UNIDAD DIDÁCTICA 5-3

INTRODUCCIÓN AL ANÁLISIS DE LA VARIANZA

1. Introducción

En esta parte del Unidad Didáctica 5 se estudia el Análisis de la Varianza (ANOVA), que constituye, sin lugar a dudas, una de las técnicas más valiosas de la Inferencia Estadística.

Desarrollado hacia 1930 por R.A. FISHER, cuando trabajaba en una estación de investigación agraria en Inglaterra, el ANOVA es la técnica básica para el estudio de observaciones que dependen de varios factores, siendo la herramienta fundamental en el análisis de los modelos de Regresión Lineal y de Diseño de Experimentos.

2. Ejemplo intuitivo

La **idea básica** del **ANOVA** consiste en **descomponer la variabilidad total observada en unos datos en una serie de términos, asociados a los efectos de cada factor estudiado y a sus posibles interacciones, más una parte residual con la que después se compararán las primeras para investigar su posible significación estadística**.

Con el fin de precisar estas ideas, vamos a ver un sencillo ejemplo intuitivo.

Ejemplo: Se desea estudiar los efectos que la variedad cultivada y la dosis de abonado tienen sobre el rendimiento de un cultivo de trigo (expresado en Qms/Ha). Se van a comparar dos variedades (A y B) y tres dosis de abonado (10, 20 y 30 Kg de N por Ha). Se dispone de 12 parcelas similares, y se decide plantar 2 parcelas con cada una de las 6 combinaciones posibles de variedad y dosis.

- El rendimiento del trigo es la variable respuesta. Se trata de una variable aleatoria.
- Se desea estudiar el efecto sobre esta variables de dos factores:
  - la variedad, que es un factor cualitativo para el que se analizan dos variantes (variedad A y variedad B)
  - la dosis de abonado, que es una factor cuantitativo para el que se analizan 3 niveles (10, 20 y 30)
- En total hay 6 tratamientos posibles , obtenidos al combinar las 2 variedades con las 3 dosis. A cada tratamiento le corresponde una población en la que está definida la variable respuesta. Así, la primera población estaría constituida por todas las parcelas hipotéticas que podrían plantarse con la variedad A y abonarse con 10 kgs N/Ha)

- Las dos parcelas asignadas a cada tratamiento constituyen una muestra de la población correspondiente. El conjunto de los resultados de las 6 muestras constituyen los datos disponibles para el estudio.
- Las cuestiones que se desea investigar son:

A) En promedio para las 3 dosis estudiadas ¿hay diferencia entre los rendimiento medios de las dos variedades? (¿Existe un efecto del factor Variedad sobre la media de la respuesta?)

B) En promedio para las dos variedades estudiadas, ¿varía el rendimiento medio al variar la dosis de abonado? (¿Existe un efecto del factor Dosis sobre la media de la respuesta?)

C) El efecto del abonado sobre el rendimiento ¿es diferente en la variedad B que en la variedad A? O, lo que es equivalente, ¿la diferencia de rendimiento medios entre las dos variedades se modifica al variar la dosis de abonado? (¿Existe una interacción entre los efectos de los dos factores Variedad y Dosis?)

Veamos unos posibles resultados (rendimientos por parcela en Qm/Ha) en algunos casos hipotéticos extremos, con el fin de introducir la idea de la descomposición de la variabilidad total en términos asociados a los diferentes efectos posibles:

	Dosis Abonado		
	10	**20**	**30**
Var: A	20 20	20 20	20 20
Var: B	20 20	20 20	20 20

Caso 1º: los 12 valores x_{ijk} han resultado idénticos. ¡No hay variabilidad en los datos! La variabilidad total se mide en el ANOVA por la Suma de Cuadrados Total, que es la suma de los cuadrados de las desviaciones de los datos respecto a la media de todos ellos). En este caso 1º se tendrá que $SC_{total} = \sum_{i,j,k} \left(x_{ijk} - \bar{x} \right)^2$ es cero

Consecuencia: ni la variedad ni la dosis de abonado influyen en el rendimiento

	Dosis Abonado		
	10	**20**	**30**
Var: A	20 20	20 20	20 20
Var: B	30 30	30 30	30 30

Caso 2º: ¡Sí que hay variabilidad en los datos! $SC_{total} = \sum_{i,j,k} \left(x_{ijk} - \bar{x} \right)^2 = 300$ (la fórmula cómoda para calcular la SC_{total} se ve más adelante)

Cuando se analiza esta variabilidad (de ahí el nombre de Análisis de la Varianza) se constata que se debe exclusivamente al efecto de la variedad sobre la media (la B da valores sistemáticamente más altos que la A), no habiendo ningún efecto de la dosis.En la tabla resumen del ANOVA se calcularía, mediante unas fórmulas que se ven más adelante[1], una

[1] Estas fórmulas no se usarán en este curso, pero sí que hay que saber rellenar datos faltantes en una tabla ANOVA.

suma de cuadrados debida al efecto de la variedad (SC_{var}) y una suma de cuadrados debida al efecto de la dosis (SC_{dosis}). Aplicando estas formulas a los datos de este 2º caso se obtendría:

SC_{total}=300 SC_{var}= 300 SC_{dosis}=0

	Dosis Abonado		
	10	**20**	**30**
Var: A	20 20	25 25	30 30
Var: B	30 30	35 35	40 40

Caso 3º : la SC_{total} calculada por la fórmula que se ve más adelante, es ahora:

$$SC_{total} = \sum_{i,j,k} \left(x_{ijk} - \bar{x} \right)^2 = 500$$

Cuando se analiza esta variabilidad se constata que parte de ella se debe al efecto de la variedad sobre la media (La B da valores 10 unidades más altos que la A), pero que también hay una parte debido al efecto de la dosis (los valores aumentan 5 unidades al pasar de 10 a 20 y otras 5 unidades al pasar de 20 a 30). Por otra parte se constata que no hay interacción entre ambos factores, porque la diferencia entre A y B es la misma sea cual sea la dosis, y el efecto de aumentar la dosis es el mismo en ambas variedades.

Calculando, mediante las fórmulas que se verán más adelante, las Sumas de Cuadrados asociadas a cada efecto se obtendría:

SC_{total}=500 SC_{var}= 300 SC_{dosis}=200 $SC_{interaccion}$=0

	Dosis Abonado		
	10	**20**	**30**
Var: A	20 20	25 25	30 30
Var: B	30 30	40 40	50 50

Caso 4º: la variabilidad total, calculada por la fórmula que se ve más adelante, es ahora

$$SC_{total} = \sum_{i,j,k} \left(x_{ijk} - \bar{x} \right)^2 = 1175$$

Cuando se analiza esta variabilidad se constata que una parte de ella se debe al efecto de la variedad sobre la media (La B da valores más altos que la A), y parte se debe al efecto de la dosis (los valores aumentan al aumentar la dosis). Pero por otra parte se constata que hay una interacción entre ambos factores: la diferencia entre B y A es mayor a dosis altas que bajas, y el efecto de aumentar la dosis es más marcado en la variedad B que en la A.

Calculando, mediante las fórmulas que se verán más adelante, las Sumas de Cuadrados asociadas a cada efecto se obtendría:

SC_{total}=1175 SC_{var}= 675 SC_{dosis}=450 $SC_{interaccion}$=50

Sin embargo, todos los casos vistos hasta ahora no son realistas, porque suponen que en las 6 poblaciones estudiadas no hay variabilidad (dado que los dos valores muestreados en cada tratamiento han resultado idénticos). En la realidad las diferencias existentes entre las parcelas dentro de cada población (distinta fertilidad, desigual ataque de plagas, etc...) harán que los resultados obtenidos, pese a tener la misma variedad y la misma dosis de abonado, no sean idénticos.

	Dosis Abonado		
	10	20	30
Var: A	19 21	26 24	28 32
Var: B	30 31	39 43	49 52

Caso 5º: Este caso es el único realista. Los efectos de los factores vienen parcialmente enmascarados por la variabilidad residual originada por factores no controlados (Las parejas de parcelas con idéntica variedad y abonado no son exactamente iguales y, por tanto, no dan exactamente los mismos rendimientos).

La variabilidad total se debe ahora, no sólo al efecto de la variedad, la dosis y su interacción, sino además a una variabilidad residual asociada a todos los restantes factores no controlados que influyen sobre los rendimientos (diferencias entre parcelas de una misma población)

Calculando, mediante las fórmulas que se verán más adelante, las Sumas de Cuadrados asociadas a cada efecto se obtendría:

SC_{total}=1261.67 SC_{var}= 736.33 SC_{dosis}=450.17 $SC_{interac}$=50.17 SC_{resid}=25.0

Construcción del Cuadro Resumen del Análisis de la Varianza

Paralelamente a esta descomposición de la SC_{total} en sus componentes, se realiza una descomposición de los "grados de libertad" totales, que son siempre el número de datos menos 1 (12-1=11 en el ejemplo), en los grados de libertad asociados a cada término. Los grados de libertad asociados al efecto de un factor son siempre el número de variantes del factor menos 1 (gl_{var}=2-1=1 y gl_{dosis}=3-1=2), mientras que los de una interacción son el producto de los grados de libertad de los factores correspondientes (gl_{interc}=1x2=2), quedando como grados de libertad residuales los restantes (gl_{resid}=11-1-2-2=6)

La comparación de la "varianza" asociada a cada efecto con la varianza residual permite estudiar si dicho efecto es o no significativo. Dichas varianzas se estiman dividiendo cada Suma de Cuadrados por sus correspondientes grados de libertad, obteniéndose unos estadísticos a los que se denomina Cuadrados Medios (El CM_{total}, que no es más que la varianza de los datos, no se acostumbra a calcular).

El $CM_{residual}$ es una estimación de la σ^2 existente en las poblaciones muestreadas, asumiendo que dichas poblaciones tienen todas la misma σ^2 (o del promedio de dichas varianzas en el caso de que difieran de unas poblaciones a otras).

El CM asociado a cada efecto es también una estimación (independiente de la anterior) de dicha σ^2 si dicho efecto no existe en la población, pero tiende a ser mayor que σ^2 en el caso de que exista un efecto real poblacional.

Para ver si el CM de un efecto es significativamente mayor que el $CM_{residual}$, lo que implicaría la existencia de un efecto real a nivel poblacional, se comprueba si el cociente $CM_{efecto}/CM_{residual}$ (al que se denomina F-ratio) es demasiado elevado para ser una F de Fisher con los grados de libertad correspondientes, calculándose para ello el valor *p-value* asociado (que no es más que la probabilidad de que una F, con los grados de libertad correspondiente, sea mayor que el valor obtenido para la F_{ratio}). Contra menor sea este *p-vale*, más fuerte será la evidencia respecto a al existencia poblacional del efecto correspondiente.

Los resultados se disponen en una tabla resumen del ANOVA, tal como se recoge a continuación para los datos del Caso 5º:

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Variedad	736.333	1	736.333	176.72	0.0000
Dosis	450.167	2	225.083	54.02	0.0001
Variedad*Dosis	50.167	2	25.0833	6.02	0.0368
RESIDUAL	25.0	6	4.16667		
TOTAL	1261.67	11			

Del análisis de la tabla se desprende que los efectos simples de Variedad y de Dosis son muy significativos estadísticamente (p-values menores que 0.01) y que el efecto de la interacción Variedad*Dosis es también significativo estadísticamente (p-value inferior a 0.05 aunque superior a 0.01).

3. Terminología

3.1. Respuesta

Es aquella variable sobre la que se quiere estudiar el posible efecto de los factores estudiados. Es siempre una variable aleatoria definida sobre las poblaciones correspondientes a los diferentes tratamientos (ver aptdo. 3.4). En general se estudian los posibles efectos de los factores sobre la media de la respuesta, pero también es posible utilizar el ANOVA para investigar efectos sobre la varianza de la respuesta.

Ejemplos de variables respuesta son: *througput* de un sistema informático, retardo de los mensajes de un multicomputador, tiempo de inserción de objetos mediante una determinada estructura de datos, consumo de energía de un proceso, etc.

3.2. Factores

Son los diferentes parámetros del proceso estudiado, cuyos posibles efectos sobre la variable respuesta se desea investigar.

Pueden ser de naturaleza cualitativa (cuando las diferentes alternativas para el factor son cualitativamente diferentes) o cuantitativa (cuando las diferentes alternativas para el factor corresponden a diferentes niveles o cantidades de algo).

Ejemplos factores cualitativos: temperatura, tráfico de una red, número de canales virtuales, tamaño de memoria, etc.

Ejemplos factores cuantitativos: proveedor, variedad, topología de una red, tipo de procesador, número de buffers, etc.

3.3. Variantes y niveles

Variantes: son las diferentes alternativas que se plantean en el estudio de un factor cualitativo. Por ejemplo, para el factor tipo de procesador, las variantes podrían ser: *Intel Xeon X5670*, *Intel Xeon X5680* e *Intel Xeon W3680*

Niveles: son las diferentes alternativas que se plantean en el estudio de un factor cuantitativo. Por ejemplo, para el factor temperatura los niveles podrían ser 35ºC, 45ºC, 55ºC y 65ºC.

3.4. Tratamientos

Cada combinación que se obtiene seleccionando una variante (o nivel) de cada uno de los factores estudiados constituye un tratamiento. A cada tratamiento le corresponde una población sobre la que se distribuye la variable respuesta. Por ejemplo, en un experimento en el que se quiere estudiar el efecto de los factores TEMPERATURA y TIPO CPU, un tratamiento sería procesador *Intel Xeon W3680* a 35ºC.

3.5. Planes factoriales equilibrados

Sea un estudio en el que se van a investigar los efectos de K factores sobre una variable respuesta.

Factor 1: se plantean n_1 niveles o variantes

Factor 2: se plantean n_2 niveles o variantes

…………………………………………………

Factor k se plantean n_k niveles o variantes

El número de posibles tratamientos será = $n_1 \times n_2 \times \ldots \times n_K$

Un Plan Factorial Equilibrado (o Balanceado) es aquél en el que para cada tratamiento posible se dispone de un <u>mismo número</u> n_r de datos

Si $n_r = 1$ el Plan Factorial se denomina **no replicado**.

Las expresiones para el cálculo de las Sumas de Cuadrados (que es el único cálculo laborioso en la construcción del cuadro resumen del ANOVA) son sencillas en el caso de los planes factoriales equilibrados. Los cálculos son generalmente mucho más complicados en el caso de planes factoriales no equilibrados, exigiendo el recurso a software estadístico (como Statgraphics).

4. Concepto de efecto simple y de interacción doble

Cuando se estudia sólo un único factor, el "efecto" del mismo sobre la media de la variable estudiada hace referencia a la existencia de diferencias entre las medias de las poblaciones asociadas a las diferentes variantes del factor.

Cuando se estudian simultáneamente varios factores, aparecen los conceptos de "efectos simples" y de "interacciones", que vamos a precisar sobre el siguiente ejemplo

Ejemplo: en una red de multicomputadores se desea investigar el efecto, sobre los retardos de los mensajes transmitidos en la red, de dos factores:

- Algoritmo de encaminamiento de mensajes utilizado: con dos variantes 1 (algoritmo no adaptativo) y 2 (algoritmo adaptativo)
- Nivel de carga de la red: con tres niveles 1 (bajo), 2 (medio) y 3 (alto)

Efecto simple de un factor:

Se define sobre el promedio de las condiciones estudiadas de los restantes factores.

Así, en nuestro ejemplo, el efecto simple del factor " Algoritmo de encaminamiento" se medirá por la diferencia entre los retardos medios obtenidos con uno u otro algoritmo, para el promedio de los tres niveles de carga estudiados.

De forma análoga, el efecto simple del factor " Nivel de carga de la red " se medirá por las diferencias entre los retardos medios obtenidos con los tres niveles de carga, para el promedio de los dos algoritmos de encaminamiento.

Interacciones dobles

Existirá una interacción doble entre ambos factores, si el efecto de uno de ellos es diferente según la variante considerada del otro factor.

Así, en nuestro ejemplo, existirá interacción entre los dos factores si, por ejemplo, la diferencia de los retardos medios entre los dos algoritmos es muy marcada si el nivel de carga de la red es alto, pero es pequeña o inexistente si el nivel de carga es bajo.

De forma simétrica, existiría interacción entre los dos factores si, por ejemplo, la diferencia de los retardos medios entre un nivel de carga alto y un nivel de carga bajo fuera mucho más marcada trabajando con el algoritmo A que con el algoritmo B.

Hay que tener en cuenta que si el efecto de un primer factor depende de la variante considerada del segundo, también necesariamente el efecto del segundo factor dependerá de la variante considerada del primero. De hecho, ambos fenómenos no son más que las dos caras de una misma moneda: la existencia de interacción entre los dos factores.

Nota: en estudios con más de dos factores, puede también plantearse la existencia de interacciones de orden superior: triples, cuádruples, etcétera... Así, existiría una interacción triple entre tres factores A, B y C si, por ejemplo, hubiera una interacción doble entre A y B cuando C está a nivel bajo, pero no existiera dicha interacción A*B cuando C está a nivel alto. En general las interacciones de orden superior a dos se presentan poco en la práctica, siendo además difíciles de interpretar, por lo que no serán consideradas en este tema.

5. Construcción del cuadro resumen del ANOVA

5.1. Filas y columnas

Como se ha visto en el ejemplo del apartado 2 el cuadro resumen del ANOVA (que se repite a continuación) tiene las siguientes filas y columnas:

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Variedad	736.333	1	736.333	176.72	0.0000
Dosis	450.167	2	225.083	54.02	0.0001
Variedad*Dosis	50.167	2	25.0833	6.02	0.0368
RESIDUAL	25.0	6	4.16667		
TOTAL	1261.67	11			

Filas: una primera fila (que es idéntica en todos los ANOVA) con el encabezamiento de cada columna, una fila para cada efecto simple y para cada interacción doble, una fila "Residual" y una fila "Total".

Columnas: una primera columna con los nombres de las filas, una columna con la Sumas de Cuadrados, otra con los grados de libertad, otra con los Cuadrados Medios, otra con los F_{ratio} y otra con los p-value

5.2. Cálculo de las Sumas de Cuadrados

El único aspecto laborioso para obtener el cuadro resumen es el cálculo de las Sumas de Cuadrados. Las fórmulas para obtener dichas sumas de cuadrados, en el caso de planes factoriales equilibrados, se exponen a continuación, aunque no se utilizarán en este curso.

Fórmulas generales

Sea un plan factorial equilibrado con dos factores A (con I variantes) y B (con J variantes) y con N replicaciones en cada uno de los IxJ tratamientos posibles. Sea x_{ijn} una observación genérica (donde i = 1 ... I, j = 1 ... J y n = 1, ... N)

Se definen los siguientes **totales**:

$T_{ij} = \sum_n x_{ijn}$ $\quad T_{i0} = \sum_j T_{ij}$ $\quad T_{0j} = \sum_i T_{ij}$ $\quad TG = \sum_{ijn} x_{ijn}$ y el **sustraendo general** $SG = TG^2/(IJN)$

Las expresiones de las **sumas de cuadrados** y sus respectivos **grados de libertad** son:

$SC_{Total} = \sum_{ijn} x_{ijn}^2 - SG$ que tiene IJN - 1 grados de libertad

$SC_{factorA} = \frac{\sum_i T_{i0}^2}{JN} - SG$ que tiene I-1 grados de libertad

$SC_{factorB} = \frac{\sum_i T_{0j}^2}{IN} - SG$ que tiene J-1 grados de libertad

$SC_{A*B} = \frac{\sum_{ij} T_{ij}^2}{N} - SG - SC_{factorA} - SC_{factorB}$ que tiene (I-1)(J-1) grados de libertad

La $SC_{residual}$ y sus grados de libertad se obtienen por diferencia:

$SC_{residual} = SC_{Total} - SC_{factorA} - SC_{factorB} - SC_{A*B}$ $\quad gl_{residual} = IJ(N-1)$

Ejemplo

Se realizó un estudio en una acería, para analizar la influencia sobre el alargamiento máximo hasta la rotura (variable ε_{max} medida en porcentaje) en barras corrugadas de acero, de la calidad del acero (dos calidades: B400SD y B500SD) y del diámetro de la barra (3 diámetros: 8 mm, 16 mm y 24 mm).

Para cada calidad y diámetro se seleccionaron aleatoriamente 5 barras, cada una de una colada diferente, determinándose en cada una el valor de ε_{max} mediante un ensayo de tracción-deformación. Los resultados obtenidos se recogen en la siguiente tabla; se han incluido también en la misma, en negrita, los totales de los 5 datos de cada casilla, de las filas, de las columnas y el total general.

Valores del alargamiento a la rotura (ε_{max} en %)
en barras corrugadas de acero

	8 mm	16 mm	24 mm	
B400SD	15.29	16.91	17.23	
	15.89	16.99	17.81	
	16.02	17.27	17.74	
	16.56	16.85	18.02	
	15.46	16.35	18.37	**252.76**
	79.22	**84.37**	**89.17**	
B500SD	12.07	12.92	13.30	
	12.42	13.01	12.82	
	12.73	12.21	12.49	
	13.02	13.49	13.55	
	12.05	14.01	14.53	**194.62**
	62.29	**65.64**	**66.69**	
	141.51	**150.01**	**155.86**	**447.38**

Los cálculos de las sumas de cuadrado y grados de libertad serían los siguientes:

$$SG = \frac{TG^2}{2x3x5} = \frac{447.38^2}{30} = 6671.63$$

SC_{total} = $15.29^2 + 15.89^2 + ... + 13.55^2 + 14.53^2$ - SG = 131.807   con 30-1=29 gl

$$SC_{calidad} = \frac{252.76^2 + 194.62^2}{15} - SG = 112.675 \qquad \text{con } 2\text{-}1 = 1 \text{ gl}$$

$$SC_{diamtero} = \frac{141.51^2 + 150.01^2 + 155.86^2}{10} - SG = 10.413 \qquad \text{con } 3\text{-}1 = 2 \text{ gl}$$

$$SC_{cal*diam} = \frac{79.22^2 + 84.37^2 + \ldots + 66.69^2}{5} - SG - SC_{calidad} - SC_{diametro} = 1.604$$

con gl = (2-1)x(3-1) = 2

Finalmente la $SC_{residual}$ y sus grados de libertad se calculan por diferencia:

$SC_{residual} = SC_{total} - SC_{calidad} - SC_{diametro} - SC_{cal*diam} = 7.115$

$gl_{residual} = 29 - 1 - 2 - 2 = 24$

5.3. Obtención del resto de la tabla

Los Cuadrados Medios se obtienen dividiendo cada Suma de Cuadrados por sus respectivos grados de libertad (Nota: generalmente el CM_{Total} no se calcula)

La F_{ratio} para cada efecto se calcula dividiendo su Cuadrado Medio por el $CM_{residual}$

Finalmente el p-value correspondiente a cada efecto se obtendría buscando en una tabla de la F la probabilidad de que una F de Fisher, con los grados de libertad correspondientes, sea mayor que el valor obtenido para el F_{ratio}.

Nota: en general las tablas disponibles no son lo suficientemente detalladas como para permitir hallar exactamente los p-values. En la práctica lo que se hace es ver si el F_{ratio} obtenido es mayor que el valor correspondiente al 5% en la tabla F (en cuyo caso el p-value sería inferior al 5% y el efecto sería "significativo") y si también es mayor que el valor correspondiente al 1% en la tabla F (en cuyo caso el p-value sería inferior al 1% y el efecto sería "muy significativo")

Seguidamente se recoge el cuadro resumen del ANOVA para el ejemplo estudiado:

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:calidad	112.675	1	112.675	380.08	0.0000
B:diametr	10.4132	2	5.20658	17.56	0.0000
INTERACTIONS					
AB	1.6035	2	0.80175	2.70	0.0873
RESIDUAL	7.1148	24	0.29645		
TOTAL (CORRECTED)	131.807	29			

6. Interpretación del cuadro resumen. Test F

Como hemos visto, en un ANOVA se muestrean tantas poblaciones como tratamientos existen. La hipótesis clásica del ANOVA (homogeneidad) asume que todas estas poblaciones (cuyas medias pueden ser diferentes si los factores tienen efectos reales) tienen la misma varianza σ^2. A esta varianza, que se debe al efecto de todos los factores no controlados que pueden influir en los resultados, se le denomina varianza residual del modelo.

Se demuestra que el $CM_{residual}$ que aparece en el cuadro resumen, es una estimación de esta varianza residual del modelo.

¿Y qué pasa con los Cuadrados Medios asociados a los diferentes efectos estudiados (simple o interacciones)?... Eso depende de que el efecto correspondiente exista o no a nivel poblacional:

- Se demuestra que si el efecto es inexistente en las poblaciones estudiadas (o sea, si es cierta la hipótesis nula que postula que el efecto del factor o interacción analizado es nulo), el CM_{efecto} es también una estimación (independiente de la anterior) de la varianza residual.

  En este caso, la $F_{ratio} = CM_{efecto}/CM_{residual}$ es una variable aleatoria que sigue una distribución F de Fisher, con los grados de libertad correspondientes (los grados de libertad del efecto (gl_e) del efecto en el numerador y los residuales (gl_r) en el denominador)

- Sin embargo, si realmente el efecto existe a nivel poblacional (o sea, si no es cierta la hipótesis nula que postula que el efecto estudiado es inexistente), se demuestra que el CM_{efecto} es una variable aleatoria cuya media es mayor que la varianza residual.

  En este caso, la $F_{ratio} = CM_{efecto}/CM_{residual}$ es una variable aleatoria cuya media es mayor que la de una distribución $F_{ge,gr}$ de Fisher.

Para dilucidar entre las dos alternativas anteriores, lo que se hace es ver si el valor obtenido para la F_{ratio} es "demasiado grande" frente a lo que cabría esperar de una variable $F_{ge,gr}$ de Fisher. Hay dos formas equivalentes de ver esto:

1. Si se dispone de una tabla F, se obtienen de la misma los valores F(5%) y F(1%) que son aquellos que tienen, respectivamente, una probabilidad del 5% y del 1% de ser superados por la variable $F_{ge,gr}$.

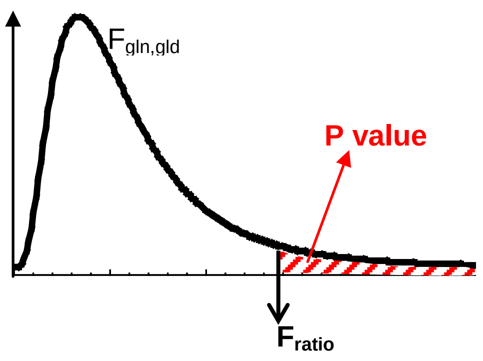
   1.1. Si F_{ratio} > F(1%) se rechaza la hipótesis nulo y se dice que el efecto es muy significativo estadísticamente.

   1.2. Si F(1%) > F_{ratio} > F(5%) se rechaza la hipótesis nula y se dice que el efecto es significativo estadísticamente.

   1.3. Si F_{ratio} < F(5%) se acepta la hipótesis nula y se dice que el efecto no es estadísticamente significativo. A veces, si F(5%) > F_{ratio} > F(10%) se califica el efecto como "dudoso"

IMPORTANTE: el hecho de que un efecto no resulte estadísticamente significativo <u>no quiere decir que se ha demostrado que el efecto es inexistente</u>, sino sólo que la hipótesis de que el efecto es nulo es compatible con los datos disponibles (como también lo serían otras hipótesis como, quizás, que el efecto es "poco importante"

2. Los software estadísticos lo que hacen es calcular, mediante métodos numéricos la probabilidad <u>Prob($F_{ge,gr}$ > F_{ratio}) a la que se denomina "p-value"</u>. De acuerdo con lo que se acaba de exponer:



   2.1. Si p-value <0.01 se rechaza la hipótesis nula y se dice que el efecto es muy significativo estadísticamente.

   2.2. Si 0.05 > p-value > 0.01 se rechaza la hipótesis nula y se dice que el efecto es significativo estadísticamente.

   2.3. Si p-value > 0.05 se acepta la hipótesis nula y se dice que el efecto no es estadísticamente significativo[3]. A veces si 0.10 > p-value > 0.05 se califica el efecto como "dudoso"

En definitiva, un efecto es tanto más significativo cuanto:

- Mayor es el CM_{efecto} en relación al $CM_{resdual}$
- Mayor es, en consecuencia, su F_{ratio}
- Menor es, en consecuencia, su p-value(ver figura anterior)

Así, volviendo al cuadro resumen del ANOVA del ejemplo de los aceros

```
--------------------------------------------------------------------------------
Source                  Sum of Squares     Df     Mean Square     F-Ratio     P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:calidad                     112.675      1         112.675      380.08      0.0000
 B:diametr                     10.4132      2         5.20658       17.56      0.0000

INTERACTIONS
 AB                             1.6035      2         0.80175        2.70      0.0873

RESIDUAL                        7.1148     24         0.29645
--------------------------------------------------------------------------------
TOTAL (CORRECTED)              131.807     29
--------------------------------------------------------------------------------
```

la conclusión sería que el efecto de la calidad y del diámetro sobre el valor medio del ε_{mas} son muy significativos estadísticamente y que la interacción doble entre los efectos de ambos factores es dudosa.

7. Interpretación de factores cualititativos. Intervalos LSD

Cuando el test F resulta significativo para un factor cualitativo con I > 2 variantes, hay que precisar entre cuáles de las variantes del factor existen diferencias significativas entre las medias. En efecto, un valor significativo de la F-ratio sólo indica que no es cierto que la I medias sean iguales, o sea que al menos una de las medias difiere de las restantes, pero no precisa cuáles son las que difieren entre sí

Una forma sencilla de precisar esta cuestión es mediante el establecimiento de intervalos LSD (siglas de "Least Signficative Difference") para la media de cada variante.

Si la media muestral correspondiente a la variante i del factor es $\bar{x}_i$, la desviación típica con la que está estimada dicha media será igual a la raíz cuadrada del CM_{resid} (que como hemos visto es una estimación de la varianza existente en las poblaciones estudiadas), dividida por el número n de datos a partir de los que se ha calculado la media.

El intervalo LSD para dicha media viene dado por la expresión[2]:

$$\bar{x}_i \pm \frac{\sqrt{2}}{2} t^{\alpha}_{glresid} \sqrt{\frac{CM_{resid}}{n}}$$

La diferencia entre la media de dos tratamientos será significativa si los respectivos intervalos LSD no se solapan entre sí.

Nota: el intervalo obtenido, intervalo LSD, no es un intervalo de confianza para las medias correspondientes. Su utilización es sólo la comparación de medias,

[2] En este curso no se pedirá calcular los intervalos LSD "a mano", pero sí su significado e interpretación.

<u>Ejemplo</u>

Una factoría de motores tiene 2 proveedores de los cigüeñales que mecaniza. Un tercer proveedor ofrece sus cigüeñales algo más caros argumentando sus mejores propiedades dinámicas, concretamente que su equilibrado dinámico (número de gramos de material que hay que eliminar hasta conseguir que el centro de gravedad de la pieza coincida con el eje de giro) es menor.

La factoría decide hacer una prueba comparando 10 cigüeñales del nuevo proveedor (código=A) con 10 de cada uno de sus 2 proveedores tradicionales (códigos B y C). Los resultados obtenidos se recogen en la tabla siguiente. (Los datos se adjuntan en el archivo Statgraphics *equidina.sf3*)

PROVEEDOR		
A	B	C
23	35	50
28	36	43
21	29	36
27	40	34
35	43	45
41	49	52
37	51	52
30	28	43
32	50	44
36	52	34

La tabla resumen del Anova (obtenida mediante Statgraphics) es:

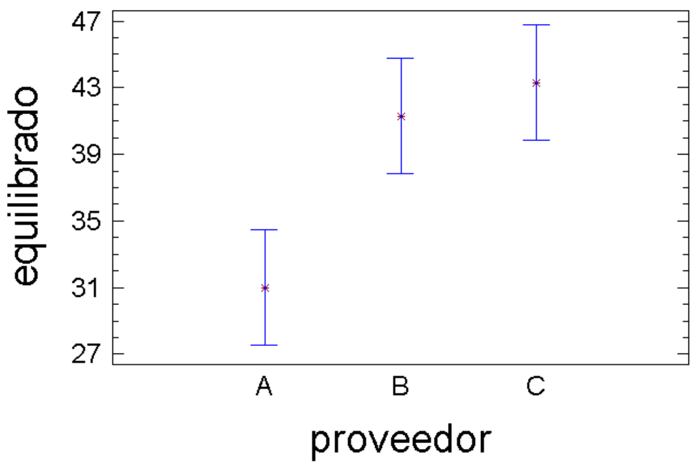
```
--------------------------------------------------------------------------------
Source              Sum of Squares     Df     Mean Square     F-Ratio     P-Value
--------------------------------------------------------------------------------

proveedor                   871.27      2          435.63        7.65      0.0023

RESIDUAL                   1538.20     27          56.970
--------------------------------------------------------------------------------
TOTAL                      2409.47     29
--------------------------------------------------------------------------------
```

que pone de manifiesto (p-value < 0.01) que el efecto del Proveedor es muy significativo.

Los valores medios de los 10 datos de cada proveedor resultan:

media proveedor A = 31.0   media proveedor A = 41.3   media proveedor A = 43.3

Los intervalos LSD para las medias se representan en la siguiente figura:



Puede constatarse que existe una diferencia significativa entre la media del proveedor A y las de los otros dos proveedores, no siendo significativa la diferencia al respecto entre estos dos últimos.

8. Interpretación de los efectos de factores cuantitativos

En el ejemplo de los proveedores, en el que el factor estudiado es de tipo cualitativo, el objetivo del estudio es determinar cuál es el mejor proveedor de los tres estudiados. En el caso de un factor cuantitativo, como por ejemplo el estudio de 4 posibles temperaturas para mejorar el rendimiento de un proceso químico ¿el objetivo perseguido es también seleccionar la mejor temperatura entre las cuatro ensayadas?

<u>En general el estudio del efecto de un factor cuantitativo, no persigue como objetivo la comparación de los niveles concretos ensayados, sino la investigación de la naturaleza de la función de respuesta que relaciona el valor medio de la variable estudiada con el nivel del factor ensayado</u>.

En el caso del ejemplo mencionado, interesa hallar respuestas a preguntas del siguiente tipo:

- ¿Afecta la temperatura (en el margen de valores estudiados) al rendimiento medio del proceso? (En principio el test F del Anova permite contestar esta cuestión)
- En el margen de valores estudiado ¿existe un efecto lineal positivo (o negativo) de la temperatura, es decir una tendencia a crecer (o a decrecer) el rendimiento medio al aumentar la temperatura?
- ¿Existe también un efecto cuadrático de la temperatura? (por ejemplo, a medida que aumenta la temperatura el incremento del rendimiento es cada vez menor, lo que implicaría un efecto cuadrático negativo)

La comparación de los intervalos LSD para las medias de los niveles concretos ensayados del factor no son útiles para responder a estas preguntas, y pueden incluso conducir a conclusiones engañosas.

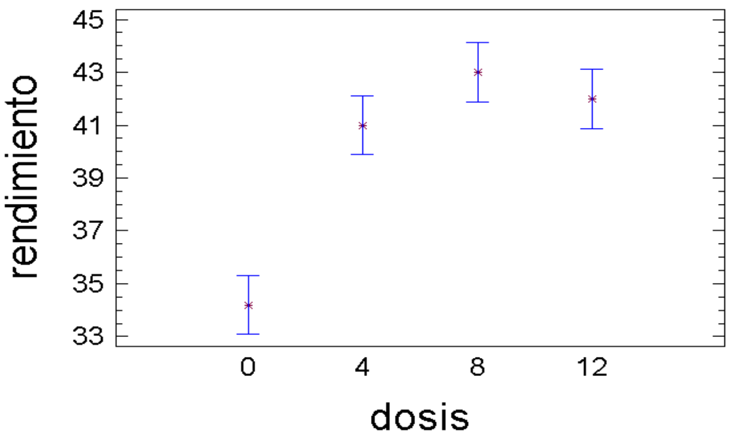
Ejemplo: En la siguiente tabla se recoge el rendimiento en azúcar (Qm/acre) en 20 parcelas de remolacha en las que se utilizaron 4 dosis diferentes de un abono complejo (0, 4, 8 y 12 Qm/acre)

0 Qm/a	4 Qm/a	8 Qm/a	12 Qm/a
37	39	45	42
35	42	41	44
33	41	44	40
34	43	43	43
32	40	42	41

El cuadro resumen del Anova, que se recoge a continuación, indica que el efecto de la dosis de abonado es muy significativo.

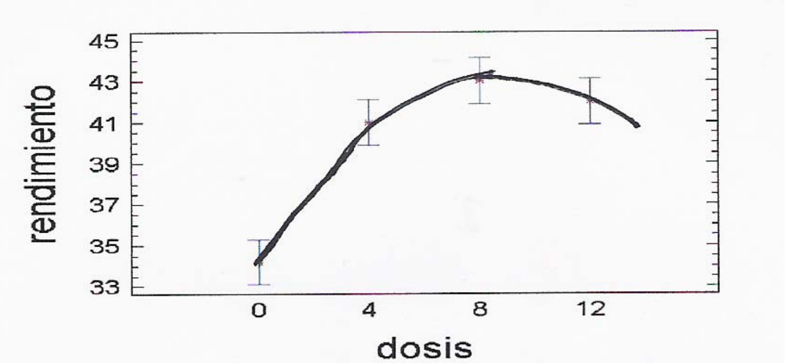
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
dosis	238.15	3	79.3833	28.35	0.0000
RESIDUAL	44.80	16	2.8		
TOTAL	282.95	19			

El gráfico de intervalos LSD para las medias de las 4 dosis de abonado es



Si se interpretase como si de un factor cualitativo se tratase, la conclusión sería que la media de la dosis 0 es significativamente menor que las 3 restantes, pero que no hay diferencias significativas entre las medias de las dosis 4, 8 y 12. Alguien podría concluir que no abonar es peor que abonar, pero que puestos a abonar da lo mismo la dosis que se utilice, conclusión obviamente absurda.

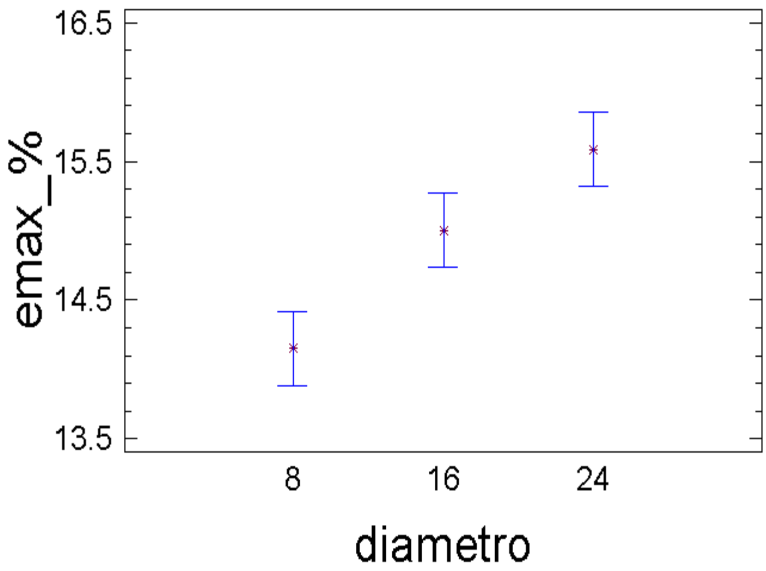
La forma de interpretar el efecto de un factor cuantitativo es analizar un gráfico con las medias de los diferentes niveles (puede utilizarse el propio gráfico de intervalos LSD) para ver la forma que tendría una curva sencilla que pasase, aproximadamente, por dichos valores medios.



Se aprecia en la curva ajustada en la figura que, en el rango de dosis estudiado, el rendimiento medio tiene una tendencia general a crecer al hacerlo la dosis (efecto lineal positivo) pero que este crecimiento es cada vez más lento llegando incluso a decrecer para dosis muy altas (curvatura negativa). Se aprecia en la figura que el rendimiento máximo se obtendría para una dosis cercana a 8.

Nota: este ajuste puede precisarse más ajustando un modelo de regresión como los que se estudiarán en la última parte de la asignatura.

Como otro ejemplo se recoge a continuación el gráfico de intervalos LSD para las medias de los 3 diámetros estudiados en el ejemplo de los aceros

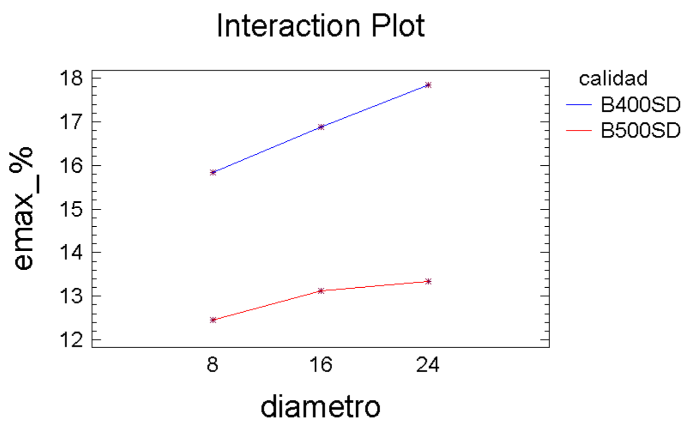


Se aprecia en dicho gráfico que el diámetro tiene un efecto claramente lineal y positivo sobre el valor medio de los ε_{max}.

9. INTERPRETACIÓN DE LAS INTERACCIONES DOBLES

La forma más sencilla de interpretar las interacciones dobles es utilizando un gráfico en el que se representen los valores medios obtenidos para las variantes (o niveles) de un factor, diferenciados según las variantes (o niveles) del otro factor. Estos gráficos los proporciona Statgraphics, teniendo la opción de elegir cuál de los dos factores se pone en el eje de abscisas.

Como ejemplo se recoge a continuación el gráfico de la interacción calidad*diámetro del ejemplo de los aceros (dicha interacción resultó dudosa en la tabla del Anova)



Se precia en el gráfico que el efecto (lineal) del diámetro sobre los ε_{max} es ligeramente más marcado en los aceros B400SD que en los B500SD

10. Residuos: concepto y análisis

En cualquier análisis estadístico real tiene una gran importancia práctica en el análisis de los residuos de los datos.

El residuo de un dato puede definirse como: la diferencia entre el valor realmente obtenido (el dato) y el valor medio predecible para el tratamiento correspondiente a esa observación a partir del conjunto de los datos disponibles:

residuo = dato - (predicción media para el tratamiento)

El residuo refleja cómo han afectado a la observación en cuestión el conjunto de factores residuales, es decir los no contemplados explícitamente en el estudio.

Todos los software estadísticos, y concretamente Statgraphics, proporcionan como complemente de muchos análisis estadísticos (y en particular de los Anova) los valores de los residuos de cada observación y diferentes gráficos para analizarlos.

Si se cumplen todas las hipótesis básicas del Anova los residuos deben fluctuar aproximadamente como los valores de una variable normal de media cero y varianza igual a la σ^2 residual. Un residuo anormalmente elevado en una observación indica que algo "raro" ha pasado en la misma que debe ser estudiado y, en su caso corregido.

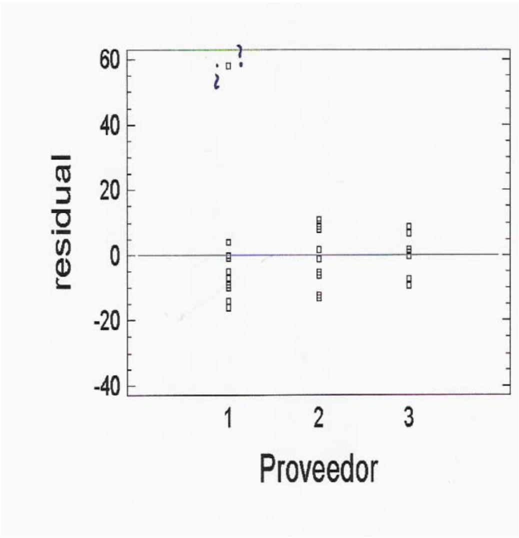
<u>Ejemplo</u>: Como ejemplo de las graves consecuencias que puede generar un valor anómalo y de la utilidad del análisis de los residuos para detectarlo vamos a suponer que al introducir los datos del ejemplo del equilibrado dinámico de los cigüeñales el 5º dato del proveedor A se ha metido como "95" al confundirse el "3" del valor real ("35") con un 9.

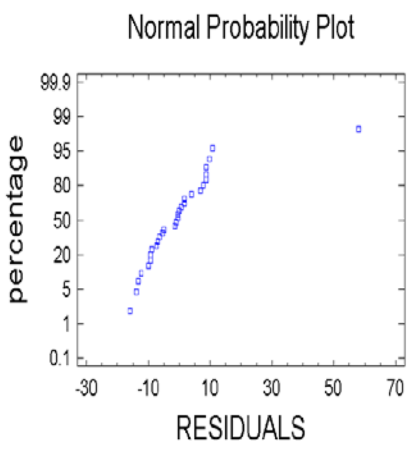
El cuadro resumen del Anova es ahora

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:proveedor	207.267	2	103.633	0.53	0.5934
RESIDUAL	5258.2	27	194.748		
TOTAL (CORRECTED)	5465.47	29			

Comparando los resultados con los obtenidos con el dato correcto se aprecian grandes diferencias: la $SC_{proveedor}$ se ha reducido a la cuarta parte, la $SC_{residual}$ se ha más que triplicado y la F_{ratio} ha pasado de 7.65 (muy significativa) a 0.53 (no significativa)

Al hacer un gráfico de los residuos en función del proveedor se detecta con mucha claridad la presencia de un residuo anómalo, que también se detectaría en una representación en papel probabilístico normal de los residuos.





Ejercicios resueltos

R. Romero y L.R. Zúnica "Métodos Estadísticos en Ingeniería" SPUPV 637

Ver boletín correspondiente en PoliformaT (EST GII: Recursos / 04 | Ejercicios)

Para saber más

- **One-way ANOVA. Monash University**

  http://www.buseco.monash.edu.au/mkt/resources/applets/one-way-anova.html

Mediante este applet se puede realizar el ANOVA de 1 factor con 3 o 4 niveles o variantes. Hay tres paneles:

- ✓ Panel de datos que muestra las poblaciones y los datos de la muestra.
- ✓ La tabla resumen del ANOVA.
- ✓ La f(x) de la distribución F que muestra el valor F calculado, el valor-p y el valor crítico.

Fuentes

Métodos Estadísticos en Ingeniería (Romero Villafranca, Rafael) etsinf

DEIOAC VALENCIA