L'anàlisi de la variància (ANOVA)

Josep Gibergans Bàguena

P08/05057/02313

Universitat Oberta de Catalunya
www.uoc.edu

Índex

L'anàlisi de la variància (ANOVA)

1. Introducció

Suposem que ens plantegem el problema de comparar la vida mitjana de dues classes de bombetes A i B; agafem una mostra de bombetes de la classe A i mesurem els temps de vida. Fem el mateix amb una mostra de bombetes de la classe B.

Normalment, les mitjanes dels temps de vida $\bar{x}_A$ i $\bar{x}_B$ no seran iguals. La diferència pot ser deguda a l'atzar o al fet que les bombetes d'una classe són de qualitat superior a les de l'altra. Precisament és això el que volem saber, si hi ha una diferència de qualitat o si no n'hi ha.

La tècnica del contrast d'hipòtesi serveix per a comprovar si una determinada hipòtesi sobre un fet lligat a un experiment aleatori es pot acceptar o s'ha de rebutjar.

Establim les hipòtesis:

- Hipòtesi nul·la: les bombetes de la marca A tenen una vida mitjana igual a la vida mitja de les bombetes de la marca B:

$$H_0 : \mu_A = \mu_B \qquad (H_0 : \mu_A - \mu_B = 0)$$

- Hipòtesi alternativa: les bombetes de la marca A i les de la marca B no tenen la mateixa vida mitjana:

$$H_1 : \mu_A \neq \mu_B \qquad (H_1 : \mu_A - \mu_B \neq 0)$$

Ara podem fer un contrast de la diferència de mitjanes per decidir si tots dos tipus de bombetes provenen de poblacions de bombetes amb vides mitjanes iguals.

Imaginem, però, que en lloc de comparar dues classes de bombetes, en volem comparar quatre. Si volem fer servir el contrast de diferències de mitjanes, cal contrastar dues a dues aquestes classes i, per tant, tenim:

$$\binom{4}{2} = 6$$

contrastos, és a dir, sis comparacions de dues mitjanes. Després cal analitzar i comparar tots els resultats. Evidentment, no és una tasca fàcil.

Tot això ens indica que aquesta manera de procedir no és la més adequada per a tractar aquest tipus de problemes. Utilitzarem una nova tècnica que es coneix com l'anàlisi de la variància, que serveix per a estudiar la generalització d'aquest problema en cas que tinguem més de dues mostres.

L'anàlisi de la variància (ANOVA) d'un conjunt de mostres consisteix a contrastar la hipòtesi nul·la "totes les mitjanes poblacionals d'on provenen les mostres són iguals", contra la hipòtesi alternativa "no totes les mitjanes són iguals" amb un nivell de significació α prefixat.

Teoria de l'anàlisi de la variància

L'abreviatura ANOVA prové de l'anglès *ANalysis Of VAriance* (anàlisi de la variància).
La teoria i metodologia de l'anàlisi de la variància va ser desenvolupada i introduïda per R.A. Fisher durant els primer anys de la segona dècada del segle xx.

El nom d'anàlisi de la variància que fa servir ANOVA prové del fet que, tot i que comparem mitjanes, l'estadístic de contrast que fa servir ANOVA es basa en el quocient de dos estimadors de la variància.

Exemples d'experiments amb l'ANOVA

A continuació presentem alguns exemples d'experiments en què es fa servir l'ANOVA:

a) Comparacions de vides mitjanes de tot tipus de dispositius per a diferents marques.

b) Comparacions entre el nombre de connexions a un servidor en diferents franges horàries.

c) Comparació de les qualificacions entre estudiants que han cursat una assignatura amb professors diferents.

d) Comparació del nombre d'accidents mitjà per a diferents intervals d'edat dels conductors.

e) Comparació entre les vendes mitjanes mensuals de diferents grans magatzems.

Hi ha moltes situacions experimentals en què hi ha dos o més factors d'interès alhora. Per exemple, es podrien analitzar tres tipus de benzina fixant-nos en dos factors: el consum i el nivell de contaminació. Per a tractar aquest problemes s'utilitza l'**anàlisi de la variància amb factors múltiples**, però aquest problema no l'estudiarem.

En aquesta sessió ens ocuparem de la situació en què volem comparar un nombre *k* de mostres (o grups) a partir d'una única característica de l'individu observat (variable o factor).

2. La informació mostral

La taula següent registra la notació que utilitzarem al llarg de la sessió:

	Població 1	Població 2	...	Població *j*	...	Població *k*
Mitjana	μ_1	μ_2	...	μ_j	...	μ_k
Variància	σ_1^2	σ_2^2	...	σ_j^2	...	σ_k^2

	Mostra 1	Mostra 2	...	Mostra *j*	...	Mostra *k*
	x_{11}	x_{12}		x_{1j}		x_{1k}
	x_{21}	x_{22}		x_{2j}		x_{2k}
	...	...		...		...
	...	...		x_{ij}		...
	...			...		$x_{n_k k}$
	$x_{n_1 1}$			...		
		$x_{n_2 2}$		$x_{n_j j}$		
Mitjana	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_j$		$\bar{x}_k$
Variància	s_1^2	s_2^2		s_j^2		s_k^2

Subíndexs

Cada observació x_{ij} porta dos subíndexs que ens informen que es tracta de l'observació *i*-èsima de la mostra *j*-èsima.

No hi ha cap motiu perquè les mostres tinguin la mateixa mida, de manera que amb n_j indicarem la grandària de la mostra *j*-èsima. Calcularem la mitjana de la mostra *j*-èsima mitjançant l'expressió següent:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} = \frac{x_{1j} + x_{2j} + \ldots + x_{n_j j}}{n_j}$$

La variància mostral d'aquesta mostra *j*-èsima vindrà donada per:

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1}$$

Si ara considerem el conjunt de totes les observacions format pels individus de totes les mostres, aquest estarà format per un nombre d'individus igual a la suma dels individus de totes les mostres, és a dir:

$$n = n_1 + n_2 + \ldots + n_k$$

I d'aquest conjunt global també podem calcular la mitjana global:

Mitjana global i mitjana de les mitjanes

És important no confondre la mitjana global amb la mitjana de les mitjanes. Només són el mateix en cas que les mostres tinguin la mateixa mida.

$$\bar{x} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n} = \frac{\sum_{j=1}^{k} n_j \bar{x}_j}{n}$$

i també la variància global:

$$s^2 = \frac{\sum_{j=1}^{\kappa} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2}{n - 1}$$

Exemple de les tres marques d'ordinadors

Considerem que es fa un experiment per a comparar el temps que triguen tres marques d'ordinadors de diferent marca a carregar un mateix sistema operatiu.

Es pren una mostra de quatre ordinadors de la marca A, és a dir, es mesura el temps (en segons) que triguen a carregar el sistema operatiu quatre ordinadors d'aquesta marca. De la marca B es prenen sis mesures i cinc de la marca C. La taula següent registra els resultats de l'experiment:

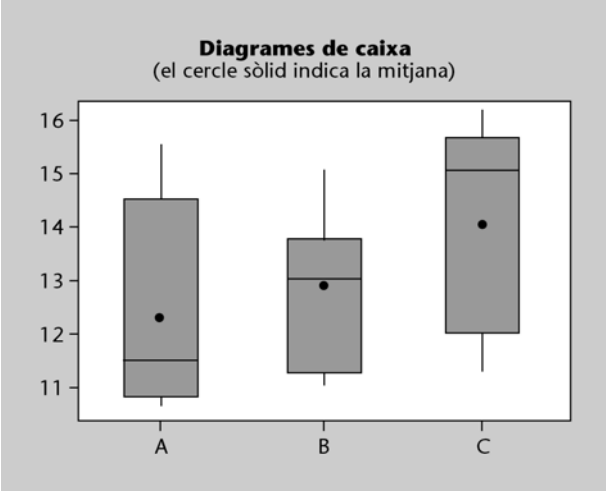
Marca A	10,7	11,2	12,0	15,5		
Marca B	13,4	11,5	11,2	15,1	13,3	12,9
Marca C	11,5	12,7	15,4	16,1	15,2	

Fent servir la notació que hem presentat anteriorment, tenim:

	Mostra j = 1	**Mostra j = 2**	**Mostra j = 3**
	$x_{11} = 10,7$	$x_{12} = 13,4$	$x_{13} = 11,5$
	$x_{21} = 11,2$	$x_{22} = 11,5$	$x_{23} = 12,7$
	$x_{31} = 12,0$	$x_{32} = 11,2$	$x_{33} = 15,4$
	$x_{41} = 15,5$	$x_{42} = 15,1$	$x_{43} = 16,1$
		$x_{52} = 13,3$	$x_{53} = 15,2$
		$x_{62} = 12,9$	
Mitjana	$\bar{x}_1 = 12,35$	$\bar{x}_2 = 12,90$	$\bar{x}_3 = 14,18$
Variància	$s_1^2 = 4,70$	$s_2^2 = 2,02$	$s_3^2 = 3,90$

Mirant aquests resultats, podem pensar que les mostres dels ordinadors A i B poden provenir de poblacions amb la mateixa mitjana, atès que les mitjanes mostrals 12,35 i 12,90, respectivament, són bastant properes. La mitjana mostral de la marca C és 14,18; aquesta està més allunyada de les altres, però presenta una major dispersió que les anteriors; no és tan fàcil, doncs, pensar si aquesta mostra prové d'una població amb la mateixa mitjana que els ordinadors de les marques A i B.

És possible representar aquesta situació mitjançant els diagrames de caixa de les tres mostres:



Observació

La tasca de comparar més de dues mostres no és fàcil.

3. La variabilitat de la mostra global: les sumes de quadrats

Hem vist que podem considerar el conjunt global format per tots els elements de les mostres i, després, calcular la mitjana d'aquest conjunt global. A continuació, intentarem d'explicar a què són degudes les diferències entre els valors de les observacions x_{ij} i el valor de la mitjana global $\bar{x}$. Entendrem per **variabilitat** la diferència entre els valors observats i la mitjana. Veurem que aquesta variabilitat és deguda a dos factors:

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

1) Variabilitat dintre de cada mostra: diferència entre l'observació i la mitjana de la mostra $(x_{ij} - \bar{x}_j)$.

2) Variabilitat entre les mostres: diferència entre la mitjana de la mostra i la mitjana global $(\bar{x}_j - \bar{x})$.

Si hi ha molta variabilitat entre les mostres podrem pensar que és degut al fet que es tracta de mostres estretes de poblacions diferents o simplement a l'origen aleatori de les mostres. A continuació, veurem com podem separar aquests dos efectes provocats per la variabilitat dintre de cada mostra i per la variabilitat entre les mostres.

Observació

Les mitjanes mostrals poden ser diferents pel fet que provenen de poblacions amb mitjanes diferents o simplement per l'origen aleatori de les mostres.

Si sumem al quadrat la darrera expressió totes les observacions mitjançant un doble sumatori, un per a les mostres i l'altre per a les observacions de cada mostra, tenim:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)(\bar{x}_j-\bar{x})$$

en què podem veure que el darrer sumand és zero:

$$\sum_{j=1}^{\kappa}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)(\bar{x}_j-\bar{x}) = \sum_{j=1}^{\kappa}(\bar{x}_j-\bar{x})\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j) = 0$$

ja que:

$$\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j) = n_j\bar{x}_j - n_j\bar{x}_j = 0$$

Amb tot, la variabilitat de la mostra global es pot descompondre en dues parts:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^2$$

$$SQT = SQD + SQE$$

Suma de quadrats totals (SQT) =
= suma de quadrats dintre les mostres (SQD) +
+ suma de quadrats entre mostres (SQE)

Considerem cadascun d'aquests sumands:

Obtenció de la variància de la mostra global

És immediat veure que si dividim SQT/(n – 1) tenim la variància de la mostra global.

- La Suma de Quadrats Totals (**SQT**) ens informa de la variabilitat de la mostra global.

- La Suma de Quadrats Dintre les mostres (**SQD**) és una mesura de la variació dins les mostres.

- La Suma de Quadrats Entre mostres (**SQE**) és una mesura de la variació entre les mostres; la calculem a partir de la diferència entre las mitjanes de les mostres i la mitjana total. Si les mitjanes són molt diferents, aleshores aquesta quantitat és gran.

Si dividim SQD i SQE per $n - k$ i $k - 1$, respectivament, obtenim els estadístics següents:

$$MQD = \frac{SQD}{n-k};\ MQE = \frac{SQE}{k-1}$$

que, com veurem tot seguit, són necessaris per a fer l'ANOVA.

4. Hipòtesis sobre les dades per a fer l'ANOVA

Per poder fer una anàlisi d'aquests tipus, cal tenir les hipòtesis següents:

1) Les k mostres han de ser **aleatòries i independents** entre si.

2) Les poblacions han de ser **normals.**

3) Les variàncies de les k poblacions han de ser **idèntiques**:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$$

Sota aquestes hipòtesis i quan es compleix $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$, és a dir, si les mitjanes poblacionals són totes iguals, les sumes de quadrats *SQE* i *SQD* es distribueixen segons distribucions χ^2 amb $(k - 1)$ i $(n - k)$ graus de llibertat, respectivament.

Per a mostres d'una població normal $N(\mu, \sigma)$ sempre s'acompleix que:

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

té una distribució χ^2 amb $n - 1$ graus de llibertat.

I com que són independents, una important conseqüència és que el quocient entre aquests estadístics:

$$f = \frac{MQE}{MQD} = \frac{SQE/(k-1)}{SQD/(n-k)}$$

es distribueix segons una distribució F de Snedecor amb $(k - 1)$ graus de llibertat al numerador i $(n - k)$ al denominador.

Quocient de variables aleatòries

Si X és una variable aleatòria que té una distribució χ^2 amb n graus de llibertat; Y és una altra variable aleatòria que té una distribució χ^2 amb m graus de llibertat i X i Y són independents, aleshores la variable:

$$Z = \frac{X/n}{Y/m}$$

es distribueix segons una F de Snedecor amb n i m graus de llibertat al numerador i denominador, respectivament.

A continuació, veurem com podem fer servir aquesta descomposició de la variabilitat de les dades mostrals per a construir un contrast d'hipòtesi que ens permeti de prendre una decisió sobre la igualtat de les mitjanes de les poblacions de procedència de les mostres de l'estudi.

5. L'ANOVA és un contrast d'hipòtesi

L'estadístic de contrast que farem servir en l'anàlisi de la variància es basa en el fet de comparar els dos orígens de la variabilitat de les mostres que hem trobat en l'apartat anterior: la variació entre les mostres i la variació dintre les mostres. Suposarem que es compleixen les hipòtesis del model. Una vegada fets aquests supòsits, procedirem de la manera següent:

1) Plantejarem les nostres hipòtesis:

- Hipòtesi nul·la: H_0: totes les mitjanes són iguals:

$$\mu_1 = \mu_2 = \ldots = \mu_k = \mu$$

- Hipòtesi alternativa: H_1: no totes les mitjanes són iguals

2) Fixarem un nivell significatiu α.

3) Calcularem l'estadístic de contrast a partir de les sumes de quadrats:

$$f = \frac{SQE/(k-1)}{SQD/(n-k)}$$

que, com hem vist en l'apartat anterior, si es compleix la hipòtesi nul·la (igualtat de mitjanes) és una observació d'una distribució F de Snedecor amb $n - k$ graus de llibertat al denominador i $k - 1$ graus de llibertat al numerador.

4) Finalment, podem actuar de dues maneres:

El *p*-valor

El p-valor és la probabilitat del resultat observat o d'un de més allunyat si la hipòtesi nul·la és certa.

a) A partir del *p*-valor. Aquest valor *és:* $p = P(F > f)$*:*

- Si $p \leq \alpha$ es rebutja la hipòtesi nul·la H_0
- Si $p > \alpha$ no es rebutja la hipòtesi nul·la H_0

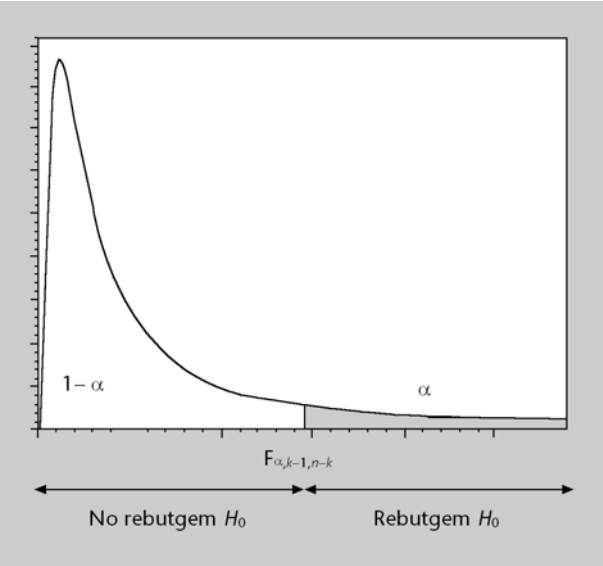
b) A partir del valor crític $F_{\alpha,\, k-1,\, n-k}$ que separa la regió d'acceptació de la regió de rebuig.

No rebutjar o rebutjar la H_0

No rebutjar la H_0 no vol dir exactament que acceptem la hipòtesi, sinó simplement que res no s'oposa a pensar que la H_0 pugui ser certa.

Rebutjar la H_0 no vol dir necessàriament que totes les mitjanes siguin diferents, sinó que vol dir que alguna (potser totes) és diferent d'una altra.

- Si $f > F_{\alpha,k-1,\, n-k}$ es rebutja la hipòtesi nul·la H_0
- Si $f \leq F_{\alpha,k-1,\, n-k}$ no es rebutja la hipòtesi nul·la H_0



Si volem determinar quins són els grups que presenten unes diferències prou significatives, farem proves *t* de Student per a comparació de mitjanes tal com es plantejava en l'inici de la sessió.

6. Construcció de la taula de l'ANOVA

En aquest apartat estem interessats a presentar una forma convenient i habitual de presentar el càlculs i resultats de l'anàlisi de la variància. Aquesta manera de sintetitzar aquests càlculs és en forma de taula, anomenada *taula de l'ANOVA*. Podem dur a terme tots els càlculs de la taula a partir de les mitjanes $\bar{x}_j$ i variàncies s_j^2 de les diferents mostres o de la mitjana $\bar{x}$ de la mostra global.

Podem escriure les sumes de quadrats que ens fan falta per a calcular l'estadístic de contrast de la manera següent:

$$SQE = \sum_{j=1}^{\kappa}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{\kappa} n_j(\bar{x}_j - \bar{x})^2$$

$$SQD = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{k}(n_j - 1)s_j^2$$

$$SQT = SQD + SQE$$

Taula de l'anàlisi de la variància				**Regla de decisió**
Font de variació	**Suma de quadrats**	**Graus de llibertat**	**Mitjana de quadrats**	**Estadístic de prova**
Entre grups	$SQE = \sum_j n_j(\bar{x}_j - \bar{x})^2$	$k - 1$	SQE/(k – 1)	$f = \frac{SQE/(k-1)}{SQD/(n-k)}$
Dins els grups	$SQD = \sum_j\sum_i (x_{ij} - \bar{x}_j)^2$	$n - k$	SQD/(n – k)	
Total	$SQT = \sum_j\sum_i (x_{ij} - \bar{x})^2$	$n - 1$	-	

k = nombre de mostres, n = grandària de la mostra total

Exemple de les tres marques d'ordinadors (II)

Tornem a l'exemple de la comparació de temps de càrrega d'un sistema operatiu per a tres marques diferents d'ordinadors.

Farem una anàlisi de la variància amb nivell significatiu $\alpha = 0{,}05$ per a determinar si podem considerar que les mitjanes dels temps dels tres ordinadors són iguals.

Hipòtesi nul·la: $H_0 : \mu_1 = \mu_2 = \mu_3$

Hipòtesi alternativa: H_1: no totes les mitjanes són iguals

Ja havíem calculat:

Mostra	***A*** $j = 1$	***B*** $j = 2$	***C*** $j = 3$
Grandària	$n_1 = 4$	$n_2 = 6$	$n_3 = 5$
Mitjana	$\bar{x}_1 = 12{,}35$	$\bar{x}_2 = 12{,}90$	$\bar{x}_3 = 14{,}18$
Variància	$s_1^2 = 4{,}70$	$s_2^2 = 2{,}02$	$s_3^2 = 3{,}90$

De manera que la mitjana de la mostra global és:

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \bar{x}_3 n_3}{n_1 + n_2 + n_3} = \frac{12{,}35 \cdot 4 + 12{,}90 \cdot 6 + 14{,}18 \cdot 5}{4 + 6 + 5} = 13{,}18$$

A continuació podem trobar les sumes de quadrats:

- $SQD = \sum_{j=1}^{\kappa}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{\kappa}(n_j - 1)s_j^2 = (4-1)4{,}70 + (6-1)2{,}02 + (5-1)3{,}90 = 39{,}80$
- $SQE = \sum_{j=1}^{\kappa} n_j(\bar{x}_j - \bar{x})^2 = 4(12{,}35 - 13{,}18)^2 + 6(12{,}90 - 13{,}18)^2 + 5(14{,}18 - 13{,}18)^2 = 8{,}226$

I construir la taula de l'ANOVA:

Taula de l'anàlisi de la variància				Regla de decisió
Font de variació	**Suma de quadrats**	**Graus de llibertat**	**Mitjana de quadrats**	**Estadístic de prova**
Entre grups	SQE = 8,226	3 – 1 = 2	8,226/2 = 4,113	$f = \frac{4,113}{3,32} = 1,24$
Dins els grups	SQD = 39,80	15 – 3 = 12	39,80/12 = 3,32	
Total	SQT = 48,026	15 – 1 = 14	-	

$k = 3$ (nombre de mostres), $n = 15$ (grandària de la mostra total)

Estadístic de contrast: $f = \frac{SQE/(k-1)}{SQD/(n-k)} = \frac{4,113}{3,32} = 1,24$.

Aquest estadístic segueix una distribució F de Snedecor amb $k - 1 = 2$ i $n - k = 12$ graus de llibertat al numerador i denominador, respectivament.

1) A partir del *p*-valor:

$$P(F > f) = P(F > 1,24) = 0,3239 > 0,05$$

Per tant, no rebutgem la hipòtesi nul·la.

2) A partir del valor crític:

Per a un nivell significatiu $\alpha = 0,05$, tenim un valor crític:

$$F_{0,05;2;12} = 3,89$$

Si comparem aquest valor amb l'estadístic de contrast, $f = 1,24$, tenim que $1,24 < 3,89$ i, per tant, no rebutgem la hipòtesi nul·la.

Així, doncs, podem concloure que no hi ha una diferència significativa entre els temps que triguen les tres marques d'ordinadors a carregar el sistema operatiu.

7. Resum

En aquesta sessió hem presentat la tècnica d'anàlisi de la variància (ANOVA) per a la comparació de les mitjanes per a més de dues mostres. Hem comprovat que la variació de les observacions és deguda a dos factors: la variabilitat dintre de cada mostra i la variabilitat entre les mostres. Hem expressat aquestes variacions numèricament mitjançant sumes de quadrats. Amb les sumes de quadrats hem pogut trobar un estadístic de prova per a contrastar la igualtat de les mitjanes de les mostres. Finalment, hem après a resumir tots els càlculs en l'anomenada *taula de l'ANOVA*.

Exercicis

1. Considerem quatre companyies (A, B, C i D) les accions de les quals cotitzen en borsa. Seleccionem aleatòriament les cotitzacions d'aquestes accions durant diferents instants de temps al llarg d'un mes. Així, doncs, per a la companyia A s'observa la cotització en cinc instants aleatoris, per a la B s'observa en quatre, per a la C s'observa en sis i, finalment, per a la companyia D, en cinc.

En la taula següent es dóna la cotització en cèntims d'euro de les diferents accions en els instants de temps seleccionats:

A	670, 840, 780, 610, 900
B	600, 800, 690, 650
C	800, 810, 730, 690, 750, 720
D	970, 840, 930, 790, 920

Contrasteu el nivell del 5% si les cotitzacions mitjanes de les accions de cadascuna de les quatre companyies es poden considerar iguals. Feu la taula d'anàlisi de la variància.

2. Els estudiants de segon curs d'una escola universitària d'enginyeria van estar repartits de forma aleatòria en tres grups. A cada grup es va ensenyar estadística amb una estratègia docent diferent. Al final dels curs tots els alumnes van fer el mateix examen. Aleatòriament, es van seleccionar algunes qualificacions obtingudes per alguns alumnes dels tres grups. Els resultats són els següents:

Grup 1: $n = 5$, $\sum x_i = 116$ $\sum x_i^2 = 2.728$

Grup 2: $n = 5$ $\sum x_i = 126$ $\sum x_i^2 = 3.294$

Grup 3: $n = 7$ $\sum x_i = 171$ $\sum x_i^2 = 4.193$

a) Sobre quins supòsits podreu fer una anàlisi de la variància?

b) Feu una anàlisi de la variància i digueu si podeu assegurar a un nivell significatiu del 0,05 que el resultat obtingut depèn de la tècnica d'ensenyament utilitzada.

Solucionari

1. En aquest problema estem interessats a comparar quatre poblacions, farem servir una anàlisi de la variància o ANOVA. Sabem que l'ANOVA ens permet de comparar les mitjanes de diversos grups.

Hipòtesi nul·la: H_0: les mitjanes són iguals: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Hipòtesi alternativa: H_1: les mitjanes no són iguals.

Per a construir la taula de l'anàlisi de la variància:

Primer calculem les mitjanes i les variàncies de cada mostra:

	J	n_j	Mitjana $\bar{x}_j$	Variància s_j^2
A	1	5	760,00	14.250,00
B	2	4	685,00	7.233,33
C	3	6	750,00	2.200,00
D	4	5	890,00	5.350,00

I la mitjana total de les mostres:

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \bar{x}_3 n_3 + \bar{x}_4 n_4}{n_1 + n_2 + n_3 + n_4} = 774{,}50$$

A continuació, podem trobar les sumes de quadrats:

- $SQD = \sum_{j=1}^{\kappa} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{\kappa} (n_j - 1) s_j^2 = \sum_{j=1}^{4} (n_j - 1) s_j^2$

  = (5 – 1) 14.250,00 + (4 – 1) 7.233,33 + (6 – 1) 2.200,00 + (5 – 1) 5.350,00 =
  = 111.100,00

En aquesta expressió hem tingut en compte que: $s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$

- $SQE = \sum_{j=1}^{\kappa} n_j (\bar{x}_j - \bar{x})^2 = 5(760{,}00 - 774{,}50)^2 + 4(685{,}00 - 774{,}50)^2 + 6(750{,}00 -$
  $- 774{,}50)^2 + + 5(890{,}00 - 774{,}50)^2 = 103.395{,}00$

Ara ja podem construir la taula:

Taula de l'anàlisi de la variància			
Font de la variació	**Suma de quadrats**	**Graus de llibertat**	**Mitjana de quadrats**
Entre grups	$SQE = \sum_j n_j (\bar{x}_j - \bar{x})^2$ SQE = 103.395,00	$k - 1$ 4 – 1 = 3	SQE / (k – 1) = = 103.395,00/3 = 34.465,00
Dins els grups	$SQD = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$ SQD = 111.100,00	$n - k$ 20 – 4 = 16	SQD/($n - k$) = 111.100,00/16 = = 6.943,75
Total	$SQT = \sum_j \sum_i (x_{ij} - \bar{x})^2$	20 – 1 = 19	-

k = nombre de grups = 4, n = grandària de la mostra total = 20

Estadístic de contrast: $f = \frac{SQE/(k-1)}{(SQD)/(n-k)} = \frac{34.465,00}{6.943,75} = 4,96$

L'estadístic segueix una distribució F de Snedecor amb $k - 1 = 3$ i $n - k = 16$ graus de llibertat.

Calculem el p-valor:

$$P(F > f) = P(F > 4,96) = 0,0127 < 0,05,$$

de manera que rebutgem H_0. Amb una confiança del 95%, hi ha diferència significativa entre les quatre companyies.

2.

a) Per a poder aplicar aquesta tècnica amb fiabilitat, són necessàries les restriccions prèvies que presentem a continuació:

1. Les mostres han de ser independents.

2. Les poblacions (o subpoblacions) segueixen distribucions normals; altrament, la mostra triada ha de ser prou gran (més de trenta observacions a cada submostra). En el nostre cas, les mostres són menors que trenta. Per tant, per a poder aplicar l'ANOVA, hem de suposar que les poblacions segueixen distribucions normals.

3. La variància per a cada població (o subpoblació) és la mateixa.

b) En aquest problema estem interessats a comparar tres poblacions i farem servir una anàlisi de la variància o ANOVA. Sabem que l'ANOVA ens permet de comparar les mitjanes de diversos grups.

Hipòtesi nul·la: H_0: les mitjanes són iguals: $\mu_1 = \mu_2 = \mu_3$

Hipòtesi alternativa: H_1: les mitjanes no són iguals

Per a realitzar la taula de l'anàlisi de la variància, en primer lloc, calculem les mitjanes i les variàncies de cada mostra:

	n_j	Mitjana $\bar{x}_j$	Variància s_j^2
Grup 1 ($j = 1$)	5	23,2	9,2
Grup 2 ($j = 2$)	5	25,6	4,3
Grup 3 ($j = 3$)	7	24,4	4,25

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1} = \frac{116}{5} = 23,2$$

$$s_1^2 = \frac{\sum (x_{1i} - \bar{x}_1)^2}{n_1 - 1} = \frac{1}{n_1 - 1}\left[\sum x_{1i}^2 - n(\bar{x}_1)^2\right] = \frac{1}{5-1}\left[2.728 - 5 \cdot 23,2^2\right] = 9,2$$

$$\bar{x}_2 = \frac{\sum x_{2i}}{n_2} = \frac{128}{5} = 25,6$$

$$s_2^2 = \frac{\sum (x_{2i} - \bar{x}_2)^2}{n_2 - 1} = \frac{1}{n_2 - 1}\left[\sum x_{2i}^2 - n(\bar{x}_2)^2\right] = \frac{1}{5-1}\left[3.294 - 5 \cdot 25,6^2\right] = 4,3$$

$$\bar{x}_3 = \frac{\sum x_{3i}}{n_3} = \frac{171}{7} = 24,4$$

$$s_3^2 = \frac{\sum (x_{3i} - \bar{x}_3)^2}{n_3 - 1} = \frac{1}{n_3 - 1}\left[\sum x_{3i}^2 - n(\bar{x}_3)^2\right] = \frac{1}{7-1}\left[4.193 - 7 \cdot 24,4^2\right] = 4,25$$

A continuació, podem trobar les sumes de quadrats mitjançant aquestes expressions:

$$SQD = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{k}(n_j - 1)s_j^2 = \sum_{j=1}^{3}(n_j - 1)s_j^2 =$$
$$(5 - 1)\ 9,2 + (5 - 1)\ 4,3 + (7 - 1)\ 4,25 = 79,5$$

En aquesta expressió hem tingut en compte que:

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$$

$$SQE = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 = 5(23,2 - 24,4)^2 + 5(25,6 - 24,4)^2 + 7(24,4 - 24,4)^2 = 14,40$$

Ara ja podem calcular l'estadístic de contrast. Primer construïm la taula d'anàlisi de la variància:

Taula de l'anàlisi de la variància			
Font de la variació	**Suma de quadrats**	**Graus de llibertat**	**Mitjana de quadrats**
Entre grups	$SQE = \sum_j n_j(\bar{x}_j - \bar{x})^2$ SQE = 14,40	$k - 1$ $3 - 1 = 2$	SQE/($k - 1$) = 14,40 / 2 = 7,202
Dins els grups	$SQD = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$ SQD = 79,50	$n - k$ $17 - 3 = 14$	SQD/($n - k$) = 79,5 / 14 = 5,679
Total	$SQT = \sum_j \sum_i (x_{ij} - \bar{x})^2$	$n - 1 = 16$	-

k = nombre de grups = 3, n = grandària de la mostra total = 17

Estadístic de contrast: $f = \frac{SQE/(k-1)}{(SQD)/(n-k)} = 1,268$

L'estadístic segueix una distribució F de Snedecor amb $k - 1 = 2$ i $n - k = 14$ graus de llibertat. Tenim un p-valor:

$$P(F > f) = P(F > 1{,}268) = 0{,}3118 > 0{,}05$$

No rebutgem H_0, de manera que podem assegurar a un nivell significatiu del 0,05 que el resultat obtingut no depèn de l'estratègia docent utilitzada.