform one-way ANOVA.

```
p = anova1(X)
```

```
p =

    0.0023
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	53.7238	4	13.4309	6.05	0.0023
Error	44.408	20	2.2204		
Total	98.1318	24			