ANEXO I.

ANÁLISIS DE LA VARIANZA.

El análisis de la varianza (o Anova: Analysis of variance) es un método para comparar dos o más medias. Cuando se quiere comparar más de dos medias es incorrecto utilizar repetidamente el contraste basado en la *t de Student* por dos motivos:

En primer lugar, y como se realizarían simultánea e independientemente varios contrastes de hipótesis, la probabilidad de encontrar alguno significativo por azar aumentaría. En cada contraste se rechaza la H_0 si la *t* supera el nivel crítico, para lo que, en la hipótesis nula, hay una probabilidad α. Si se realizan *m* contrastes independientes, la probabilidad de que, en la hipótesis nula, ningún estadístico supere el valor crítico es $(1 - \alpha)^m$, por lo tanto, la probabilidad de que alguno lo supere es $1 - (1 - \alpha)^m$, que para valores de α próximos a 0 es aproximadamente igual a α*m. Una primera solución, denominada *método de Bonferroni*, consiste en bajar el valor de α, usando en su lugar α / m, aunque resulta un método muy conservador.

Por otro lado, en cada comparación la hipótesis nula es que las dos muestras provienen de la misma población, por lo tanto, cuando se hayan realizado todas las comparaciones, la hipótesis nula es que todas las muestras provienen de la misma población y, sin embargo, para cada comparación, la estimación de la varianza necesaria para el contraste es distinta, pues se ha hecho en base a muestras distintas.

El método que resuelve ambos problemas es el *anova*, aunque es algo más que esto: es un método que permite comparar varias medias en diversas situaciones; muy ligado, por tanto, al diseño de experimentos y, de alguna manera, es la base del análisis multivariante.

1. Bases del análisis de la varianza.

Supónganse *k* muestras aleatorias independientes, de tamaño *n*, extraídas de una única población normal. A partir de ellas existen dos maneras independientes de estimar la varianza de la población σ^2:

1) Una llamada *varianza dentro de los grupos* (ya que sólo contribuye a ella la varianza dentro de las muestras), o *varianza de error*, o *cuadrados medios del error*, y habitualmente representada por *MSE* (*M*ean *S*quare *E*rror) o *MSW* (*M*ean *S*quare *W*ithin) que se calcula como la media de las *k* varianzas muestrales (cada varianza muestral es un estimador centrado de σ^2 y la media de *k* estimadores centrados es también un estimador centrado y más eficiente que todos ellos). *MSE* es un cociente: al numerador se le llama *suma de cuadrados del error* y se representa por *SSE* y al denominador *grados de libertad* por ser los términos independientes de la suma de cuadrados.

2) Otra llamada *varianza entre grupos* (sólo contribuye a ella la varianza entre las distintas muestras), o *varianza de los tratamientos*, o *cuadrados medios de los tratamientos* y representada por *MSA* o *MSB* (*M*ean *S*quare *B*etween). Se calcula a partir de la varianza de las medias muestrales y es también un cociente; al numerador se le llama *suma de cuadrados de los tratamientos* (se le representa por *SSA)* y al denominador (*k*-1) grados de libertad.

MSA y *MSE*, estiman la varianza poblacional en la hipótesis de que las *k* muestras provengan de la misma población. La distribución muestral del cociente de dos estimaciones independientes de la varianza de una población **normal** es una *F* con los grados de libertad correspondientes al numerador y denominador respectivamente, por lo tanto se puede contrastar dicha hipótesis usando esa distribución.

Si en base a este contraste se rechaza la hipótesis de que *MSE* y *MSA* estimen la misma varianza, se puede rechazar la hipótesis de que las *k* medias provengan de una misma población.

Aceptando que las muestras provengan de poblaciones con la misma varianza, este rechazo implica que las medias poblacionales son distintas, de modo que con un único contraste se contrasta la igualdad de *k* medias.

Existe una tercera manera de estimar la varianza de la población, aunque no es independiente de las anteriores. Si se consideran las *kn* observaciones como una única muestra, su varianza muestral también es un estimador centrado de σ^2:

Se suele representar por *MST*, se le denomina *varianza total* o *cuadrados medios totales*, es también un cociente y al numerador se le llama *suma de cuadrados total* y se representa por *SST*, y el denominador (*kn* -1) grados de libertad.

Los resultados de un anova se suelen representar en una tabla como la siguiente:

Fuente de variación	G.L.	*SS*	*MS*	*F*
Entre grupos Tratamientos	k-1	*SSA*	*SSA* /(k-1)	*MSA* /*MSE*
Dentro Error	(n-1)k	*SSE*	*SSE* /k(n-1)	
Total	kn-1	*SST*		

F se usa para realizar el contraste de la hipótesis de medias iguales. La región crítica para dicho contraste es $F > F_{\alpha\,(k-1,(n-1)k)}$

2. Algunas propiedades.

Es fácil ver en la tabla anterior que

$GL_{error} + GL_{trata} = (n - 1)\,k + k - 1 = k + k - 1 = nk - 1 = GL_{total}$

No es tan inmediato, pero las sumas de cuadrados cumplen la misma propiedad, llamada *identidad* o *propiedad aditiva* de la suma de cuadrados:

SST = SSA + SSE

El análisis de la varianza se puede realizar con tamaños muestrales iguales o distintos, sin embargo es recomendable iguales tamaños por dos motivos:

1) La *F* es insensible a pequeñas variaciones en la asunción de igual varianza, si el tamaño es igual.

2) Igual tamaño minimiza la probabilidad de error tipo II.

3. Pruebas para la homocedasticidad.

Para que este contraste de hipótesis, basado en la *F*, lo sea de la igualdad de medias es necesario que todas las muestras provengan de una población con la misma varianza σ^2, de la que *MSE* y *MSA* son estimadores. Por lo tanto es necesario comprobarlo antes de realizar el contraste. Del mismo modo que no se puede usar repetidamente la prueba basada en la *t* para comparar más de dos medias, tampoco se puede usar la prueba basada en la *F* para comparar más de dos varianzas. La prueba más usada para contrastar si varias muestras son homocedásticas (tiene la misma varianza) es la prueba de Bartlett.

La prueba se basa en que, en **la hipótesis nula de igualdad de varianzas** y poblaciones normales, un estadístico calculado a partir de las varianzas muestrales y *MSE* sigue una distribución χ^2_{k-1} .

Otras pruebas para contrastar la homocedasticidad de varias muestras son la de Cochran y la de la *F* del cociente máximo, ambas similares y de cálculo más sencillo pero restringidas al caso de iguales tamaños muestrales. La de Cochran es particularmente útil para detectar si una varianza es mucho mayor que las otras.

En el caso de que las muestras no sean homocedásticas, no se puede, en principio, realizar el análisis de la varianza.

Existen, sin embargo, soluciones alternativas: Sokal y Rohlf describen una prueba aproximada, basada en unas modificaciones de las fórmulas originales.

Hay situaciones en que la heterocedasticidad es debida a falta de normalidad. En estos casos existen transformaciones de los datos que estabilizan la varianza: la raíz cuadrada en el caso de Poisson, el arco seno de la raíz cuadrada de *p* para la binomial, el logaritmo cuando la desviación estándar es proporcional a la media.

En la práctica, si las pruebas de homocedasticidad obligan a rechazar la hipótesis nula, se prueba si con alguna de estas transformaciones los datos son homocedásticos, en cuyo caso se realiza el anova con los datos transformados.

Hay que tener en cuenta que estas pruebas van "*al revés*" de lo habitual. La hipótesis nula es lo que se quiere probar, en consecuencia hay que usarlas con precaución.

4. Modelos de análisis de la varianza.

El anova permite distinguir dos modelos para la hipótesis alternativa:

Modelo I o de *efectos fijos* en el que la H_1 supone que las *k* muestras son muestras de *k* poblaciones distintas y fijas.

Modelo II o de *efectos aleatorios* en el que se supone que las *k* muestras, se han seleccionado aleatoriamente de un conjunto de *m>k* poblaciones.

La manera más sencilla de distinguir entre ambos modelos es pensar que, si se repitiera el estudio un tiempo después, en un modelo I las muestras serían iguales (no los individuos que las forman) es decir corresponderían a la misma situación, mientras que en un modelo II las muestras serían distintas.

Aunque las suposiciones iniciales y los propósitos de ambos modelos son diferentes, los cálculos y las pruebas de significación son los mismos y sólo difieren en la interpretación y en algunas pruebas de hipótesis suplementarias.

4.1 Modelo I o de efectos fijos.

Un valor individual se puede escribir en este modelo como

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i=1, \ldots, k \quad y \quad j=1, \ldots, n$$

μ es la media global, σ_i es la constante del efecto, o efecto fijo, que diferencia a las *k* poblaciones. También se puede escribir:

$$\mu_i = \mu + \alpha_i \quad \varepsilon_{ij}$$

que representa la desviación de la observación *j*-ésima de la muestra *i*-ésima, con respecto a su media. A este término se le suele llamar *error aleatorio* y, teniendo en cuenta las suposiciones iniciales del análisis de la varianza son *k* variables (una para cada muestra), todas con una distribución normal de media o y varianza σ^2.

La hipótesis nula en este análisis es que todas las medias son iguales

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$
H_1: al menos una es diferente

que puede escribirse en términos del modelo como:

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
$H_1:$ al menos dos α_i distintas de 0

Como en H_0 se cumplen las condiciones del apartado anterior se tratará de ver como se modifican las estimaciones de la varianza en H_1.

En H_0 *MSA* y *MSE* son estimadores centrados de σ^2, es decir y usando el superíndice 0 para indicar el valor de las variables en H_0

$E[MSA^0] = \sigma^2$

$E[MSE^0] = \sigma^2$

Se puede ver que *MSE* es igual en la hipótesis nula que en la alternativa. Por lo tanto:

$E[MSE] = E[MSE^0] = \sigma^2$

Sin embargo al valor esperado de *MSA* en la hipótesis alternativa se le añade un término con respecto a su valor en la hipótesis nula

$$E[MSA] = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k} \alpha_i^2$$

Al segundo sumando dividido por *n* se le llama *componente de la varianza añadida por el tratamiento*, ya que tiene forma de varianza, aunque estrictamente no lo sea pues α_i no es una variable aleatoria.

La situación, por lo tanto, es la siguiente: en H_0, *MSA* y *MSE* estiman σ^2; en H_1, *MSE* estima σ^2 pero *MSA* estima $\sigma^2 + n\sigma_\alpha^2$. Contrastar la H_0 es equivalente a contrastar la existencia de la componente añadida o, lo que es lo mismo, que *MSE* y *MSA* estimen, o no, la misma varianza.

El estadístico de contraste es *F=MSA/MSE* que, en la hipótesis nula, se distribuye según una *F* con $k - 1$ y $(n - 1)k$ grados de libertad. En caso de rechazar la H_0, *MSA - MSE* estima $n\sigma_\alpha^2$.

4.2 Modelo II o de efectos aleatorios.

En este modelo se asume que las *k* muestras son muestras aleatorias de *k* situaciones distintas y aleatorias. De modo que un valor aislado Y_{ij} se puede escribir como:

$$Y_{ij} = \mu + A_i + \varepsilon_{ij} \quad i = 1, \ldots k \quad y \quad j = 1, \ldots n$$

donde μ es la media global, ε_{ij} son variables (una para cada muestra) distribuidas normalmente, con media 0 y varianza α^2 (como en el modelo I) y A_i es una variable distribuida normalmente, independiente de las $\square_{ij}$, con media 0 y varianza σ_a^2.

La diferencia con respecto al modelo I es que en lugar de los efectos fijos α_i ahora se consideran efectos aleatorios A_i.

Igual que en el modelo I se encuentra que *MSE* no se modifica en la H_1 y que al valor esperado de *MSA* se le añade el término de *componente añadida* (que aquí es una verdadera varianza ya que A_i es una variable aleatoria):

$$\frac{n}{k-1} \sum_{i=1}^{k} A_i^2 = n\sigma_a^2$$

Para llegar a este resultado se utiliza la asunción de independencia entre A_i y ε_{ij} y es, por tanto, muy importante en el modelo y conviene verificar si es correcta en cada caso.

Por tanto, en H_0 tanto *MSA* como *MSE* estiman σ^2, mientras que en H_1, *MSE* sigue estimando σ^2 y *MSA* estima $\sigma^2 + n\sigma_a^2$. La existencia de esta componente añadida se contrasta con $F = MSA/MSE$ y en caso afirmativo, la varianza de A_i se estima como:

$$\sigma_a^2 = \frac{1}{n}(MSA - MSE)$$

5. Pruebas "*a posteriori*".

En general, en un modelo II el interés del investigador es averiguar si existe componente añadida y en su caso estimarla.

Sin embargo, en un modelo I, lo que tiene interés son las diferencias entre los distintos grupos.

Las pruebas "a posteriori" son un conjunto de pruebas para probar todas las posibles hipótesis del tipo $\mu_i - \mu_j = 0$.

Existen varias, (Duncan, Newman-Keuls, LSD): todas ellas muy parecidas. Usan el rango (diferencia entre medias) de todos los pares de muestras como estadístico y dicho rango debe superar un cierto valor llamado *mínimo rango significativo* para considerar la diferencia significativa.

La principal diferencia con respecto a la *t de Student* radica en que usan *MSE* como estimador de la varianza, es decir un estimador basado en todas las muestras.

Una manera semigráfica habitual de representar los resultados es dibujar una línea que una cada subconjunto de medias adyacentes entre las que no haya diferencias significativas.

6. Análisis de la varianza de dos factores.

Es un diseño de *anova* que permite estudiar simultáneamente los efectos de dos fuentes de variación.

Una observación individual se representa como:

Y_{ijk} $i=1, \ldots, a$ $j=1, \ldots, b$ $k=1, \ldots, n$

El primer subíndice indica el nivel del primer factor, el segundo el nivel del segundo factor y el tercero la observación dentro de la muestra. Los factores pueden ser ambos de efectos fijos (se habla entonces de *modelo I*), de efectos aleatorios *(modelo II)* o uno de efectos fijos y el otro de efectos aleatorios *(modelo mixto)*. El modelo matemático de este análisis es:

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ Modelo I

$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}$ Modelo II

$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \varepsilon_{ijk}$ Modelo mixto

donde μ es la media global, α_i o A_i el efecto del nivel i del 1º factor, β_j o B_j el efecto del nivel j del 2º factor y ε_{ijk} las desviaciones aleatorias alrededor de las medias, que también se asume que están normalmente distribuidas, son independientes y tienen media 0 y varianza σ^2.

A las condiciones de muestreo aleatorio, normalidad e independencia, este modelo añade la de aditividad de los efectos de los factores.

A los términos $(\alpha\beta)_{ij}$, $(AB)_{ij}$, $(\alpha B)_{ij}$, se les denomina *interacción* entre ambos factores y representan el hecho de que el efecto de un determinado nivel de un factor sea diferente para cada nivel del otro factor.

7. Identidad de la suma de cuadrados.

La suma de cuadrados total en un *anova de 2 vías*, es:

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{...})^2$$

(donde para representar las medias se ha usado la convención habitual de poner un punto (.) en el lugar del subíndice con respecto al que se ha sumado) que dividida por sus grados de libertad, *abn* - 1, estima la varianza σ^2 en el supuesto de que las *ab* muestras provengan de una única población. Se puede demostrar que

$$SST = SSA + SSB + SSAB + SSE$$

que es la llamada *identidad de la suma de cuadrados* en un anova de dos factores. Los sucesivos sumandos reciben respectivamente el nombre de suma de cuadrados del 1º factor (tiene *a* -1 grados de libertad y recoge la variabilidad de los datos debida exclusivamente al 1º factor), del 2º factor (con *b* -1 grados de libertad y recoge la variabilidad de los datos debida exclusivamente al 2º factor), de la interacción (con (*a* - 1)(*b* - 1) grados de libertad, recoge la variabilidad debida a la interacción) y del error (con *ab*(*n* - 1) grados de libertad, recoge la variabilidad de los datos alrededor de las medias de cada muestra).

Los resultados de un análisis de la varianza de dos factores se suelen representar en una tabla como la siguiente:

Fuente de variación	***GL***	***SS***	***MS***
1º factor	a - 1	SSA	SSA/(a - 1)
2º factor	b - 1	SSB	SSB/(b - 1)
Interacción	(a - 1)(b - 1)	SSAB	SSAB/[(a - 1)(b - 1)]
Error	ab(n - 1)	SSE	SSE/[ab(n - 1)]
Total	abn - 1	SST	

Los grados de libertad también son aditivos.

En ocasiones se añade una primera línea llamada *de tratamiento* o *de subgrupos* cuyos grados de libertad y suma de cuadrados son las sumas de los del primer, segundo factor y la interacción, que corresponderían a la suma de cuadrados y grados de libertad del tratamiento de un análisis de una vía en que las *ab* muestras se considerarán como muestras de una clasificación única.

Para plantear los contrastes de hipótesis hay que calcular los valores esperados de los distintos cuadrados medios.

8. Contrastes de hipótesis en un análisis de la varianza de dos factores.

Del mismo modo que se hizo en el anova de una vía, para plantear los contrastes de hipótesis habrá que calcular los valores esperados de los distintos cuadrados medios. Los resultados son:

8.1 Modelo I.

MS	*Valor esperado*
MSA	$\sigma^2 + \frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2$
MSB	$\sigma^2 + \frac{na}{b-1}\sum_{j=1}^{b}\beta_j^2$
MSAB	$\sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2$
MSE	σ^2

Por lo tanto, los estadísticos *MSAB*/*MSE*, *MSA*/*MSE* y *MSB*/*MSE* se distribuyen como una *F* con los grados de libertad correspondientes y permiten contrastar, respectivamente, las hipótesis:

i) no existe interacción (*MSAB*/*MSE*)

$$H_0: (\alpha\beta)_{ij} = 0 \quad i=1,\ldots,a \quad j=1,\ldots,b$$

ii) no existe efecto del primer factor, es decir, diferencias entre niveles del primer factor (*MSA*/*MSE*)

$$H_0: \mu_{1\cdot} = \ldots = \mu_{a\cdot}$$

iii) no existe efecto del segundo factor (*MSB*/*MSE*)

$$H_0: \mu_{\cdot 1} = \ldots = \mu_{\cdot b}$$

Si se rechaza la primera hipótesis de no interacción, no tiene sentido contrastar las siguientes. En este caso lo que está indicado es realizar un análisis de una vía entre las *ab* combinaciones de tratamientos para encontrar la mejor combinación de los mismos.

8.2 Modelo II.

MS	*Valor esperado*
MSA	$\sigma^2 + n\sigma^2_{ab} + nb\sigma^2_a$
MSB	$\sigma^2 + n\sigma^2_{ab} + na\sigma^2_b$
MSAB	$\sigma^2 + n\sigma^2_{ab}$
MSE	σ^2

donde σ^2_a, σ^2_b y σ^2_{ab} son, respectivamente las componentes añadidas por el primer factor, por el segundo y por la interacción, que tienen la misma forma que los del modelo I, sin más que cambiar α_i y β_j por A_i y B_j, respectivamente.

La interacción se contrasta, como en el modelo I, con *MSAB*/*MSE*, si se rechaza la hipótesis nula se contrastarían cada uno de los factores con *MSA*/*MSAB* y *MSB*/*MSAB*.

En un modelo II, como no se está interesado en estimar los efectos de los factores sino sólo la existencia de la componente añadida, **sí** tiene sentido contrastar la existencia de la misma para cada factor incluso aunque exista interacción.

Aquí el problema se plantea cuando no se puede rechazar la hipótesis nula y se concluye que no existe interacción: entonces tanto *MSE* como *MSAB* estiman σ^2, entonces ¿cuál se elige para contrastar la componente añadida de los factores?

En principio, parece razonable escoger su media (la media de varios estimadores centrados es también un estimador centrado y más eficiente), sin embargo si se elige *MSAB* se independiza el contraste para los factores de un posible error tipo II en el contraste para la interacción. Hay autores que por ello opinan que es mejor usar *MSAB,* pero otros proponen promediar si se puede asegurar baja la probabilidad para el error tipo II. La media de los cuadrados medios se calcula dividiendo la suma de las sumas de cuadrados por la suma de los grados de libertad.

Ejemplo.

A partir de la siguiente tabla de un *anova* de 2 factores modelo II, realizar los contrastes adecuados.

Fuente de variación	***G.L.***	***SS***	***MS***
1º factor	4	315,8	78,95
2º factor	3	823,5	274,5
Interacción	12	328,9	27,41
Error	100	2308,0	23,08
Total	119	3776,2	

Se empezaría contrastando la existencia de interacción: f = 27,41/23,08 = 1,188 como $F_{0,05(12,100)}$ = 1,849 no se puede, al nivel de significación del 95%, rechazar la hipótesis nula y se concluye que no existe interacción.

Si usamos *MSAB* para contrastar los factores:

1º factor: f = 78,95/27,41 = 2,880 como $F_{0,05(4,12)}$ = 3,26 no se rechaza la hipótesis nula y se concluye la no existencia de componente añadida por este factor.

2º factor: f = 274,5/27,41 = 10,015 como $F_{0,05(3,12)}$ = 3,49 se rechaza la hipótesis nula y se acepta la existencia de componente añadida por este factor.

El resultado del análisis es: no existe componente añadida por la interacción, tampoco por el 1º factor y sí existe componente añadida por el 2º.

La estimación de esta componente es: como a partir de los grados de libertad de la tabla podemos calcular $a = 5$, $b = 4$ y $n = 6$ resulta que la estimación de $n\,a\,\sigma_b^2$ es 274,5 - 27,41 = 247,09; por lo tanto $\sigma_b^2 = 247{,}09 : 30 = 8{,}24$ que representa un 35,7% de componente añadida por el segundo factor.

Si se hubiera optado por promediar, los cuadrados medios promediados son (328,9+2308,0)/(12+100)=23,54 con 112 grados de libertad y hubiera resultado significativo también el 1º factor.

La salida de un paquete estadístico, por ejemplo, el Statgraphics, para un Anova de 2 factores Modelo II

Analysis of Variance for Hum.Relativa - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:altura	2381,75	3	793,917	6,14(1)	0,0851
D:bosque	2145,13	1	2145,13	16,60(1)	0,0267
INTERACTIONS					
AB	387,625	3	129,208	7,30(0)	0,0012
RESIDUAL	425,0	24	17,7083		
TOTAL (CORRECTED)	5339,5	31			

F-ratios are based on the following mean squares:
(0) Residual
(1) AB

8.3 Modelo mixto.

Supóngase el primer factor de efectos fijos y el segundo de efectos aleatorios, lo que no supone ninguna perdida de generalidad, ya que el orden de los factores es arbitrario.

MS	Valor esperado
MSA	$\sigma^2 + n\sigma^2_{ab} + \frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2$
MSB	$\sigma^2 + na\sigma^2_b$
MSAB	$\sigma^2 + n\sigma^2_{ab}$
MSE	σ^2

Se contrastan la interacción y el factor aleatorio con el término de error, si la interacción fuera significativa no tiene sentido contrastar el efecto fijo y si no lo fuera, el efecto fijo se contrasta con el término de interacción o con el promedio de interacción y error.

9. Tamaños muestrales desiguales en un Anova de dos factores.

Aunque los paquetes estadísticos suelen hacer el anova de dos factores, tanto en el caso de tamaños muestrales iguales como desiguales, conviene resaltar que el análisis es bastante más complicado en el caso de tamaños desiguales. La complicación se debe a que con tamaños desiguales hay que ponderar las sumas de cuadrados de los factores con los tamaños muestrales y no resultan ortogonales (su suma no es la suma de cuadrados total) lo que complica no sólo los cálculos sino también los contrastes de hipótesis.

Por esto, cuando se diseña un análisis factorial de la varianza se recomienda diseñarlo con tamaños iguales. Hay ocasiones en que, sin embargo, por la dificultad de obtener los datos o por pérdida de alguno de ellos es inevitable recurrir al análisis con tamaños desiguales. Algunos autores recomiendan, incluso, renunciar a alguno de los datos para conseguir que todas las muestras tengan el mismo tamaño. Evidentemente esta solución es delicada pues podría afectar a la aleatoriedad de las muestras.

10. Casos particulares: Anova de dos factores sin repetición.

En ciertos estudios en que los datos son difíciles de obtener o presentan muy poca variabilidad dentro de cada subgrupo es posible plantearse un anova sin repetición, es decir, en el que en cada muestra sólo hay una observación (n=1). Hay que tener en cuenta que, como era de esperar con este diseño, no se puede calcular *SSE*. El término de interacción recibe el nombre de *residuo* y que, como no se puede calcular *MSE*, no se puede contrastar la hipótesis de existencia de interacción.

Esto último implica también que:

a) en un modelo I, para poder contrastar las hipótesis de existencia de efectos de los factores no debe haber interacción (si hubiera interacción no tenemos término adecuado para realizar el contraste).

b) en un modelo mixto existe el mismo problema para el factor fijo.

Bloques completos aleatorios.

Otro diseño muy frecuente de *anova* es el denominado de *bloques completos aleatorios* diseñado inicialmente para experimentos agrícolas pero actualmente muy extendido en otros campos. Puede considerarse como un caso particular de un *anova* de dos factores sin repetición o como una extensión al caso de k muestras de la comparación de medias de dos muestras emparejadas. Se trata de comparar k muestras emparejadas con respecto a otra variable cuyos efectos se quieren eliminar.

En este diseño a los datos de cada individuo se les denomina *bloque* y los datos se representan en una tabla de doble entrada análoga a la del *anova* de clasificación única en la que las a columnas son los tratamientos y las b filas los bloques, el elemento Y_{ij} de la tabla corresponde al tratamiento i y al bloque j. Las hipótesis que se pueden plantear son:

$H_0: \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{a\cdot}$ (igualdad de medias de tratamientos)

y también, aunque generalmente tiene menos interés:

$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot b}$ (igualdad de medias de bloques)

A pesar del parecido con la clasificación única, el diseño es diferente: allí las columnas eran muestras independientes y aquí no. Realmente es un diseño de dos factores, uno de efectos fijos: los tratamientos, y el otro de efectos aleatorios: los bloques, y sin repetición: para cada bloque y tratamiento sólo hay una muestra.
El modelo aquí es:

$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ donde α_i es el efecto del tratamiento i y B_j el del bloque j. No hay término de interacción ya que, al no poder contrastar su existencia no tiene interés. Al ser un modelo mixto exige la asunción de no existencia de interacción y los contrastes se hacen usando el término *MSE* como divisor.

11 Análisis de la varianza de más de dos factores.

Es una generalización del de dos factores. El procedimiento, por lo tanto, será:

1) encontrar el modelo, teniendo en cuenta si los factores son fijos o aleatorios y todos los términos de interacción.

2) subdividir la suma de cuadrados total en tantos términos ortogonales como tenga el modelo y estudiar los valores esperados de los cuadrados medios para encontrar los estadísticos que permitan realizar los contrastes de hipótesis.

Un modelo de tres factores fijos, por ejemplo, será:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + (\alpha\beta\delta)_{ijk} + \varepsilon_{ijkl}$$

Los tres primeros subíndices para los factores y el cuarto para las repeticiones, nótese que aparecen términos de interacción de segundo y tercer orden, en general en un modelo de *k* factores aparecen términos de interacción de orden 2, 3,... hasta *k* y el número de términos de interacción de orden *n* será el número combinatorio $C_{k,n}$. Este gran número de términos de interacción dificulta el análisis de más de dos factores, ya que son difíciles de interpretar y complican los valores esperados de los cuadrados medios por lo que también resulta difícil encontrar los estadísticos para los contrastes. Por estas razones no se suele emplear este tipo de análisis y cuando interesa estudiar varios factores a la vez se recurre a otros métodos de análisis multivariante.