

El análisis de la varianza (ANOVA)

Josep Gibergans Bàguena

P08/75057/02313

Índice

Sesión 1

El análisis de la varianza (ANOVA)	5
1. Introducción	5
2. La información muestral	6
3. La variabilidad de la muestra global: las sumas de cuadrados	9
4. Hipótesis sobre los datos para llevar a cabo el ANOVA	10
5. El ANOVA es un contraste de hipótesis	11
6. Construcción de la tabla del ANOVA	12
7. Resumen	14
Ejercicios	15

El análisis de la varianza (ANOVA)

1. Introducción

Supongamos que nos planteamos el problema de comparar la vida media de dos clases de bombillas A y B; cogemos una muestra de bombillas de la clase A y medimos los tiempos de vida. Hacemos lo mismo con una muestra de bombillas de la clase B.

Normalmente, las medias de los tiempos de vida \bar{x}_A y \bar{x}_B no serán iguales. La diferencia puede ser debida al azar o al hecho de que las bombillas de una clase son de calidad superior a las de la otra. Precisamente es eso lo que queremos saber, si hay una diferencia de calidad o si no la hay.

La técnica del contraste de hipótesis sirve para comprobar si una determinada hipótesis sobre un hecho vinculado a un experimento aleatorio se puede aceptar o se tiene que rechazar.

Establecemos las hipótesis:

- Hipótesis nula: las bombillas de la marca A tienen una vida media igual a la vida media de las bombillas de la marca B:

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

- Hipótesis alternativa: las bombillas de la marca A y las de la marca B no tienen la misma vida media:

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

Ahora podemos hacer un contraste de la diferencia de medias para decidir si los dos tipos de bombillas provienen de poblaciones de bombillas con vidas medias iguales.

Sin embargo, imaginemos que en lugar de comparar dos clases de bombillas, queremos comparar cuatro. Si queremos utilizar el contraste de diferencias de medias, es preciso contrastar dos a dos estas clases y, por tanto, tenemos:

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

contrastes, es decir, seis comparaciones de dos medias. Después hay que analizar y comparar todos los resultados. Evidentemente, no es una tarea fácil.

Todo esto nos indica que esta manera de proceder no es la más adecuada para tratar este tipo de problemas. Utilizaremos una nueva técnica que se conoce como el análisis de la varianza, que sirve para estudiar la generalización de este problema en caso de que tengamos más de dos muestras.

El análisis de la varianza (ANOVA) de un conjunto de muestras consiste en contrastar la hipótesis nula “todas las medias poblacionales de las que provienen las muestras son iguales”, contra la hipótesis alternativa “no todas las medias son iguales” con un nivel de significación α prefijado.

Teoría del análisis de la varianza


La abreviatura ANOVA proviene del inglés *ANalysis Of VAriance* (‘análisis de la varianza’). La teoría y metodología del análisis de la varianza fueron desarrolladas e introducidas por R.A. Fisher durante los primeros años de la segunda década del siglo xx.

El número de análisis de la varianza que utiliza ANOVA proviene del hecho de que, a pesar de que comparamos medias, el estadístico de contraste que utiliza ANOVA se basa en el cociente de dos estimadores de la varianza.

Ejemplos de experimentos con el ANOVA

A continuación presentamos algunos ejemplos de experimentos en los que se utiliza el ANOVA:

- a) Comparaciones de vidas medias de todo tipo de dispositivos para diferentes marcas.
- b) Comparaciones entre el número de conexiones a un servidor en diferentes franjas horarias.
- c) Comparación de las cualificaciones entre estudiantes que han cursado una asignatura con profesores diferentes.
- d) Comparación del número medio de accidentes para diferentes intervalos de edad de los conductores.
- e) Comparación entre las ventas medias mensuales de diferentes grandes almacenes.

Se dan muchas situaciones experimentales en las que hay dos o más factores de interés al mismo tiempo. Por ejemplo, se podrían analizar tres tipos de gasolina fijándonos en dos factores: el consumo y el nivel de contaminación. Para tratar este problema se utiliza el **análisis de la varianza con factores múltiples**, pero aquí no lo estudiaremos. 

En esta sesión nos ocuparemos de la situación en la que queremos comparar un número k de muestras (o grupos) a partir de una única característica del individuo observado (variable o factor).

2. La información muestral

La tabla siguiente registra la notación que utilizaremos a lo largo de la sesión:

	Población 1	Población 2	...	Población j	...	Población k
Media	μ_1	μ_2	...	μ_j	...	μ_k
Varianza	σ_1^2	σ_2^2	...	σ_j^2	...	σ_k^2

	Muestra 1	Muestra 2	...	Muestra j	...	Muestra k
	x_{11}	x_{12}		x_{1j}		x_{1k}
	x_{21}	x_{22}		x_{2j}		x_{2k}

		x_{ij}		...
		$x_{n_k k}$
	$x_{n_1 1}$...		
		$x_{n_2 2}$		$x_{n_j j}$		
Media	\bar{x}_1	\bar{x}_2		\bar{x}_j		\bar{x}_k
Varianza	s_1^2	s_2^2		s_j^2		s_k^2

Subíndices

Cada observación x_{ij} lleva dos subíndices que nos informan de que se trata de la observación i -ésima de la muestra j -ésima.

No existe ningún motivo para que las muestras tengan el mismo tamaño, de manera que con n_j indicaremos el tamaño de la muestra j -ésima. Calcularemos la media de la muestra j -ésima mediante la expresión siguiente:

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

La varianza muestral de esta muestra j -ésima vendrá dada por:

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

Si ahora consideramos el conjunto de todas las observaciones formado por los individuos de todas las muestras, éste estará formado por un número de individuos igual a la suma de los individuos de todas las muestras, es decir:

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

Y de este conjunto global también podemos calcular la media global:

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

así como la varianza global:

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

Media global y media de las medias

Es importante no confundir la media global con la media de las medias. Sólo son lo mismo en caso de que las muestras tengan el mismo tamaño.

Ejemplo de las tres marcas de ordenadores

Consideremos que se lleva a cabo un experimento para comparar el tiempo que tardan tres marcas de ordenadores de diferente marca en cargar un mismo sistema operativo.

Se toma una muestra de cuatro ordenadores de la marca A, es decir, se mide el tiempo (en segundos) que tardan en cargar el sistema operativo cuatro ordenadores de esta marca. De la marca B se toman seis medidas y cinco de la marca C. La tabla siguiente registra los resultados del experimento:

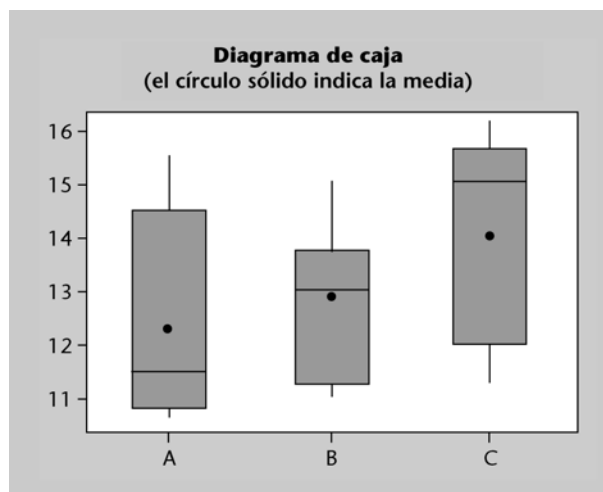
Marca A	10,7	11,2	12,0	15,5		
Marca B	13,4	11,5	11,2	15,1	13,3	12,9
Marca C	11,5	12,7	15,4	16,1	15,2	

Utilizando la notación que hemos presentado anteriormente, tenemos:

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

Observando estos resultados, podemos pensar que las muestras de los ordenadores A y B pueden provenir de poblaciones con la misma media, dado que las medias muestrales 12,35 y 12,90, respectivamente, son bastante cercanas. La media muestral de la marca C es 14,18; ésta está más alejada de las otras, pero presenta una mayor dispersión que las anteriores; no es tan fácil, pues, pensar si esta muestra proviene de una población con la misma media que los ordenadores de las marcas A y B.

Es posible representar esta situación mediante los diagramas de caja de las tres muestras:



Observación

La tarea de comparar más de dos muestras no es fácil.

3. La variabilidad de la muestra global: las sumas de cuadrados

Hemos visto que podemos considerar el conjunto global formado por todos los elementos de las muestras y, después, calcular la media de este conjunto global. A continuación intentaremos explicar a qué se deben las diferencias entre los valores de las observaciones x_{ij} y el valor de la media global \bar{x} . Entenderemos por **variabilidad** la diferencia entre los valores observados y la media. Veremos que esta variabilidad se debe a dos factores:

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

1) Variabilidad dentro de cada muestra: diferencia entre la observación y la media de la muestra ($x_{ij} - \bar{x}_j$).

2) Variabilidad entre las muestras: diferencia entre la media de la muestra y la media global ($\bar{x}_j - \bar{x}$).

Si existe mucha variabilidad entre las muestras, podremos pensar que este hecho se debe a que se trata de muestras extraídas de poblaciones diferentes o simplemente al origen aleatorio de las muestras. A continuación veremos cómo podemos separar estos dos efectos provocados por la variabilidad dentro de cada muestra y por la variabilidad entre las muestras.

Si sumamos al cuadrado la última expresión, todas las observaciones mediante un doble sumatorio, uno para las muestras y otro para las observaciones de cada muestra, tenemos:

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

donde podemos ver que el último sumando es cero:

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

ya que:

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

Con todo, la variabilidad de la muestra global se puede descomponer en dos partes:

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

Observación

Las medias muestrales pueden ser diferentes por el hecho de que provienen de poblaciones con medias diferentes o simplemente por el origen aleatorio de las muestras.

$$SCT = SCD + SCE$$

$$\begin{aligned} \text{Suma de cuadrados totales (SCT)} &= \\ &= \text{suma de cuadrados dentro de las muestras (SCD)} + \\ &+ \text{suma de cuadrados entre muestras (SCE)} \end{aligned}$$

Consideramos cada uno de estos sumandos:

- La Suma de Cuadrados Totales (SCT) nos informa de la variabilidad de la muestra global.
- La Suma de Cuadrados Dentro de las muestras (SCD) es una medida de la variación dentro de las muestras.
- La Suma de Cuadrados Entre muestras (SCE) es una medida de la variación entre las muestras; la calculamos a partir de la diferencia entre las medias de las muestras y la media total. Si las medias son muy diferentes, entonces esta cantidad es grande.

Obtención de la varianza de la muestra global

Es inmediato ver que si dividimos $SCT/(n-1)$, tenemos la varianza de la muestra global.

Si dividimos SCD y SCE por $n-k$ y $k-1$, respectivamente, obtenemos los estadísticos siguientes:

$$MCD = \frac{SCD}{n-k}; \quad MCE = \frac{SCE}{k-1}$$

que, como veremos a continuación, son necesarios para llevar a cabo el ANOVA.

4. Hipótesis sobre los datos para llevar a cabo el ANOVA

Para poder llevar a cabo un análisis de este tipo, hay que tener las hipótesis siguientes:

- 1) Las k muestras deben ser **aleatorias e independientes** entre sí.
- 2) Las poblaciones deben ser **normales**.
- 3) Las varianzas de las k poblaciones deben ser **idénticas**:

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

Bajo estas hipótesis y cuando se cumple $\mu_1 = \mu_2 = \dots = \mu_k = \mu$, es decir, si las medias poblacionales son todas iguales, las sumas de cuadrados SCE y SCD se distribuyen según distribuciones χ^2 con $(k-1)$ y $(n-k)$ grados de libertad, respectivamente.

Para muestras de una población normal $N(\mu, \sigma)$ siempre se cumple que:

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

tiene una distribución χ^2 con $n - 1$ grados de libertad.

Y dado que son independientes, una importante consecuencia es que el cociente entre estos estadísticos:

$$f = \frac{MCE}{MCD} = \frac{SCE/(k-1)}{SCD/(n-k)}$$

se distribuye según una distribución F de Snedecor con $(k - 1)$ grados de libertad en el numerador y $(n - k)$ en el denominador.

A continuación veremos cómo podemos utilizar esta descomposición de la variabilidad de los datos muestrales para construir un contraste de hipótesis que nos permita tomar una decisión sobre la igualdad de las medias de las poblaciones de procedencia de las muestras del estudio.

Cociente de variables aleatorias

Si X es una variable aleatoria que tiene una distribución χ^2 con n grados de libertad, Y es otra variable aleatoria que tiene una distribución χ^2 con m grados de libertad y X e Y son independientes, entonces la variable:

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

se distribuye según una F de Snedecor con n y m grados de libertad en el numerador y denominador, respectivamente.

5. El ANOVA es un contraste de hipótesis

El estadístico de contraste que utilizaremos en el análisis de la varianza se basa en el hecho de comparar los dos orígenes de la variabilidad de las muestras que hemos encontrado en el apartado anterior: la variación entre las muestras y la variación dentro de las muestras. Supondremos que se cumplen las hipótesis del modelo. Una vez hechos estos supuestos, procederemos de la manera siguiente:

1) Plantearemos nuestras hipótesis:

- Hipótesis nula: H_0 : todas las medias son iguales:

$$\mu_1 = \mu_2 = \dots = \mu_k = \mu$$

- Hipótesis alternativa: H_1 : no todas las medias son iguales.

2) Fijaremos un nivel significativo α .

3) Calcularemos el estadístico de contraste a partir de las sumas de cuadrados:

$$f = \frac{SCE/(k-1)}{SCD/(n-k)}$$

que, como hemos visto en el apartado anterior, si se cumple la hipótesis nula (igualdad de medias) es una observación de una distribución F de Snedecor con $n - k$ grados de libertad en el denominador y $k - 1$ grados de libertad en el numerador.

4) Finalmente, podemos actuar de dos maneras:

a) A partir del p -valor. Este valor es: $p = P(F > f)$:

- Si $p \leq \alpha$, se rechaza la hipótesis nula H_0 .
- Si $p > \alpha$, no se rechaza la hipótesis nula H_0 .

b) A partir del valor crítico $F_{\alpha, k-1, n-k}$, que separa la región de aceptación de la región de rechazo:

- Si $f > F_{\alpha, k-1, n-k}$, se rechaza la hipótesis nula H_0 .
- Si $f \leq F_{\alpha, k-1, n-k}$, no se rechaza la hipótesis nula H_0 .

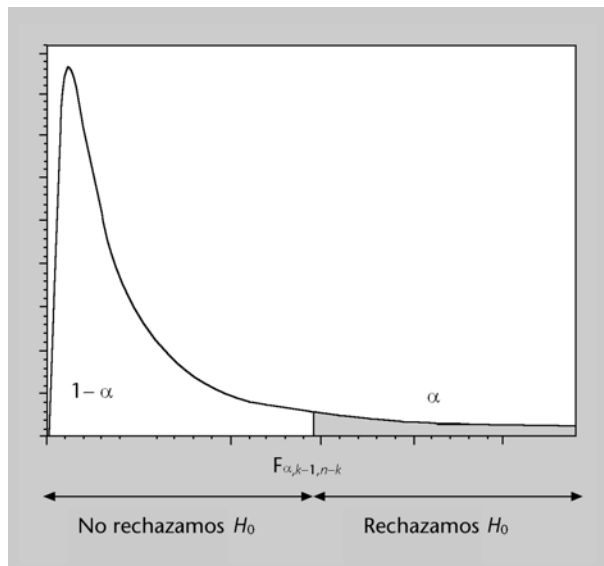
El p -valor

El p -valor es la probabilidad del resultado observado o de otro más alejado si la hipótesis nula es cierta.

No rechazar o rechazar la H_0

No rechazar la H_0 no significa exactamente que aceptemos la hipótesis, sino simplemente que nada se opone a pensar que la H_0 pueda ser cierta.

Rechazar la H_0 no significa necesariamente que todas las medias sean diferentes, sino que significa que alguna (quizá todas) es diferente de otra.



Si queremos determinar cuáles son los grupos que presentan unas diferencias lo bastante significativas, haremos pruebas t de Student para comparación de medias, tal como se planteaba al inicio de la sesión.

6. Construcción de la tabla del ANOVA

En este apartado estamos interesados en plantear una forma conveniente y habitual de presentar los cálculos y resultados del análisis de la varianza. Esta manera de sintetizar estos cálculos es en forma de tabla, llamada *tabla del ANOVA*. Podemos llevar a cabo todos los cálculos de la tabla a partir de las medias \bar{x}_j y varianzas s_j^2 de las diferentes muestras o de la media \bar{x} de la muestra global.

Podemos escribir las sumas de cuadrados que necesitamos para calcular el estadístico de contraste de la manera siguiente:

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

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

$$SCT = SCD + SCE$$

Tabla del análisis de la varianza				Regla de decisión
Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	Estadístico de prueba
Entre grupos	$SCE = \sum_j n_j (\bar{x}_j - \bar{x})^2$	$k - 1$	$SCE / (k - 1)$	$f = \frac{SCE / (k - 1)}{SCD / (n - k)}$
Dentro de los grupos	$SCD = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$	$n - k$	$SCD / (n - k)$	
Total	$SCT = \sum_j \sum_i (x_{ij} - \bar{x})^2$	$n - 1$	-	

k = número de muestras, n = tamaño de la muestra total

Ejemplo de las tres marcas de ordenadores (II)

Volvamos al ejemplo de la comparación de tiempo de carga de un sistema operativo para tres marcas diferentes de ordenadores.

Haremos un análisis de la varianza con nivel significativo $\alpha = 0,05$ para determinar si podemos considerar que las medias de los tiempos de los tres ordenadores son iguales.

Hipótesis nula: $H_0: \mu_1 = \mu_2 = \mu_3$.

Hipótesis alternativa: H_1 : no todas las medias son iguales.

Ya habíamos calculado:

Muestra	A $j = 1$	B $j = 2$	C $j = 3$
Tamaño	$n_1 = 4$	$n_2 = 6$	$n_3 = 5$
Media	$\bar{x}_1 = 12,35$	$\bar{x}_2 = 12,90$	$\bar{x}_3 = 14,18$
Varianza	$s_1^2 = 4,70$	$s_2^2 = 2,02$	$s_3^2 = 3,90$

De manera que la media de la muestra global es:

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

A continuación podemos encontrar las sumas de cuadrados:

$$\bullet \quad SCD = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^k (n_j - 1) s_j^2 = (4 - 1)4,70 + (6 - 1)2,02 + (5 - 1)3,90 = 39,80$$

$$\bullet \quad SCE = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 = 4(12,35 - 13,18)^2 + 6(12,90 - 13,18)^2 + 5(14,18 - 13,18)^2 = 8,226$$

Y construir la tabla del ANOVA:

Tabla del análisis de la varianza				Regla de decisión
Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	Estadístico de prueba
Entre grupos	SCE = 8,226	$3 - 1 = 2$	$8,226/2 = 4,113$	$f = \frac{4,113}{3,32} = 1,24$
Dentro de los grupos	SCD = 39,80	$15 - 3 = 12$	$39,80/12 = 3,32$	
Total	SCT = 48,026	$15 - 1 = 14$	-	

$$\text{Estadístico de contraste: } f = \frac{SCE/(k-1)}{SCD/(n-k)} = \frac{4,113}{3,32} = 1,24$$

Este estadístico sigue una distribución F de Snedecor con $k - 1 = 2$ y $n - k = 12$ grados de libertad en el numerador y en el denominador, respectivamente.

1) A partir del p -valor:

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

Por tanto, no rechazamos la hipótesis nula.

2) A partir del valor crítico:

Para un nivel significativo $\alpha = 0,05$, tenemos un valor crítico:

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

Si comparamos este valor con el estadístico de contraste, $f = 1,24$, tenemos que $1,24 < 3,89$ y, por tanto, no rechazamos la hipótesis nula.

Así pues, podemos concluir que no hay una diferencia significativa entre los tiempos que tardan las tres marcas de ordenadores en cargar el sistema operativo.

7. Resumen

En esta sesión hemos presentado la técnica de análisis de la varianza (ANOVA) para la comparación de las medias para más de dos muestras. Hemos comprobado que la variación de las observaciones se debe a dos factores: la variabilidad dentro de cada muestra y la variabilidad entre las muestras. Hemos expresado estas variaciones de forma numérica mediante sumas de cuadrados. Con las sumas de cuadrados hemos podido encontrar un estadístico de prueba para contrastar la igualdad de las medias de las muestras. Finalmente, hemos aprendido a resumir todos los cálculos en la llamada *tabla del ANOVA*.

Ejercicios

1. Consideremos cuatro compañías (A, B, C y D), cuyas acciones cotizan en bolsa. Seleccionamos de forma aleatoria las cotizaciones de estas acciones durante diferentes instantes de tiempo a lo largo de un mes. Así pues, para la compañía A se observa la cotización en cinco instantes aleatorios, para la B se observa en cuatro, para la C se observa en seis y, finalmente, para la compañía D, en cinco.

En la tabla siguiente se da la cotización en céntimos de euro de las diferentes acciones en los instantes de tiempo seleccionados:

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

Contrastad el nivel del 5% si las cotizaciones medias de las acciones de cada una de las cuatro compañías se pueden considerar iguales. Confeccionad la tabla de análisis de la varianza.

2. Los estudiantes de segundo curso de una escuela universitaria de ingeniería estuvieron repartidos de forma aleatoria en tres grupos. En cada grupo se enseñó estadística con una estrategia docente diferente. Al final del curso todos los alumnos hicieron el mismo examen. Se seleccionaron de forma aleatoria algunas cualificaciones obtenidas por algunos alumnos de los tres grupos. Los resultados son los siguientes:

$$\text{Grupo 1: } n = 5 \quad \sum x_i = 116 \quad \sum x_i^2 = 2.728$$

$$\text{Grupo 2: } n = 5 \quad \sum x_i = 126 \quad \sum x_i^2 = 3.294$$

$$\text{Grupo 3: } n = 7 \quad \sum x_i = 171 \quad \sum x_i^2 = 4.193$$

- ¿Sobre qué supuestos podréis hacer un análisis de la varianza?
- Haced un análisis de la varianza e indicad si podéis asegurar a un nivel significativo del 0,05 que el resultado obtenido depende de la técnica de enseñanza utilizada.

Solucionario

1. En este problema estamos interesados en comparar cuatro poblaciones, y utilizaremos un análisis de la varianza o ANOVA. Sabemos que el ANOVA nos permite comparar las medias de varios grupos.

Hipótesis nula: H_0 : las medias son iguales: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Hipótesis alternativa: H_1 : las medias no son iguales.

Para construir la tabla del análisis de la varianza:

Primero calculamos las medias y las varianzas de cada muestra:

	J	n_j	Media \bar{x}_j	Varianza 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

Y la media total de las muestras:

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

A continuación podemos encontrar las sumas de cuadrados:

$$\begin{aligned} \bullet \quad SCD &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^k (n_j - 1) s_j^2 = \sum_{j=1}^4 (n_j - 1) s_j^2 \\ &= (5 - 1) 14.250,00 + (4 - 1) 7.233,33 + (6 - 1) 2.200,00 + (5 - 1) 5.350,00 = \\ &= 111.100,00 \end{aligned}$$

En esta expresión hemos tenido en cuenta que: $s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$

$$\begin{aligned} \bullet \quad SCE &= \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 = 5(760,00 - 774,50)^2 + 4(685,00 - 774,50)^2 + 6(750,00 - \\ &- 774,50)^2 + 5(890,00 - 774,50)^2 = 103.395,00 \end{aligned}$$

Ahora ya podemos construir la tabla:

Tabla del análisis de la varianza			
Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
Entre grupos	$SCE = \sum_j n_j (\bar{x}_j - \bar{x})^2$ $SCE = 103.395,00$	$k - 1$ $4 - 1 = 3$	$SCE / (k - 1) =$ $= 103.395,00 / 3 = 34.465,00$
Dentro de los grupos	$SCD = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$ $SCD = 111.100,00$	$n - k$ $20 - 4 = 16$	$SCD / (n - k) = 111.100,00 / 16 =$ $= 6.943,75$
Total	$SC T = \sum_j \sum_i (x_{ij} - \bar{x})^2$	$20 - 1 = 19$	-

k = número de grupos = 4, n = tamaño de la muestra total = 20

$$\text{Estadístico de contraste: } f = \frac{SCE/(k-1)}{(SCD)/(n-k)} = \frac{34.465,00}{6.943,75} = 4,96$$

El estadístico sigue una distribución F de Snedecor con $k - 1 = 3$ y $n - k = 16$ grados de libertad.

Calculamos el p -valor:

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

de manera que rechazamos H_0 . Con una confianza del 95%, existe diferencia significativa entre las cuatro compañías.

2.

a) Para poder aplicar esta técnica con fiabilidad, son necesarias las restricciones previas que presentamos a continuación:

1. Las muestras deben ser independientes.
2. Las poblaciones (o subpoblaciones) siguen distribuciones normales; por otra parte, la muestra elegida debe ser lo bastante grande (más de treinta observaciones en cada submuestra). En nuestro caso, las muestras son menores que treinta. Por tanto, para poder aplicar el ANOVA, debemos suponer que las poblaciones siguen distribuciones normales.
3. La varianza para cada población (o subpoblación) es la misma.

b) En este problema estamos interesados en comparar tres poblaciones y utilizaremos un análisis de la varianza o ANOVA. Sabemos que el ANOVA nos permite comparar las medias de varios grupos.

Hipótesis nula: H_0 : las medias son iguales: $\mu_1 = \mu_2 = \mu_3$

Hipótesis alternativa: H_1 : las medias no son iguales

Para realizar la tabla del análisis de la varianza, en primer lugar calculamos las medias y las varianzas de cada muestra:

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

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

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

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

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

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

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

A continuación podemos encontrar las sumas de cuadrados mediante estas expresiones:

$$SCD = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^k (n_j - 1) s_j^2 = \sum_{j=1}^3 (n_j - 1) s_j^2 = (5 - 1) 9,2 + (5 - 1) 4,3 + (7 - 1) 4,25 = 79,50$$

En esta expresión hemos tenido en cuenta que:

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

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

Ahora ya podemos calcular el estadístico de contraste. Primero construimos la tabla de análisis de la varianza:

Tabla del análisis de la varianza			
Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
Entre grupos	$SCE = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2$ $SCE = 14,40$	$k - 1$ $3 - 1 = 2$	$SCE / (k - 1) = 14,40 / 2 = 7,202$
Dentro de los grupos	$SCD = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$ $SCD = 79,50$	$n - k$ $17 - 3 = 14$	$SCD / (n - k) = 79,5 / 14 = 5,679$
Total	$SC T = \sum_j \sum_i (x_{ij} - \bar{x})^2$	$n - 1 = 16$	-

k = número de grupos = 3, n = tamaño de la muestra total = 17

Estadístico de contraste: $f = \frac{SCE/(k-1)}{(SCD)/(n-k)} = 1,268$

El estadístico sigue una distribución F de Snedecor con $k - 1 = 2$ y $n - k = 14$ grados de libertad. Tenemos un p -valor:

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

No rechazamos H_0 , de manera que podemos asegurar a un nivel significativo del 0,05 que el resultado obtenido no depende de la estrategia docente utilizada.

