Tema 9

Análisis de la Varianza (ANOVA)

Conceptos generales

La técnica del **Análisis de la Varianza** consiste en descomponer la variabilidad de una población (representada por su varianza) en diversos sumandos según los factores que intervengan en la creación de esa variabilidad. Por ejemplo, si estudiamos la varianza que presenta una colección de calificaciones que provienen de tres asignaturas en cuatro cursos distintos a lo largo de los últimos años, la varianza total se puede descomponer en cuatro sumandos:

- Parte proveniente del factor asignatura
- Componente aportado por los distintos cursos
- Influencia de la evolución temporal en los últimos años
- Varianza propia (interna) de la población.

Este ejemplo sería bastante complejo, porque depende de **tres factores**: asignatura, curso y año. Son mucho más frecuentes los ejemplos de **un solo factor** (por ejemplo, tres métodos distintos aplicados simultáneamente a alumnado del mismo nivel) o de **dos factores** (lo sería la influencia del sexo y la edad en un rendimiento)

Lo original del Análisis de la Varianza es que su verdadero objetivo no es la variabilidad, sino otros contrastes, como la igualdad de medias o el ajuste en un problema de Regresión. Lo veremos en los casos que vamos a estudiar.

Análisis de Varianza de un factor

Modelo y supuestos

Supongamos la existencia de varias muestras distintas que corresponden a los resultados obtenidos en una población bajo la influencia de distintos niveles de un factor. La palabra niveles no se debe interpretar en sentido ordinal. Pueden ser niveles distintos métodos de enseñanza, lugares de nacimiento o sexo. Se consideran igualmente válidos niveles cualitativos o cuantitativos (fijos).

Para fijar ideas, supongamos un experimento consistente en medir los minutos transcurridos en la desaparición de un dolor después de la administración de tres tipos de analgésicos a una muestra de pacientes con migraña

Analgésico	A	B	C
	12	11	14
	18	12	18
	14	13	17
	10	12	21
	21	8	17
	15	15	16
		18	19

		11	21

En este caso el factor es el tipo de analgésico, que actúa a través de tres niveles distintos A. B y C.

En la tabla se observa que dentro de cada nivel existe bastante variabilidad (cada paciente tendrá su forma de reaccionar), y que parece que también existen diferencias entre unos niveles y otros. Si calculáramos las medias nos resultaría

m_1=15; m_2=12,5; m_3=17.875

Si las medias fueran iguales, negaríamos que existan diferencias en el efecto de los distintos analgésicos, pero como no lo son, deberemos plantearnos un contraste de hipótesis para la igualdad de medias.

De hecho, el verdadero contraste que se propone el Análisis de la Varianza es el de **igualdad de medias**. Plantearemos la hipótesis nula:

H_0: $m_1=m_2=m_3$

Pero en realidad la contrastaremos descomponiendo la varianza. Para ello supondremos que cada medida de minutos se puede descomponer en tres sumandos:

$a_{ij} = m + a_i + e_{ij}$

a_{ij}: Es la medida real que se observa en los sujetos (12, 18, 13, 10...) y se considera descompuesta en tres factores aditivos

m: Es la media general de todo el experimento. En el ejemplo equivaldría a 15,14.

a_i: Mide la influencia del factor, mediante la diferencia entre la media de cada columna y la media general. En el ejemplo se darían estas diferencias: a_1=15-15,14=-0,14 a_2=12,5-15,14=-2,64 a_3=17,875.15,14=2,735

e_{ij}: Mide la variación propia de cada individuo.

Para entenderlo mejor descompondremos dos datos:

La medida 8 de la segunda columna equivale a 8=15,14-2,64-4,5. En esta suma 15,14 es la media general del experimento, -2,64 la influencia del medicamento B y -4,5 la diferencia aportada por el individuo, que ha reaccionado muy rápido.

La medida 19 de la tercera columna equivale a 19=15,14+2,735+1,125. El factor medicamento aporta 2,735, porque es el más lento en actuar, y el individuo 1,125, que no es tan rápido como el anterior.

Modelo

El conjunto de supuestos más aceptado en este caso, porque permite inferencias muy simples, es el siguiente:

Se trabaja sobre una variable aleatoria Y_{ij}, a la que se le supone descompuesta de la siguiente forma:

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ donde μ es la media de la población, α_i la influencia del factor, que equivale a la diferencia entre la media general y la del grupo. Finalmente, ε_{ij} se corresponde con la diferencia propia de cada individuo.

Se supone que todas las Y son **normales e independientes**.

Llamamos n_i al número de sujetos por grupo, y N al número total, con lo que $n_1+n_2+n_3\ldots=N$

Estimadores

La media general μ se estima mediante $$m = \frac{\sum_i \sum_j Y_{ij}}{N}$$

Las medias de cada grupo o nivel de forma similar: $$m_i = \frac{\sum_j Y_{ij}}{n_i}$$

La influencia del factor (α_i) se estima mediante la diferencia $\alpha_i = m_i - m$

Para la estimación de la varianza deberemos antes abordar la operación fundamental del Análisis de la Varianza, que consiste en descomponer en sumandos la suma de cuadrados de los datos corregida con la media. Se distinguen tres sumas distintas:

Suma de cuadrados total (SCT)

Viene dada por la fórmula $$SCT = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$$

Que coincide con el numerador de la varianza total. Esta fórmula se puede simplificar mediante esta otra:

$$SCT = \sum_I \sum_J Y_{IJ}^2 - N\bar{Y}^2$$

En el ejemplo de arriba el valor sería

SCT = 5339 – 229,109504*22 = 298,59

Si esta suma la dividimos entre los grados de libertad, que son N-1, nos resultará la Media cuadrática Total. En este caso: MCT = 298,59/(22-1)= 14,22

Suma de cuadrados Intra o de error (SCE)

Representa la suma de cuadrados corregidos que se da dentro de los grupos, es decir, las diferencias de los datos entre la media de cada grupo.

$$SCE = \sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2$$

Y su expresión reducida

$$SCE = \sum_i \left(\sum_j Y_{ij}^2 - n_i \overline{Y}_i^2 \right)$$

En el ejemplo su valor sería

SCE=(1430-15^2*6)+(1312-12,5^2*8)+(2597-17,88^2*8) = 182,88

Si lo dividimos esta suma entre los grados de libertad N-n nos resultará la media cuadrática de error, que es el mejor estimador de la varianza de la población.

MCE = 182,88/(22-3)=9,63

Suma de cuadrados Inter (entre grupos)

Se define mediante la fórmula

$$SCI = \sum_i n_i (\overline{Y}_i - \overline{Y})^2$$

Aunque también se puede calcular restando, ya que se demuestra que

SCT = SCI + SCE

En el ejemplo valdría

SCI = 298,59 – 182,88 = 115,72

También se puede hallar la media cuadrática Inter dividiendo entre los grados de libertad i-1

MCI=115,72/2 =57,86

En la práctica se forman tres sumas de cuadrados:

$$S1 = \sum_i \sum_j Y_{ij}^2$$

que consiste en sumar todos los datos por separado elevados al cuadrado. En el ejemplo tendría un valor de 5339.

$$S2 = \sum_i \frac{\left(\sum_j Y_{ij}\right)^2}{n_i}$$

que equivale a sumar los datos de cada nivel, elevar al cuadrado y dividir entre el número de datos. En el ejemplo:

$S2=90^2/6+100^2/8+143^2/8= 5156,13$

Y por último, la suma S3 equivale al cuadrado de la suma total de datos dividida entre el número total de los mismos.

$S3 = 333^2/22 =5040,41.$

De esta forma, la suma de cuadrados total es la diferencia entre S1 y S3 (se puede demostrar)

SCTotal = S1 – S3 = 5339 – 5040,41 = **298,59**

De igual forma, la suma de cuadrados Intra es la diferencia entre S1 y S2

SCIntra = S1 – S2 = 5339 – 5156,13 = **182,88**

Y la otra diferencia será la suma Inter:

SCInter= S2-S3 = 5156,13 - 5040,41 = **115,72**

Análisis de Varianza de dos factores

Modelo y supuestos

Supongamos la existencia de varias muestras distintas que corresponden a los resultados obtenidos en una población bajo la influencia de distintos niveles de dos factores.

Por ejemplo, imaginemos que las medidas de la tabla siguiente se han obtenido en tres barrios distintos A,B y C y en tres niveles de edad: 10-30, 31-50, 51-70. Podemos imaginar las medidas como una valoración que se ha recogido en una encuesta:

	Barrio A	Barrio B	Barrio C
10-30	3,4,4,5,4 2,4,5,3,1	6,2,3,4,5,4 4,5,6,2,7	2,4,5,6,6 3,4,3
31-50	6,8,4,6,9 7,3,4,8,7	8,9,6,7,7 10,6,9,8,7	5,7,5,6,6 3,5,4,6
51-70	4,2,2,4,5 1,3,2,4,5	6,6,4,5,3,8 4,0,1,4	5,6,4,5,3 5,2,1,1,1,0

Al igual que en el caso de un factor, podemos descomponer las medidas en varios sumandos:

$a_{ijk} = m + a_i + b_j + ab_{ij} + e_{ijk}$

a_{ijk} es una medida cualquiera, individual, que la consideramos descompuesta en cuatro sumandos:

m: Es la media total de toda la tabla.

a_i : Mide el efecto del factor A. En el ejemplo podría ser el barrio, que influyera en la valoración efectuada por los sujetos.

b_j : Mide el efecto del otro factor B, en nuestro caso el nivel de edad.

ab_{ij} : Puede que los efectos de A y B no sean aditivos sin más, sino que exista interacción entre ellos. Este sumando mide dicha influencia mutua. Si se supone que A y B son independientes, valdrá 0, y consideraremos un modelo **sin interacción.**

e_{ijk} : Contiene las diferencias individuales. Se supone que su distribución es Normal de media 0.

La hipótesis nula en este caso es la de que todas las medias de los subgrupos son iguales.

Como en el caso anterior, el análisis se basa en sumas de cuadrados y en grados de libertad, para después dividirlos, obtener estimadores de la varianza y compararlos mediante un contraste F.

Si se ha entendido el modelo de un factor, para abordar éste hay que considerar que existen cuatro fuentes de variación en este problema.

Explicaremos cada fuente mediante la resolución que del ejemplo propuesto nos brinda la hoja de cálculo. En unos temas prácticos como estos, no llenaremos la teoría de sumatorios, remitiendo a manuales específicos el estudio detallado de los mismos.

Fuente variación	SC	G.L.	CM	F	
Factor A	29,26	2	14,63	5,05	P-valor de FA 0,165
Factor B	149,04	2	74,52	25,73	P-valor de FB 0,005 Significativa al 5%
Interacción AB	11,65	4	2,91	1,01	P-valor de FAB 0,410
Error	231,68	80	2,9		
TOTAL	421,62	88			

Fuente de variación Barrio: SCA=29,26. Esta suma representa la variabilidad de los tres grupos formados por los barrios. Se consigue de forma similar a la de un factor. Sus grados de libertad son 2, equivalentes al número de barrios menos 1.

Fuente de variación Edad: SCB= 149,04. Representa la variabilidad entre edades. Como existen tres niveles, sus grados de libertad también son cuatro.

Interacción: SCAB=11,65. En algunos modelos no se considera que haya influencia entre los dos factores. Esta decisión se debe tomar teniendo en cuenta conocimientos anteriores, y no como consecuencia de los datos obtenidos en el ANOVA. En estos temas usaremos siempre modelos con interacción.

Sus G.L. se calculan multiplicando los de los dos factores.

Error: SCE=231,68. Las sumas correspondientes al error y sus grados de libertad se suelen calcular restando los totales de los otros tres. Así se consigue más rapidez. Este sumando representa la variabilidad interna de los datos, independientemente de la influencia de los factores. Es el verdadero estimador de la varianza, y hay quien plante el ANOVA sólo para conseguir este estimador.

En el ejemplo la mejor estimación de la varianza sería 2,9.

Total: SCT=421,62. Es la suma de los factores, la interacción y el error. Su utilidad reside en facilitar los cálculos y comprobar que las sumas cuadran bien.

Todos los cuadrados medios estiman la varianza de la población, aunque el mejor estimador sea 2,9. Si aplicamos el contraste F a la comparación de estimadores, los sesgos significativos que encontremos se deberán a influencias de los factores.

En el ejemplo ha resultado significativo el factor edad.

Análisis de la regresión

Modelo y supuestos

Las técnicas de descomposición en sumas de cuadrados propias del ANOVA también se pueden aplicar a la regresión entre dos variables. El modelo teórico es el de suponer que entre dos variables X e Y existe una relación lineal de la forma:

$$\mathbf{Y_{ij} = \alpha + \beta X_i + e_{ij}}$$

En esta fórmula supondremos lo siguiente:

X es cuantitativa y presenta valores fijos, como los niveles en el modelo ANOVA. Estos valores dividen a los de Y en distintos subconjuntos. Se supone que los valores de Y en ellos son independientes entre sí (covarianza cero)

Y presenta valores aleatorios dependientes de X según la relación lineal $\alpha + \beta X$ a cuyo valor se añade un sumando aleatorio e_{ij}. Se supone que e_{ij} se distribuye normalmente y que las medias de Y en los distintos conjuntos dependen de las medias de X según la misma relación lineal.

Los valores de la varianza en los distintos subconjuntos son iguales (homocedasticidad)

Lo anterior es un breve resumen de los supuestos. En manuales de Inferencia Estadística puedes estudiarlos con más amplitud.

En los temas 5 y 7 estudiamos los estimadores de α y β y los valores pronosticados $Y' = \alpha + \beta X$. Aquí nos interesarán más bien las descomposiciones en sumas de cuadrados y las técnicas de ANOVA.

Hipótesis nula: $\beta=0$ **Hipótesis alternativa: $\beta\neq0$ (o $\beta<0$ o $\beta>0$)**

La anulación de β equivale a que todas las medias de subgrupos sean iguales, porque la recta de regresión sería horizontal, luego esta hipótesis nula coincide con la del ANOVA de igualdad de medias. Por eso nos vale esta técnica también para la regresión. Explicaremos cómo:

Suma de cuadrados total: $$SCT = \sum_i \sum_j (Y_{ij} - \bar{Y})^2$$

Tiene la misma expresión que en el ANOVA, y sus grados de libertad serán N-1, porque se ha estimado un valor, que es la media de Y.

Suma de cuadrados explicada: $$SCT = \sum_i \sum_j (Y'_{ij} - \bar{Y})^2$$

Si representamos los pronósticos del modelo de regresión como Y', se dará entonces la identidad **$Y'=\alpha + \beta X$**. La diferencias de Y' respecto a la media general representarán a la variabilidad explicada por el modelo. Sólo tiene un grado de libertad, pues todo depende del valor de β.

Suma de cuadrados no explicada (o de error): $$SCT = \sum_i \sum_j (Y_{ij} - Y'_{ij})^2$$

En ella se suman las diferencias entre los valores reales de Y y los pronosticados Y'. Es decir, se suman los cuadrados de $\mathbf{e_{ij}}$. Representa, pues, la suma de errores, y de ahí su nombre. Le quedarán N-2 grados de libertad, por lo que el estimador de la varianza de la población será el cociente de esa suma entre N-2.

Estas tres sumas se pueden estructurar de forma similar al caso de ANOVA con un factor. Lo veremos con un ejemplo:

Se ha sometido a unos sujetos a unas horas de entrenamiento para una prueba en la que el número de aciertos depende en gran parte del manejo de un mando de juegos para ordenador de nuevo diseño. En la siguiente tabla se recogen los distintos niveles de tiempo de entrenamiento y las puntuaciones obtenidas en un determinado juego.

Tiempo en minutos	Resultados en puntos de 0 a 10
10	3 3 4 5 4 6 4
15	4 5 4 6 5 7 8
20	4 6 6 5 7 8 7 5 6
25	5 9 9 8 6 7 10 4 7
30	4 8 8 9 6 8 9 10 10

¿Se puede considerar que estos datos siguen un modelo de tipo lineal? ¿Cuál sería su ecuación? ¿Qué varianza presenta la población?

Aplicamos el ANOVA y nos queda:

Fuente variación	SC	G.L.	CM	F
Regresión	**69,39**	**1**	**69,39**	**28,34**
Error	**95,49**	**39**	**2,45**	
TOTAL	**164,88**	**40**	**4,12**	

P-valor de F	**0,000**	
Fcrítica al 90%	**2,84**	**Significativa**
Fcrítica al 95%	**4,09**	**Significativa**
Fcrítica al 99%	**7,33**	**Significativa**

La F=28,34 es claramente significativa, luego existe influencia de tipo lineal.

La estimación de la varianza de la población es el cuadrado medio de error, es decir, 4,12

La ecuación de la recta de regresión la obtendríamos por los métodos tradicionales y resultaría ser Y' = 2,43+0,187X

Prueba del análisis de regresión

Podemos combinar el análisis de regresión con el de varianza para probar simultáneamente la anulación de la pendiente y la hipótesis de linealidad. El esquema de cálculo sería el mismo pero añadiendo también las sumas de cuadrados INTER e INTRA.

Sólo daremos el esquema al que daría lugar el ejemplo, pues se explica por sí solo:

Análisis de la regresión
comparado con el ANOVA

Fuente variación	SC	G.L.	CM	F
INTER	**70,75**	**4**	**17,69**	**6,76**
Regresión	**69,39**	**1**	**69,39**	**26,54**
Desviación regresión	**1,36**	**3**	**0,45**	**0,17**
INTRA	**94,13**	**36**	**2,61**	
TOTAL	**164,88**	**40**	**4,12**	

		Fcrítica al 95%	
P-valor de F-INTER	**0,000**	**2,63**	**Significativa**
P-valor de F-REGR.	**0,000**	**4,11**	**Significativa**
P-valor de F-DESV.	**0,914**	**2,87**	

En él aparece somo significativa la F-INTER, luego podemos afirmar que hay efecto de los niveles. También es significativa la F-REGR. y no lo es la desviación, por lo que nos reafirmamos en que el efecto de los niveles es de tipo lineal con pendiente no nula.