Análisis de la Varianza

En este capítulo haremos una introducción al Análisis de Varianza univariante del modelo de efectos fijos con muestras independientes. La perspectiva que se seguirá será la de considerarlo como un contraste de hipótesis de igualdad de medias, dejando a un lado cualquier consideración sobre el diseño experimental mediante el que se han obtenido los datos y sobre el modelo lineal de los efectos que implica, aspectos que se dejan para posteriores asignaturas. Nuestro objetivo es que este tema sirva como nexo de unión entre la inferencia estadística y los temas de diseño experimental, dando idea de la continuidad que existe entre todos los conceptos que se engloban dentro de la estadística Matemática.

Formulación.

La situación que afronta el Análisis de la Varianza es la siguiente:

Supongamos que tenemos k muestras independientes procedentes de poblaciones Normales $N(\mu_1, \sigma)$ $N(\mu_2, \sigma)$ ... $N(\mu_k, \sigma)$ con la misma varianza y se trata de contrastar la hipótesis nula de que las medias de todas esas poblaciones son iguales

$$H_0 : \ \mu_1 = \mu_2 = \mu_k = \mu$$

Frente a la alternativa de que al menos una de las medias de esas poblaciones es distinta de las demás

$$H_1 : \exists i \ / \ \mu_i \neq \mu$$

Obviamente, la decisión debe tomarse en función de la información que proporcionan las k muestras, para cuyos valores adoptaremos la siguiente notación:

$N(\mu_1, \sigma)$	$N(\mu_2, \sigma)$	...	$N(\mu_k, \sigma)$
$\downarrow$	$\downarrow$		$\downarrow$
x_{11}	x_{21}	...	x_{k1}
x_{12}	x_{22}	...	x_{k2}
...	...	...	...
x_{1n1}	x_{2n2}	...	x_{knk}

Cada una de estas muestras tendrá su correspondiente media que de forma general podemos expresar:

$$\overline{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

e igualmente su correspondiente varianza que de forma genérica será:

$$S_i^2 = \frac{\sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i\right)^2}{n_i}$$

También podemos considerar la media conjunta de todos los valores

$$\bar{\bar{x}} = \frac{\sum_i \sum_j x_{ij}}{n}$$

y la varianza total

$$S^2 = \frac{\sum_i \sum_j \left(x_{ij} - \bar{\bar{x}}\right)^2}{n}$$

Verificándose que la varianza total se puede descomponer en la media de las varianzas, más la varianza de las medias.

En efecto, sumando y restando las medias de las muestras tenemos que:

$$S^2 = \frac{\sum_i \sum_j \left(x_{ij} - \bar{\bar{x}}\right)^2}{n} = \frac{\sum_i \sum_j \left[\left(x_{ij} - \bar{x}_i\right) + \left(\bar{x}_i - \bar{\bar{x}}\right)\right]^2}{n}$$

Desarrollando el cuadrado:

$$= \frac{\sum_i \sum_j \left[\left(x_{ij} - \bar{x}_i\right)^2 + \left(\bar{x}_i - \bar{\bar{x}}\right)^2 + 2\left(x_{ij} - \bar{x}_i\right)\left(\bar{x}_i - \bar{\bar{x}}\right)\right]}{n}$$

Por la propiedad asociativa de la suma, podemos escribirlo como:

$$= \frac{\sum_i \sum_j \left(x_{ij} - \bar{x}_i\right)^2}{n} + \frac{\sum_i \sum_j \left(\bar{x}_i - \bar{\bar{x}}\right)^2}{n}$$

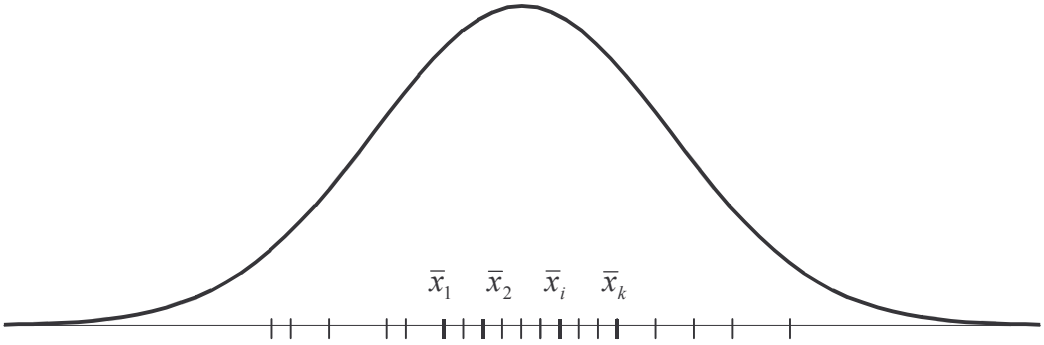
ya que el término correspondiente al doble producto es nulo por ser la suma de desviaciones a la media, llegándose a la identidad:

$$\frac{\sum_i \sum_j \left(x_{ij} - \bar{\bar{x}}\right)^2}{n} = \frac{\sum_i \sum_j \left(x_{ij} - \bar{x}_i\right)^2}{n} + \frac{\sum_i n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2}{n}$$

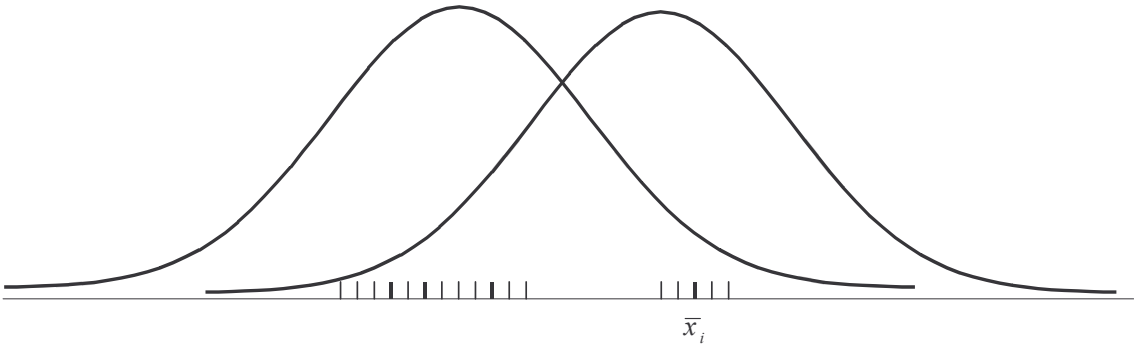
Que también podemos escribir como:

$$S^2 = \frac{\sum_i n_i S_i^2}{n} + S_{\bar{x}_i}^2$$

La idea intuitiva del Análisis de la varianza consiste en comparar la importancia relativa de ambos términos de la descomposición. Si la varianza de las medias es pequeña con relación a la media de las varianzas de las muestras esto nos indica que todas las muestras pueden proceder de una misma distribución Normal, como se muestra en la figura siguiente, y por consiguiente que se verifica la hipótesis nula.



Por el contrario, si la varianza de las medias es grande en comparación con la media de las varianzas de las muestras, situación que refleja el siguiente gráfico, es poco probable que todas las muestras procedan de una misma población Normal y es más plausible que alguna de ellas proceda de una población con distinta media.



Esta comparación entre ambos términos mediante su cociente es el fundamento del estadístico de contraste del Análisis de la varianza que deduciremos seguidamente.

Teorema de Cochran

La determinación de los estadísticos de contraste a emplear en todos los modelos de Análisis de la varianza, cualquiera que sea su complejidad, y la especificación de la distribución de estos estadísticos bajo la hipótesis nula, se basa en el teorema de Cochran que enunciamos a continuación:

Sean $y_1, y_2, \ldots, y_n$ variables aleatorias independientes con distribución Normal cero, uno y sean $Q_1, Q_2, \ldots, Q_h$ h formas cuadráticas en las y_i siendo n_j la característica de Q_j de forma que se cumple:

$$\sum_{j=1}^{h} Q_j \quad = \quad \sum_{i=1}^{n} y_i^2$$

Entonces se verifican las siguientes proposiciones:

1ª Si la suma de las características de las formas cuadráticas n_j es igual al número de variables n, cada forma cuadrática Q_j sigue una distribución Ji-cuadrado con n_j grados de libertad y las formas cuadráticas son independientes.

2ª Si cada forma cuadrática Q_j sigue una Ji-cuadrado con n_j grados de libertad, las formas cuadráticas son independientes y la suma de sus características es igual a n.

3ª Si las Q_j son independientes, cada Q_j sigue una Ji-cuadrado con n_j grados de libertad y se verifica que la suma de las n_j es igual a n.

Estadístico de Contraste.

Veamos como aplicando el anterior teorema podemos deducir el estadístico de contraste y su distribución en el caso que nos ocupa.

Cada observación x_{ij} es independiente de las restantes y sigue una distribución Normal $N(\mu_i, \sigma)$, por consiguiente el estadístico:

$$\sum_i \sum_j \left(\frac{x_{ij} - \mu_i}{\sigma} \right)^2$$

es la suma de n variables independientes N(0, 1) al cuadrado. Por otra parte, sumando y restando términos se tiene la siguiente igualdad:

$$\sum_i \sum_j \left(\frac{x_{ij} - \mu_i}{\sigma} \right)^2 = \frac{\sum_i \sum_j \left[(x_{ij} - \bar{x}_i) + [(\bar{x}_i - \bar{\bar{x}}) - (\mu_i - \mu)] + (\bar{\bar{x}} - \mu) \right]^2}{\sigma^2}$$

Desarrollando el cuadrado se obtiene:

$$= \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{\sigma^2} + \frac{\sum_i n_i [(\bar{x}_i - \bar{\bar{x}}) - (\mu_i - \mu)]^2}{\sigma^2} + \frac{n(\bar{\bar{x}} - \mu)^2}{\sigma^2}$$

ya que los dobles productos se anulan por contener todos ellos una suma de diferencias a la media. Hemos obtenido así una descomposición de la suma de los cuadrados de las n variables N(0,1) en tres formas cuadráticas. La primera de ellas contiene n sumandos pero existen entre ellos k ligaduras correspondientes a la condición de que la suma de las diferencias a la media en cada una de las k muestras es cero, por consiguiente su característica es n-k. La segunda contiene k sumandos y una ligadura, por consiguiente su característica es k-1. La tercera forma tiene un solo sumando y su característica es 1. Se cumplen por tanto las condiciones del teorema de Cochran y podemos afirmar que estas formas cuadráticas siguen distribuciones Ji-cuadrado con n-k, k-1 y 1 grados de libertad y que son independientes. Por ello el estadístico:

$$\frac{\frac{\sum_i n_i [(\bar{x}_i - \bar{\bar{x}}) - (\mu_i - \mu)]^2}{\sigma^2 (k-1)}}{\frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{\sigma^2 (n-k)}} = \frac{\frac{\sum_i n_i [(\bar{x}_i - \bar{\bar{x}}) - (\mu_i - \mu)]^2}{k-1}}{\frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{n-k}}$$

Sigue una distribución F de Snedecor con k-1 y n-k grados de libertad.

Si ahora definimos el estadístico:

$$F = \frac{\frac{\sum_i n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2}{k-1}}{\frac{\sum_i \sum_j \left(x_{ij} - \bar{x}_i\right)^2}{n-k}}$$

Tendremos que si se verifica la hipótesis nula, todas las diferencias $\mu_i - \mu$ serán nulas y por consiguiente este estadístico coincidirá con el anterior y su distribución será una F de Snedecor con k-1 y n-k grados de libertad. Por el contrario, bajo la hipótesis alternativa alguna de las diferencias $\mu_i - \mu$ será distinta de cero y este estadístico diferirá del anterior, su distribución no será la F de Snedecor mencionada sino una distribución que estará desplazada a la derecha.

Expresiones de Cálculo

Como siempre que aparecen cuadrados de diferencias a la media, que en general serán fracciones decimales, pueden encontrarse expresiones más útiles para el cálculo manual desarrollando las fórmulas iniciales. Así la suma de cuadrados que aparece en el numerador del estadístico que corresponde a la variabilidad entre grupos, adopta la siguiente expresión:

$$\sum_i n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2 = \sum_i n_i \bar{x}_i^2 + \sum_i n_i \bar{\bar{x}}^2 - 2\sum_i n_i \bar{x}_i \bar{\bar{x}} = \sum_i n_i \bar{x}_i^2 + n\bar{\bar{x}}^2 - 2n\bar{\bar{x}}^2$$

$$= \sum_i n_i \bar{x}_i^2 - n\bar{\bar{x}}^2 = \sum_i \frac{\left(\sum_j x_{ij}\right)^2}{n_i} - \frac{\left(\sum_i \sum_j x_{ij}\right)^2}{n}$$

De forma análoga la suma de cuadrados del denominador que corresponde a la variabilidad dentro de los grupos, o intragrupo, puede calcularse mediante la expresión:

$$\sum_i \sum_j \left(x_{ij} - \bar{x}_i\right)^2 = \sum_i \sum_j x_{ij}^2 + \sum_i \sum_j \bar{x}_i^2 - 2\sum_i \sum_j x_{ij} \bar{x}_i = \sum_i \sum_j x_{ij}^2 + \sum_i n_i \bar{x}_i^2 - 2\sum_i \bar{x}_i \sum_j x_{ij}$$

$$= \sum_i \sum_j x_{ij}^2 - \sum_i n_i \bar{x}_i^2 = \sum_i \sum_j x_{ij}^2 - \sum_i \frac{\left(\sum_j x_{ij}\right)^2}{n_i}$$

También es conveniente disponer de una expresión para la suma de cuadrados que representa la variabilidad total:

$$\sum_i \sum_j \left(x_{ij} - \bar{\bar{x}}\right)^2 = \sum_i \sum_j x_{ij}^2 + \sum_i \sum_j \bar{\bar{x}}^2 - 2\sum_i \sum_j x_{ij} \bar{\bar{x}} = \sum_i \sum_j x_{ij}^2 + n\bar{\bar{x}}^2 - 2\bar{\bar{x}} n\bar{\bar{x}}$$

$$= \sum_i \sum_j x_{ij}^2 - n\bar{\bar{x}}^2 = \sum_i \sum_j x_{ij}^2 - \frac{\left(\sum_i \sum_j x_{ij}\right)^2}{n}$$

Ejemplo:

Se quiere comprobar si tiene efecto sobre la adquisición de conocimientos de Análisis de Datos la modalidad de bachillerato que ha cursado el alumno. Para ello se registró el número de errores de 22 alumnos en una prueba objetiva, clasificándolos de acuerdo con el tipo de Bachillerato, se obtuvieron los siguientes datos:

A	B	C
9	6	3
8	7	4
8	5	8
9	6	3
7	5	6
7	8	3
	4	7
	7	6

La tabla siguiente muestra los cálculos básicos:

	A	B	C	Totales
$\sum_j x_{ij}$	48	48	40	136
$\sum_j x_{ij}^2$	388	300	228	916
$\frac{\left(\sum_j x_{ij}\right)^2}{n_i}$	384	288	200	872

A partir del total de los valores, 136, tenemos:

$$\frac{\left(\sum_i \sum_j x_{ij}\right)^2}{n} = \frac{136^2}{22} = 840{,}73$$

Por consiguiente las sumas de cuadrados serán:

$$SC_{entre} = 872 - 840{,}73 = 31{,}27$$

$$SC_{intra} = 916 - 872 = 44$$

$$SC_{total} = 916 - 840{,}73 = 75{,}27$$

El resto de las operaciones necesarias para obtener el estadístico aparecen en el cuadro siguiente:

Fuente de variación	SC	g.l.	CM	F
Entre grupos	31,27	2	15,64	6,74
Intra grupo	44	19	2,32	
Total	75,27			

Buscando en las tablas de la F de Snedecor con 2 y 19 grados de libertad, el valor que deja por debajo una probabilidad de 0,95 es 3,52. Por consiguiente, el valor del estadístico es mayor que el valor crítico y se rechaza la hipótesis nula de que todas las puntuaciones proceden de poblaciones con la misma media, al menos uno de los tres tipos de Bachillerato tiene un número medio de errores significativamente distinto de los otros.