Análisis de varianza para varias muestras relacionadas

Universidad Pontificia Comillas
Facultad de Ciencias Humanas y Sociales

(Última revisión, 15 de Septiembre de 2011)

índice

1. Introducción

En este modelo de análisis de varianza[1] tenemos los mismos sujetos con observaciones o puntuaciones en la misma variable pero en *condiciones distintas* o en la misma condición pero en *tiempos distintos*[2]. Como se trata de los *mismos sujetos* tenemos *muestras relacionadas* o *emparejadas*.

También puede tratarse de sujetos físicamente distintos pero *igualados* en variables relevantes (variables que controlamos con esta igualación); en este caso se trata también de muestras relacionadas. En la presentación de los datos las *filas* son los sujetos, y las *columnas* son las condiciones.

Una ventaja de utilizar a los mismos sujetos en tratamientos experimentales es que necesitaremos menos sujetos que si se trata de muestras independientes: si queremos comprobar la eficacia relativa de tres actividades distintas de aprendizaje con sujetos distintos, y estimamos en 10 el número mínimo de sujetos, necesitaremos 30 sujetos. Si utilizamos a los mismos sujetos en las tres actividades, nos bastarán 10 sujetos, y además no tenemos que preocuparnos de que los tres grupos sean equivalentes pues se trata siempre de los mismos sujetos[3].

Se trata de comprobar en qué medida la variabilidad total se debe a diferencias entre los sujetos *(filas),* a diferencias entre las condiciones *(columnas)* o a la *interacción* entre filas y columnas.

Podemos distinguir varios planteamientos sencillos que se pueden analizar mediante este modelo de análisis de varianza. Realmente se trata del mismo planteamiento, pero puede resultar sugerente verlo desde diversas perspectivas o desde diversas *preguntas de investigación* que podemos hacernos y cuyas respuestas podemos encontrar en este modelo de análisis de varianza. Se trata de un método de análisis válido siempre que se trate de *muestras relacionadas,* aunque no se trate de diseños experimentales en sentido propio.

1º Cuando los mismos sujetos van a pasar por una serie de tratamientos experimentales. La variable dependiente, la que medimos, es siempre la misma *(medidas repetidas),* como puede ser aprendizaje, satisfacción, mejoría, etc., que medimos después de cada tratamiento.

En este caso cada condición (distintos métodos, ejercicios, etc.) puede influir en los tratamientos o experiencias subsiguientes: el aprendizaje previo, el cansancio, etc., de una ocasión puede estar influyendo en los resultados de la ocasión siguiente. Este problema (derivado del *orden* en el que los sujetos pasan por las distintas experiencias) puede resolverse de dos maneras:

---

[1] También se le denomina a veces de *clasificación doble,* o de *un factor con medidas repetidas,* o de *dos factores con una muestra por grupo* (en EXCEL); el término más genérico y claro es sencillamente *análisis de varianza para muestras relacionadas*: en todas las condiciones (*en cada fila*) tenemos a los mismos sujetos o a *sujetos igualados.*

[2] Si lo que deseamos es conocer no si hay diferencias entre ocasiones, sino si se advierte una tendencia a aumentar o disminuir, tenemos un análisis de varianza específico para verificar *tendencias* que veremos más adelante.

[3] Sobre estas y otras ventajas e inconvenientes de este modelo de análisis de varianza, y sobre el número de sujetos, puede verse Pascual, Frías y García (1996:137; 203) y en monografías como la de Ximénez y San Martín (2000) que incluyen cómo llevar a cabo este análisis de varianza con el SPSS.

a) Mediante diseños *equilibrados (counterbalanced):* los sujetos pasan por los distintos tratamientos en un *orden distinto,* para neutralizar o minimizar los efectos del aprendizaje previo, cansancio, etc.[4]

b) Utilizando *sujetos distintos* en cada condición, pero *igualados* en características importantes (como podrían ser sexo, edad, rendimiento previo, etc.).

Si los tratamientos *(columnas)* son tres (por ejemplo), se divide a la muestra en bloques de tres sujetos igualados en variables que pueden afectar a la variable dependiente (los resultados, el progreso o efecto de una terapia o método, etc.). Si se tratara de ensayar tres procedimientos de aprendizaje, se podría igualar a los sujetos de cada fila en rendimiento previo, motivación, sexo, etc. Preferiblemente los tres sujetos de cada bloque se asignan *aleatoriamente* a los diversos tratamientos (en un diseño experimental en sentido propio).

2º Este modelo de análisis de varianza suele presentarse en el contexto de los diseños experimentales (los mismos sujetos pasan por diversas condiciones o experiencias), pero presentar este modelo de análisis de varianza en referencia únicamante a diseños experimentales, es de hecho muy restrictiva porque puede dejar fuera de nuestra atención otras posibilidades de interés y además muy sencillas y asequibles.

Si los mismos sujetos *valoran* (por ejemplo un una escala escala de 1 a 5) la eficacia, gusto, importancia, utilidad, etc., de una serie de conceptos del mismo ámbito (actividades, motivaciones, etc.) tenemos muestras relacionadas: los sujetos dan su valoración mediante respuestas escritas (se limitan a responder a varias preguntas) según su experiencia, sin necesidad de hacer en ese momento ningún experimento; la vida ya ha hecho que pasen por las diferentes situaciones o condiciones. Es decir, no necesitamos necesariamente que los sujetos pasen por diversas experiencias o condiciones; basta que respondan a una serie de preguntas sobre una serie de conceptos o variables del mismo ámbito. En esta situación el *orden* en el que se valoran los distintos elementos deja de ser un problema (o se puede *alterar el orden* de los ítems presentados en los cuestionarios si pensamos que el orden puede condicionar las respuestas).

3º El ejemplo que nos va a servir para introducir el método también sugiere otras posibilidades: cuando varios profesores evalúan a los mismos alumnos ¿De dónde vienen las diferencias? ¿De que los alumnos son distintos y los profesores tienden a coincidir en sus juicios? (éste sería un resultado deseable) ¿O de que los profesores son distintos en su modo de evaluar? En general siempre que tengamos un grupo de *evaluadores* que valoran a los mismos sujetos (o conceptos, etc.) este análisis nos permitirá apreciar el grado de *consistencia* o de *acuerdo* de los evaluadores al diferenciar unos sujetos (o conceptos) de otros[5].

4º Este análisis de varianza se presta de manera especial a determinar la *jerarquía de valores* de un grupo o simplemente la *jerarquía de preferencias*. Si un grupo valora una serie de conceptos (que pueden expresar valores, como libertad, igualdad, progreso económico, etc.) según su *importancia* (en una escala de *nada* a *muy importante),* podemos ver:

a) En qué medida los sujetos son *consistentes* (están más o menos de acuerdo) ordenando estos conceptos según su importancia; podemos calcular unos coeficientes

[4] Una manera práctica de hacerlo puede verse en Girden (1992:3).

[5] Un ejemplo semejante (con un planteamiento algo más complejo) puede verse en Waddington (2000): cinco profesores corrigen tres veces los trabajos (traducción del inglés al español) de 64 alumnos utilizando cada vez un método distinto de corrección; se trata de verificar con qué procedimiento los profesores difieren menos entre sí.

de fiabilidad que nos indicarán en qué grado los sujetos están de acuerdo diferenciando unos conceptos de otros.

b) Qué valores difieren entre sí (en importancia) por encima de lo que se podría esperar por azar: podemos desembocar en un *orden* que en un cierto grado refleja la jerarquía de valores (o simplemente de *preferencias)* prevalente en el grupo.

De la misma manera que pensamos en valores, podemos pensar en otras categorías: los sujetos pueden valorar *motivaciones, problemas,* etc., incluso lo que puede parecer más trivial, como *colores* o *programas de televisión*. Siempre es posible establecer una jerarquía (si la hay) de *preferencias* (término más global que valores).

En la exposición del método distinguimos dos partes:

1º El *análisis de varianza* propiamente dicho;
2º Cálculos complementarios, como los coeficientes de *asociación* y de *fiabilidad*.

Si en las *columnas* tenemos los ítems de un test o escala, ya veremos más adelante que con este planteamiento podemos calcular el mismo coeficiente de fiabilidad que habitualmente calculamos con otros procedimientos (como el coeficiente α de Cronbach). Una utilidad específica de este modelo de análisis de varianza es precisamente que puede ayudar a la comprensión del coeficiente más utilizado de fiabilidad (α de Cronbach), que posiblemente como mejor se comprende es a partir del análisis de varianza (dedicaremos un apartado específico a este tema con algunos ejemplos).

2. Análisis de varianza

En el ejemplo utilizado para exponer el método (tabla 1)[6] las *filas* son alumnos (f = 6) y las *columnas* son profesores (c = 4) que han evaluado en la misma característica a los seis alumnos. Las preguntas que nos hacemos son estas:

Las diferencias que observamos en los datos:

* ¿Se deben a que los profesores son distintos evaluando? (unos son más benévolos, otros lo son menos…)
* ¿O más bien las diferencias se deben a que los alumnos son distintos en la variable medida, y son así vistos por sus profesores con un grado suficiente de acuerdo?

Si la varianza de las *filas* (alumnos) es estadísticamente significativa (superior a lo aleatorio) tendremos un dato a favor de la unanimidad de los profesores: si hay diferencias se deben sobre todo a que los alumnos son distintos, no a que los profesores son distintos en su estilo de evaluar (pueden ser, por ejemplo, por ejemplo más o menos benévolos) o a la *interacción* profesor-alumno (algunos profesores pueden sentirse inclinados a valorar mejor o peor a determinados alumnos).

[6] Los datos están tomados del ejemplo que presenta el texto de Downie y Heath (1971); el modo de resolver el análisis de varianza que exponemos aquí es sin embargo distinto y más sencillo si se dispone de una calculadora con programación estadística. En EXCEL se hace fácilmente buscando *Análisis de datos* en *Herramientas*; este análisis se denomina en EXCEL *Análisis de varianza de dos factores con una muestra por grupo* (no incluye los contrastes posteriores, ni los coeficientes que se van a exponer); también se puede resolver en programas de Internet como indicamos en al apartado correspondiente.

	profesores (*columnas)*				total	media
	A	B	C	D	*filas*	*filas*
alumnos *(filas)*	10	6	8	7	31	7.75
	4	5	3	4	16	4.00
	8	4	7	4	23	5.75
	3	4	2	2	11	2.75
	6	8	6	7	27	6.75
	9	7	8	7	31	7.75
totales *columnas*	40	34	34	31		
medias *columnas*	6.67	5.67	5.67	5.17		

Tabla 1

De las filas y de las columnas sólo necesitamos *o las medias o los totales;* lo que resulte más cómodo. Con frecuencia las medias y desviaciones (de las filas o de las columnas) son un dato informativo de interés en sí mismo.

Aunque programas como el SPSS o incluso EXCEL nos lo pueden dar resuelto, el procedimiento que exponemos nos ayudar a ver cómo descomponemos la varianza total en las varianzas parciales que nos interesan; además si no disponemos de estos recursos, con una simple calculadora con programación estadística (para calcular medias y las desviaciones) podemos resolverlo con toda facilidad, sobe todo con muestras pequeñas.

1º Cálculos previos

1. Calculamos las medias (o simplemente los totales) de cada fila y de cada columna; ahora suponemos que hemos calculado las *medias* de filas y de las columnas.
2. Calculamos las *desviaciones típicas* (o las varianzas directamente):

del *total* (de todos los datos; N = cxf = 24 datos):	$\sigma_t =$	2.198
de las *medias* de las columnas (Mc):	$\sigma_{Mc} =$	.547
de las *medias* de las filas (Mf):	$\sigma_{Mf} =$	1.873

Si en vez de las medias de las filas o las columnas hemos calculado los *totales* de las filas y/o de las columnas, calculamos sus desviaciones típicas:

desviación típica:	de los totales de las columnas (tc):	$\sigma_{tc} =$	3.269
	de los totales de las filas (tf):	$\sigma_{tf} =$	7.492

Es importante advertir que en todos los casos *N = número de datos,* (no número de sujetos) o *número de filas por número de columnas* (c x f = 24 en este caso).

2º Cálculo de las sumas de cuadrados

Calculamos las *Sumas de Cuadrados* a partir de las desviaciones típicas ya calculadas y de N (número de datos); las fórmulas aparecen en la tabla 2; en este caso utilizamos las desviaciones típicas de las *medias* de filas y columnas. También podemos ir colocando los resultados directamente en la tabla 3.

SC de las filas:	$SC_{filas} = N\sigma^2_{Mf} =$	$(24)(1.873)^2 =$	84.19
SC de las columnas:	$SC_{columnas} = N\sigma^2_{Mc} =$	$(24)(.547)^2 =$	7.18
SC de la interacción:	$SC_{total} - (SC_{fil} + SC_{col}) =$	$(115.95) - (84.19 + 7.18) =$	24.58
SC de los totales:	$SC_{total} = N\sigma^2_t =$	$(24)(2.198)^2 =$	115.95

Tabla 2

Si en vez de calcular las medias de filas y columnas, hemos sumado el *total* de las filas y de las columnas, calcularemos las desviaciones típicas:

de los totales de las filas $\sigma_{tf} = 7.492$

de los totales de las columnas $\sigma_{tc} = 3.269$

En este caso las *Sumas de Cuadrados* son:

de las filas: $SC_f = (\sigma^2_{tf})(\frac{f}{c}) = (7.492)^2 \left(\frac{6}{4}\right) = 84.19$

de las columnas: $SC_c = (\sigma^2_{tc})(\frac{c}{f}) = (3.269)^2 \left(\frac{4}{6}\right) = = 7.13$

Si calculamos las Sumas de Cuadrados a partir de las medias nos dará el mismo resultado que si las calculamos a partir de los totales, salvo pequeñas diferencias por el redondeo de decimales (y que a efectos prácticos no tienen mayor importancia).

Si sumamos las Sumas de Cuadrados de *filas* y de *columnas* veremos que esta suma no es igual a la Suma de Cuadrados total; la variabilidad del total de las puntuaciones no se explica solamente por la variabilidad de las *filas* (diferencias entre los alumnos) y de las *columnas* (diferencias entre los profesores); nos queda la variabilidad debida a la *interacción* entre *filas* y *columnas* (alumnos y profesores). A esta fuente de variación se le denomina *residual* (también *interacción, resto),* y es la que nos queda cuando eliminamos la variabilidad debida a las diferencias sistemáticas de *filas* y de *columnas*. Esta varianza, aleatoria y no controlada, va a ser el término de comparación de las otras dos varianzas.

3º Grados de libertad

Los *grados de libertad* son:

de las filas	f-1=	(6-1)	= 5
de las columnas	c-1=	(4-1)	= 3
de la interacción	(f-1)(c-1)=	(5 x 3)	= 15
del total	N-1=	(24-1)	= 23

4º Tabla de resultados

En la tabla apropiada [3] vamos poniendo los resultados. Al consultar las tablas de la razón F nos fijaremos en los grados de libertad del numerador (de las filas y de las columnas) y del denominador (de la interacción). Esta es la norma general, pero en este caso, muestras relacionadas, esta norma puede variar como indicaremos al final, en el apartado *observaciones sobre los grados de libertad*.

origen de la variación	SC *numerador*	gl *denominador*	CM = SC/gl *Varianza* (σ^2)	F $= \sigma^2/\sigma^2_{\text{interacción}}$	p
filas (alumnos)	84.19	5	$\frac{84.19}{5} = 16.84$	$\frac{16.84}{1.64} = 10.27$	< .01
columnas (profesores)	7.18	3	$\frac{7.18}{3} = 2.39$	$\frac{2.39}{1.64} = 1.46$	> .05 (no sign.)
interacción	24.58	15	$\frac{24.58}{15} = 1.64$		
total	115.95	23			

Tabla 3

5º Interpretación

1. La variabilidad debida a diferencias entre los alumnos es significativamente superior a la debida a la interacción profesor-alumno (grados de libertad: 5 y 15, para α = .05 necesitamos F = 2.90, que es el valor que viene en las tablas, y nosotros hemos obtenido 10.27). La varianza se debe a que los alumnos son distintos (no a que los profesores son distintos); los profesores en este caso *han establecido diferencias* entre los alumnos valorándolos sin grandes discrepancias.

2. Sobre los *grados de libertad* para consultar las tablas hacemos después unas *observaciones importantes*.

El procedimiento habitual (utilizar los grados de libertad de las dos varianzas que se comparan) no siempre es el adecuado en este caso, porque no se cumplen determinados requisitos de este modelo. Como indicamos en las observaciones, los grados de libertad *más seguros* son 1 y N-1.

3. La variabilidad entre los profesores (diferencias sistemáticas de los profesores en su estilo de calificar) no es significativa (grados de libertad 3 y 15, para α = .05 necesitamos F = 3.29). Los profesores no difieren apreciablemente entre sí, y las diferencias que puede haber entre ellos (en su *estilo* de evaluar, más o menos severos) apenas contribuyen a la varianza total.

3. Observaciones sobre los grados de libertad

3.1. La condición de *esfericidad*

Cuando se mide en *varias veces sucesivas a los mismos sujetos* (y siempre que tengamos *muestras relacionadas)* como es frecuente en muchos diseños experimentales (o en estudios exploratorios), estas medidas están correlacionadas; en este caso bajan los *cuadrados medios del término del error* (el denominador de la razón F) y se obtiene con mayor facilidad un valor de F significativo. Un supuesto implícito en este modelo *(medidas repetidas),* para que los valores de F con los grados de libertad especificados antes sean válidos (es decir, que correspondan a la probabilidad indicada en las tablas), es la condición denominada de *esfericidad,* que viene a decir que las covarianzas entre cada par de tratamientos son las mismas (de ocasión a ocasión el cambio es idéntico)[7].

[7] En términos *más estadísticos:* si convertimos la matriz de correlaciones en una matriz de varianzas-covarianzas, las varianzas deberían ser idénticas entre sí y lo mismo las covarianzas.

Cuando esta condición no se cumple, y no suele cumplirse, la distribución de F que viene en las tablas no es exacta y es de hecho muy liberal: se rechaza con demasiada facilidad la Hipótesis Nula.

3.2. Cómo proceder en la práctica

La recomendación más aceptada es la siguiente:

1. Si la F no es significativa con los grados de libertad usuales [(f-1) o (c-1) y (f-1)(c-1)], el resultado no es significativo; hasta aquí es la práctica habitual.

2. Los grados de libertad *más conservadores,* y ciertamente siempre correctos (no inducen a error), son 1 y N -1 (N es aquí el *número de sujetos,* no el número de datos); si con estos grados de libertad el resultado es estadísticamente significativo se puede rechazar la Hipótesis Nula con seguridad.

En el ejemplo que nos ha servido para exponer el método, la razón F correspondiente a las filas es de 10.27; con 5 y 15 grados de libertad debemos alcanzar el valor de 2.90 según las tablas, y lo superamos con creces. Con grados de libertad 1 y N -1 (1 y 5) el valor de F que viene en las tablas es 6.61; también lo superamos (con lo que $p < .05$, pero no $p < .01$) por lo que podemos rechazar la Hipótesis Nula con el criterio más conservador.

3. Si vemos que el resultado no es significativo con 1 y N -1 grados de libertad, todavía puede serlo con los ajustes en los grados de libertad.

Hay dos ajustes posibles, uno más conservador ($\hat{\varepsilon}$, de Box, también denominado de Greenhouse-Geisser) y otro más liberal ($\tilde{\varepsilon}$, de Huynh y Feldt). El estadístico ε *(épsilon)* expresa en qué medida se apartan los datos del requisito de *esfericidad.* A mayor valor de ε, los datos se apartan menos del modelo teórico (de la condición de *esfericidad).*

Estos ajustes son de cálculo laborioso[8] pero están programados en programas informáticos como el SPSS. Al menos conviene conocer a) cómo se utilizan estos coeficientes y b) cómo proceder en la práctica habitual cuando no disponemos de alguna de alguna de las modalidades del coeficiente ε.

a) *Cómo utilizar el coeficiente* ε

El valor de ε es siempre inferior a la unidad; cuando $\varepsilon = 1$, la *esfericidad* es perfecta. Se trata de un coeficiente por el que multiplicamos los grados de libertad de las dos varianzas que contrastamos; los grados de libertad quedan así reducidos y cuesta más rechazar la Hipótesis Nula. Si tenemos que, por ejemplo, $\varepsilon = .60$, los grados de libertad iniciales, 5 y 15, se convierten en:

$$5\ (.60) = 3$$
$$15\ (.60) = 9$$

Los grados de libertad son menos y consecuentemente necesitamos un valor mayor de F para poder rechazar la Hipótesis Nula.

[8] Pueden verse las fórmulas en diversos autores (por ejemplo, Girden, 1992:19; Kirk: 1995:281; Ximénez y San Martín, 2000:42); Kirk es el que expone de manera más sencilla cómo calcular ε a partir de la matriz de covarianzas (también nos lo da el SPSS). Ambos autores repiten la recomendación de utilizar como grados de libertad 1 y N-1 como medida de seguridad, aunque esta recomendación puede resultar muy conservadora. Una explicación de estos coeficientes puede verse también en Llobell, Frías y García (1996:158) y en García, Frías y Llobell (1999).

b) *Cómo proceder en la práctica habitual*

Lo más seguro es suponer que no se cumple la condición de *esfericidad* y proceder teniendo en cuenta estas cautelas:

1º La práctica *más segura* es utilizar como grados de libertad 1 y N -1; si con estos grados de libertad el resultado es significativo, ciertamente lo es y podemos rechazar la Hipótesis Nula. Con frecuencia, y en resultados muy claros, superamos con creces los valores que vienen en las tablas.

2º Aun así éste es un criterio muy conservador (grados de libertad 1 y N-1) y podemos aceptar como no significativos resultados que sí lo son (nos puede dar *falsos negativos*).

Si el resultado es significativo con los grados de libertad convencionales, (c-1) o (f-1) y (c-1)(f-1), pero no lo es con 1 y N -1, es entonces cuando deberíamos aplicar uno de estos dos ajustes (programados en el SPSS) (Girden, 1992:21):

Si $\varepsilon > .75$: $\tilde{\varepsilon}$ (de Huynh y Feldt, el ajuste más liberal)

Si $\varepsilon < .75$ (o si no sabemos nada sobre ε): $\hat{\varepsilon}$ (de Box o Greenhouse-Geisser, el ajuste más conservador)

4. Contrastes posteriores

Si nos interesan los contrastes posteriores, suele recomendarse el contraste de Tukey[9]; una propuesta más segura (sobre todo con muestras pequeñas y cuando no se cumple la condición de esfericidad) que vemos recomendada[10] son los contrastes de Bonferroni (o Dunn-Bonferroni)[11].

Frecuentemente cuando tenemos muestras relacionadas lo que más nos puede interesar son los coeficientes a) de asociación (η o más habitualmente η^2) y b) de fiabilidad expuestos más adelante. En este ejemplo no tiene interés comprobar entre qué profesores hay diferencias significativas, posiblemente incluso aun cuando la razón F de los profesores fuera significativa tampoco tendría especial interés; en otros planteamientos sí puede interesar ver qué *columnas* o *condiciones* difieren entre sí.

5. Análisis complementarios: los coeficientes de asociación y de fiabilidad

Como en planeamientos semejantes, la razón F (como la t de Student) nos remiten a una *probabilidad de ocurrencia,* pero no a una *magnitud*.

Hay dos cálculos complementarios de interés, a) coeficientes de asociación, como en otros modelos de análisis de varianza y b) coeficientes de fiabilidad más propios de este modelo de análisis de varianza para muestras relacionadas.

---

[9] El SPSS nos da estos contrastes posteriores (Tuckey y Bonferroni)).

[10] Como Toothaker y Miller (1996:557-558), Girden (1992:29), Ximénez y San Martín (2000:49).

[11] En los contrastes de Bonferroni se utiliza la t de Student convencional (para muestras relacionadas en este caso), pero con un nivel de confianza más estricto; el nivel de confianza adoptado (por lo general .05) se divide por el número de comparaciones previstas; si nuestro nivel de confianza es $\alpha = .05$ (lo habitual) y vamos a hacer tres comparaciones, para rechazar la hipótesis nula a este nivel ($\alpha = .05$) debemos obtener una probabilidad de $p = .05/3 = .0167$. Aunque los contrastes de Bonferroni suelen ser valorados como excesivamente conservadores (Jaccard, 1998; Perneger, 1998), sí parecen más aconsejables en el caso de medidas repetidas.

5.1. Coeficientes de asociación (η^2 y $\eta^2_{parcial}$)

Como en otros planteamientos de análisis de varianza, un coeficiente muy utilizado es el coeficiente *eta* (η); es un coeficiente de asociación válido cuando una variable es categórica, no continua (como son los profesores o *columnas* en este ejemplo). También, y como en otros modelos de análisis de varianza, se utiliza ω^2, pero η^2 (o η) es de comprensión más intuitiva (una proporción) que ω^2 y más sencillo de cálculo[12]. Una razón F estadísticamante significativa, que muy probable obtenerla con muestras grandes, nos da seguridad para concluir que hay diferencias, pero no dice nada sobre la relevancia o magnitud de estas diferencias; ésta es la información que nos dan estos coeficientes, y en este sentido (expresan una magnitud, no una probabilidad) se trata de un *tamaño del efecto* que permite completar la conclusión alcanzada con la razón F propia del análisis de varianza.

Como en coeficientes análogos (como r) η elevada al cuadrado indica la proporción de varianza en la variable continua (las calificaciones de los alumnos en este ejemplo) atribuible a diferencias en la variable categórica (diferencias entre los profesores en este caso).

En el caso del análisis de varianza para muestras independientes tenemos un único coeficiente η, en el caso de muestras relacionadas tenemos dos, uno corresponde a las columnas y otro a las filas; en cada caso dividimos cada suma de cuadrados por la suma de cuadrados total (fórmulas [1] y [2]).

$$\eta_{columnas} = \sqrt{\frac{SC_{columnas}}{SC_{total}}} \qquad [1]$$

$$\eta_{filas} = \sqrt{\frac{SC_{filas}}{SC_{total}}} \qquad [2]$$

En nuestro ejemplo:

$$\eta_{columnas} = \sqrt{\frac{7.18}{115.95}} = .248 \text{ y } \eta^2 = .06$$

$$\eta_{filas} = \sqrt{\frac{84.19}{115.95}} = .726 \text{ y } \eta^2 = .53$$

Fijándonos en η^2 vemos que la proporción de varianza en la variable dependiente (las calificaciones) se explica sobre todo (53 % de la varianza) por diferencias entre los alumnos y apenas (6 %) por las diferencias entre los profesores. Todos estos coeficientes conviene interpretarlos también en términos relativos, comparando unos con otros.

Eta cuadrado parcial ($\eta^2_{parcial}$)[13] expresa la proporción de varianza de una variable dependiente explicada por la variable independiente manteniendo constantes otras fuentes de error; es decir prescindimos de otras fuentes de variabilidad. En la fórmula [1] (η, no η^2; normalmente interesa la [1] que corresponde a la variable independiente) en el denominador en vez de la suma de cuadrados total, tendremos *la suma de cuadrados de las columnas más la suma de cuadrados del error* (de la interacción) (Rosenthal y Rosnow, 1991:463):

[12] Las fórmulas de ω^2 pueden verse en muchos textos, como Ximénez y San Martín (2000:46).

[13] Este coeficiente *eta cuadrado parcial* correspondiente a las columnas (variable independiente) lo calcula el SPSS como parte del *output*.

$$\eta_{parcial} = \sqrt{\frac{SC_{columnas}}{SC_{columnas} + SC_{error}}} \qquad [3]$$

En este ejemplo, y elevando al cuadrado el coeficiente η, tenemos:

$$\eta^2_{parcial} = \frac{7.18}{7.18 + 24.58} = .226$$

Las diferencias entre profesores explican el 22.6% de la varianza, si prescindimos de otras fuentes de variabilidad (manteniendo constantes a los alumnos). Si hacemos lo mismo con los alumnos (*filas*) tendremos 84.19/(84.19+24.58) = .774; el 77.4% de la varianza queda explicado por diferencias entre los alumnos (manteniendo constantes a los profesores).

5.2. Coeficientes de fiabilidad

5.2.1. Fiabilidad de filas y columnas

Aquí hablamos de fiabilidad en el sentido de *consistencia interna;* en este ejemplo (profesores evaluando a los mismo alumnos) nos interesa comprobar el grado de consistencia o de homogeneidad (o de *grado de acuerdo*) de los profesores evaluadores *(las columnas)* al diferenciar a los alumnos *(las filas)*; en otros planteamientos nos interesa sobre todo verificar la consistencia o grado de acuerdo de las filas (sujetos) diferenciando a las columnas.

Este cálculo de la fiabilidad o consistencia de las columnas al clasificar o diferenciar a las filas (o las columnas a las filas) suele ser de interés en muchos planteamientos del análisis de varianza para muestras relacionadas; por esta razón dedicamos después a la fiabilidad un apartado adicional. Podríamos definir la fiabilidad en este caso como el *grado de acuerdo* de la filas diferenciando a las columnas o de las columnas diferenciando a las filas.

En otros casos lo que interesa es comprobar si las *filas* (los sujetos) diferencian bien a las *columnas*. Si por ejemplo los alumnos evaluaran a una serie de profesores (o de experiencias, etc.), nuestra hipótesis sería que los profesores son distintos (o las experiencias son distintas) y que los alumnos son consistentes al diferenciar las columnas. En este caso un resultado esperado hubiera sido el contrario al de este ejemplo: la varianza se debería no a que los alumnos son distintos, sino a que los profesores, o las experiencias, son distintas.

Las fórmulas de los *coeficientes de fiabilidad* son las siguientes[14]:

Fiabilidad de las *columnas* (de todas): $$r_{cc} = \frac{CM_{filas} - CM_{error}}{CM_{filas}} \qquad [4]$$

Fiabilidad de *una sola columna:* $$r_{cc} = \frac{CM_{filas} - CM_{error}}{CM_{filas} + (k-1)CM_{error}} \qquad [5]$$

Fiabilidad de las *filas:* $$r_{ff} = \frac{CM_{columnas} - CM_{error}}{CM_{columnas}} \qquad [6]$$

[14] Tomamos las fórmulas de Guilford y Fruchter (1973) y de Nunnally y Bernstein (1991) pero podemos encontrarlas en muchos textos.

Los símbolos son ya conocidos:

r_{cc} = fiabilidad de las columnas;
r_{ff} = fiabilidad de las filas;
CM_f= Cuadrados Medios (o varianza) de las filas;
CM_c= Cuadrados Medios (o varianza) de las columnas;
CM_e = Cuadrados Medios (o varianza) del término del *error;* en este caso de la *interacción;*
k = número de columnas (profesores en nuestro ejemplo)

Como podemos ver las fórmulas de la fiabilidad o *consistencia interna* de las filas [6] son iguales a las de las columnas [4], substituyendo CM_f por CM_c

En este ejemplo lo que nos interesa es el cálculo de fiabilidad (*grado de acuerdo*) de todos los profesores (las columnas); queremos comprobar en qué grado establecen diferencias entre los alumnos de manera consistente.

La fiabilidad de los profesores es en este caso:

$$r_{cc} = \frac{16.84 - 1.64}{16.84} = .903$$

Esta *cuantificación* de la consistencia añade información al valor significativo de F, y nos dice que los profesores están *muy de acuerdo* o son *muy consistentes* al evaluar a los alumnos (los *ordenan* de manera muy parecida, aunque pueden ser muy distintos en sus valoraciones absolutas).

De manera análoga, y si tiene sentido en el planteamiento de los datos, podemos calcular la fiabilidad de las filas (hasta qué punto son las filas las que discriminan *consistentemente* a las columnas). Este sería el caso si las filas (alumnos) *juzgaran a las columnas* (profesores, actividades, etc.).

Estos coeficientes son análogos al coeficiente α de Cronbach; en realidad se trata de lo mismo. En un test o escala podemos calcular el coeficiente α mediante el análisis de varianza, poniendo a los ítems en las columnas y a los sujetos en las filas.

La pregunta que nos hacemos en el caso de la fiabilidad de un test es semejante:

¿Hasta qué punto los ítems *(columnas)* son *consistentes (‘están de acuerdo’*) discriminando, diferenciando a los sujetos *(filas)* en aquello que es común a todos los ítems (lo que estamos midiendo)?

Si hay diferencias en los totales (varianza total) esperamos que se deba a que los sujetos *medidos* son distintos, no a que los *ítems* son distintos y miden cosas distintas (por eso estos coeficientes se denominan de *homogeneidad:* de homogeneidad o *consistencia interna* de las *columnas* o *jueces,* en este caso). Si en el ejemplo que nos ha servido para exponer el método suponemos que los cuatro profesores son ítems de una escala, y calculamos el coeficiente α de Cronbach con la fórmula habitual, llegaremos al mismo resultado.

Esta relación entre fiabilidad y análisis de varianza la explicamos con más detalle en el apartado siguiente.

5.2.2. Relación entre fiabilidad *(consistencia interna)* y análisis de varianza para muestras relacionadas.

Explicamos con más detalle esta relación por el interés que tiene la fiabilidad tanto en psicometría *(fiabilidad de los tests)* como en muchos planteamientos experimentales: nos puede interesar verificar la consistencia o fiabilidad de una serie de jueces cuando evalúan una serie de sujetos u objetos. Posiblemente la fiabilidad, en su sentido más habitual (coeficientes de *consistencia interna* referidos a tests y escalas) se entiende mejor a través del análisis de varianza.

Para entender el concepto de fiabilidad en este contexto es útil la analogía con el concepto de la *unanimidad de unos jueces* evaluando (o con más propiedad *ordenando* o *clasificando* de más a menos…) a una serie de sujetos.

Para *ordenar* o *diferenciar* bien hace falta:

1º Que los *jueces* sean *coherentes* entre sí, es decir, tengan el mismo criterio, se fijen en lo mismo, estén básicamente de acuerdo…

2º Que los *sujetos* sean distintos según el *criterio compartido* por los jueces… (se ordena mejor a los muy diferentes)

Vamos a pensar en dos situaciones distintas.

Situación 1ª

Los *jueces* son los *ítems* de un test (y aquí está la *analogía,* en pensar que los ítems van a *juzgar* a una serie de sujetos): todos los *ítems-jueces* participan del mismo criterio (es decir *miden lo mismo),* y tenemos este resultado (tabla 4):

	ítems						
sujetos	1º	2º	3º	4º	5º	6º	*medias de los sujetos (filas)*
1	2	2	1	2	2	1	1.67
2	1	1	2	1	1	2	1.33
3	3	4	4	3	3	4	3.50
4	4	3	3	4	4	3	3.50
5	5	5	5	5	5	5	5.00
6	6	6	6	6	6	6	6.00
medias (y σ) de los ítems (columnas)	3.5 1.707	3.5 1.707	3.5 1.707	3.5 1.707	3.5 1.707	3.5 1.707	

Tabla 4

Qué vemos en estos resultados:

a) Las medias de los ítems *(columnas)* son semejantes (idénticas en este caso);
b) Las medias de los sujetos *(filas)* son muy distintas.

Además vemos que las correlaciones entre los *ítems-jueces* son altas o al menos claras: coinciden en asignar puntuaciones altas o bajas a los mismos sujetos (los sujetos 1 y 2 puntúan bajo en todos los ítems, los sujetos 3 y 4 puntúan en el centro de la escala en todos los ítems, y los sujetos 5 y 6 puntúan alto en todos los ítems).

Es claro que sin un suficiente *grado de acuerdo en los jueces* o sin suficientes *diferencias entre los sujetos,* no se darían estas relaciones. Los coeficientes de fiabilidad vienen a expresar el grado de relación global entre los jueces (por eso se denominan de *consistencia interna*… entre los ítems o jueces). Es más, sin diferencias entre los evaluados

no se puede comprobar *empíricamente* (mediante coeficientes de correlación) que los jueces *miden lo mismo.*

Conclusión: los ítems (que son los *jueces*…) *tienen un alto de acuerdo ordenando* o *diferenciando* a los sujetos, estableciendo diferencias con el mismo criterio; las *columnas* (ítems) son *fiables* cuando clasifican a las *filas*. De este *grado de acuerdo* (comprobado) *deducimos* que miden lo mismo. Aunque refiriéndonos a los ítems de tests y escalas, de la mera relación entre los ítems (que si es alta desembocará en un coeficiente de fiabilidad alto) no podemos *sin más* concluir que los ítems *miden lo mismo* (hacen falta *consideraciones conceptuales* sobre la formulación de los ítems; los sujetos pueden quedar *ordenados* de manera semejante en todos los ítems y a pesar de eso es posible que las formulaciones de los ítems no reflejen con nitidez un mismo rasgo bien definido).

Esta *suficiente grado de acuerdo* de los ítems (expresión un tanto impropia tratándose de ítems) es lo que *cuantificamos* mediante el coeficiente α de Cronbach, pero podemos llegar al mismo resultado con los coeficientes ya vistos y que son posteriores al análisis de varianza. Lo que sucede es que si lo que queremos es calcular la fiabilidad de un test, las fórmulas habituales de la fiabilidad son más cómodas que el análisis de varianza.

Situación 2ª

Los jueces son unos *sujetos* que con *idéntico criterio* (utilidad, eficacia, etc.) *valoran* una serie de ítems: métodos, actividades, profesores, etc.

Tenemos este resultado hipotético (tabla 5):

	ítems						
sujetos	1º	2º	3º	4º	5º	6º	*medias de los sujetos (filas)*
1	2	6	4	4	6	1	3.83
2	1	6	3	4	6	2	3.67
3	2	6	4	3	6	1	3.67
4	2	5	3	4	6	1	3.50
5	1	5	4	5	6	1	3.67
6	1	6	4	3	6	2	3.67
medias (y σ) de los ítems (columnas)	1.50 .50	5.67 .47	3.67 .47	4.00 .63	6.00 .00	1.33 .47	

Tabla 5

Esta situación ejemplifica un modelo potencialmente muy útil de investigación mediante el análisis de varianza para *muestras relacionadas.*

Qué vemos ahora:

a) Las medias de los sujetos son muy parecidas unas a otras, apenas hay diferencias;
b) Las medias de los ítems son muy desiguales;

Conclusión: tenemos algo parecido a la tabla anterior [4], pero al revés: ahora son los sujetos quienes muestran un gran acuerdo en cómo valoran los ítems (actividades, experiencias, profesores, etc.). Las medias de los sujetos son muy parecidas, pero las de los ítems son muy distintas, y son los ítems los que quedan *ordenados* en una jerarquía clara. Ahora el concepto de *fiabilidad*, de consistencia interna, lo aplicamos a los sujetos.

En la primera situación los ítems *(las columnas)* ordenan, clasifican, etc., a los sujetos *(filas)* de manera consistente; en la segunda situación son los sujetos *(filas)* quienes ordenan a

las *columnas,* valoran las *filas* con bastante unanimidad en el criterio establecido (eficacia, agrado, etc.).

Los *coeficientes de fiabilidad (homogeneidad, consistencia interna)* expresan en qué grado las columnas (ítems de un test) *miden lo mismo*[15] y diferencian a los sujetos que son distintos en aquello que tienen en común los ítems (fórmulas del coeficiente α de Cronbach, coeficientes de Kuder-Richardson). Y *a la inversa:* se puede calcular la fiabilidad de los sujetos (las filas) cuando *ordenan* los ítems (las columnas).

Estos coeficientes de fiabilidad, de las filas ordenando a las columnas y de las columnas ordenando a las filas, son los que calculamos a partir de los resultados del análisis de varianza para muestras relacionadas (relacionadas porque los sujetos son los mismos en todas las condiciones…o en todos los ítems).

En cualquier caso la fiabilidad tiende a ser alta:

a) Cuando los jueces son consistentes, participan del mismo criterio;
b) Cuando los jueces son muchos (se minimizan las inconsistencias)
c) Cuando los sujetos (u objetos) juzgados son muy diferentes entre sí (es más fácil *ordenarlos).*

Esta grado de acuerdo o consistencia interna se manifiesta en las correlaciones entre los jueces (como entre los ítems de un test); la correlación implica diferencias sistemáticas en los sujetos u objetos evaluados.

Resumiendo: fiabilidad y análisis de varianza

Pregunta común a estos planteamientos: *¿De dónde vienen las diferencias (la varianza total)?* ¿De que las filas son distintas o de que las columnas son distintas?

Una solución directa nos la da el *análisis de varianza para muestras relacionadas,* que divide (analiza) la varianza total en tres varianzas parciales:

1ª Varianza debida a que las filas son distintas…
2ª Varianza debida a que las columnas son distintas…
3ª Varianza aleatoria, debida a la interacción fila-columna (no debida a que las columnas o las filas son sistemáticamente distintas).

Resultados del análisis varianza:

1. Nos dice si las varianzas 1ª y/o 2ª son significativamente distintas de la varianza 3ª (superiores a lo que consideramos aleatorio).

2. Podemos calcular el coeficiente de fiabilidad de las filas cuando establecen diferencias entre las columnas y el coeficiente de fiabilidad de las columnas cuando establecen diferencias entre las filas. Estos coeficientes son idénticos (el mismo resultado) al coeficiente α de Cronbach.

---

15 La expresión usual *miden lo mismo* no es exacta; lo que cuantifican estos coeficientes es en qué grado los ítems están relacionados, y esta relación inter-ítem nos *confirma* (o *deducimos)* que de alguna manera miden lo mismo, expresan el mismo rasgo. También puede suceder que tengamos una fiabilidad alta con ítems que claramente no miden el mismo rasgo (con criterios conceptuales) pero que están claramente relacionados entre sí; siempre es necesaria una *evaluación cualitativa.*

5.2.3. Cuando las dos razones F (de las *filas* y de las *columnas*) son estadísticamente significativas

En cualquiera de estos planteamientos podemos encontramos con que las dos razones F (de las filas *y* de las columnas) son significativas (superiores a lo que podemos juzgar como aleatorio), aunque por lo general (no siempre), cuando tenemos este resultado, una de las razones F y uno de los coeficientes de fiabilidad suelen ser apreciablemente mayores que los otros. Aun así este resultado puede desconcertar: *¿Pueden simultáneamente clasificar, diferenciar, ordenar… las filas a las columnas y las columnas a las filas?*

Vamos a pensar en una situación de fácil comprensión: los alumnos de una clase *(filas)* evalúan a sus profesores *(columnas).* Un resultado ideal sería que las diferencias provengan sobre todo de que los profesores son distintos (y son así vistos por los alumnos) y no de que los alumnos son distintos en su manera de evaluar. En este caso tendríamos la F de las columnas *(profesores)* significativa y la fiabilidad de las filas *(alumnos)* alta: los alumnos distinguen bien unos profesores de otros[16].

Imaginemos que las dos razones F son significativas: los profesores son distintos y también los alumnos son distintos en su forma de evaluar, en ambos casos *sus* varianzas son superiores a lo que podemos encontrar por azar. Este resultado posible ¿Implica contradicción? ¿Es interpretable?

Un resultado así es posible, no es contradictorio y es interpretable. Lo que sucede es que los resultados que obtenemos y las interpretaciones que hacemos no pueden encerrarse en un *sí* o un *no,* en un *blanco* o *negro.* Es fácil verlo con un ejemplo *muy simplificado.*

Vamos a suponer que dos alumnos evalúan a dos profesores en una escala de 1 a 6 en un rasgo cualquiera. Un profesor es mejor que el otro a juicio de los dos alumnos, pero los dos alumnos tienen un *modo distinto* de evaluar:

	Profesor 1	Profesor 2
Alumno *benévolo:*	6	5
Alumno *crítico:*	2	1

Alumno benévolo. Un alumno califica al mejor profesor con un 6 y al peor profesor con un 5. Aunque distingue entre los dos profesores, este alumno tiende a juzgar con benevolencia; le gustan los dos profesores, al menos los evalúa bien, aunque uno le gusta más que el otro.

Alumno crítico. El otro alumno califica al mismo mejor profesor con un 2 y al otro profesor, que a su juicio es también peor, con un 1. Estima que un profesor es mejor que el otro, pero califica bajo a los dos; no le gusta ninguno de los dos, pero considera que uno es peor que el otro.

En un caso así (en un ejemplo real tendríamos más alumnos) las dos razones F van a ser significativas: los alumnos son *consistentes* cuando *ordenan* a los profesores según su calidad (por eso la F de los profesores es significativa), pero ellos también son *sistemáticamente distintos* en su forma de evaluar. La fiabilidad de los alumnos sería alta en este caso porque

[16] En este caso (cuestionario de evaluación del profesorado) habría que calcular la fiabilidad de *cada ítem;* también se podría calcular la fiabilidad de *subconjuntos de ítems homogéneos* (que describen el mismo rasgo) sumados en un total, como si se tratara de un test convencional.

coinciden en señalar cuáles son los *mejores* y los *peores* profesores según su juicio, aunque no coincidan en las valoraciones absolutas (tienden a coincidir en el *orden)*[17].

Cuando *en un cierto grado* se dan estas diferencias entre los alumnos, *su* razón F es significativa, es decir, la diversidad entre los alumnos *(su varianza)* es superior a lo meramente aleatorio: las diferencias entre los alumnos no son casuales, es que evalúan de distinta manera. Pero esto no quiere decir que no diferencien adecuadamente a unos profesores de otros: coinciden en el *orden* en que sitúan a sus profesores; en identificar al menos, según su propio juicio, a los mejores y peores profesores. Esto es lo que significa una fiabilidad alta de los alumnos: coincidencia básica en *diferenciar* a unos profesores de otros; en este sentido decimos que los alumnos son *unánimes,* aunque naturalmente esta unanimidad no lo es en sentido propio (la fiabilidad admite *grados;* en un caso como éste, si la fiabilidad es alta, diríamos que hay suficiente *convergencia* o *grado de acuerdo* en sus juicios).

Este ejemplo es aplicable a cualquier otra situación, y algo análogo sucede con la fiabilidad de los tests (y exámenes). Los ítems, cuando se suman en una puntuación total, pueden diferenciar adecuadamente a los sujetos; unos tienden a puntuar sistemáticamente más alto en todos ellos, otros más bajo, etc.; pero esto no quiere decir que los ítems sean de *parecida dificultad*. Una fiabilidad alta en un test querría decir que los ítems están básicamente *de acuerdo* (como si fueran *jueces*) cuando diferencian (*ordenan,* con más propiedad) a unos sujetos de otros. En el apartado siguiente aclaramos más, con un ejemplo, la relación entre el análisis de varianza y la fiabilidad de tests y escalas.

5.2.4. Análisis de varianza para muestras relacionadas y el coeficiente α de Cronbach

Es importante ver la relación existente entre estos coeficientes de fiabilidad derivados directamente del análisis de varianza, y el más conocido coeficiente de *consistencia interna* α de Cronbach, que utilizamos habitualmente para calcular la fiabilidad de tests y escalas. Se trata de lo mismo, aunque cuando nos referimos a la fiabilidad de los tests lo explicamos desde otra perspectiva[18].

Qué significa la fiabilidad *convencional* de tests y escalas también podemos entenderlo a través del análisis de varianza como ya ha quedado explicado anteriormente al inicio de este apartado[19]; en la *situación 1ª* veíamos cómo los ítems de un supuesto test eran *consistentes* cuando diferenciaban a los sujetos. Ahora lo vamos a ver con un ejemplo. Los datos son ficticios, pero nos permiten *visualizar* cómo ambos procedimientos nos llevan al mismo resultado.

En este ejemplo (tabla 6) tenemos las respuestas de cuatro sujetos a un test compuesto por seis ítems.

---

17 Un coeficiente de correlación alto entre dos elementos quiere decir que los sujetos los *ordenan* de manera semejante, pero no que tengan valores absolutos semejantes.

18 Un tratamiento más completo de la fiabilidad de los tests en Morales (2008, capítulo 6).

19 La relación entre fiabilidad y análisis de varianza la puso de relieve ya hace años Hoyt (1941, 1952) y está bien explicada en Rosenthal y Rosnow (1991:55) y en Nunnally y Bernstein (1994:274)

	ítems						
sujetos	nº 1	nº 2	nº 3	nº 4	nº 5	nº 6	total
a	6	6	5	4	6	6	33
b	6	5	5	4	3	5	28
c	3	3	2	3	4	4	19
d	4	3	2	1	1	2	13
media	4.75	4.25	3.50	3.00	3.50	4.25	23.25
desviación	1.29	1.29	1.50	1.225	1.80	1.48	7.758

Tabla 6

Estos son los datos que solemos tener cuando analizamos un test: el total de cada sujeto y la media y desviación típica de cada ítem.

Vamos a calcular en primer lugar el coeficiente α de Cronbach con la fórmula usual:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2}\right) = \left(\frac{6}{6-1}\right)\left(1 - \frac{1.29^2 + 1.29^2 + 1.50^2 + 1.225^2 + 1.80^2 + 1.48^2}{7.758^2}\right)$$

$$= (1.2)\left(1 - \frac{12.509}{60.1865}\right) = .95$$

Hacemos ahora, con los mismos datos, un análisis de varianza para muestras relacionadas con estos resultados (tabla 7):

origen de la variación	*Suma de Cuadrados*	*Grados de libertad*	*Cuadrados medios*	F
Ítems	8.375	5	1.675	2.482 p>.05
Sujetos	40.125	3	13.375	19.81 p <.05
Ítems x sujetos	10.125	15	.675	
Total	58.625	23		

Tabla 7

La varianza significativa es la que corresponde a los sujetos; son los sujetos los que difieren entre sí y no los ítems entre sí; ahora bien, si los sujetos difieren entre sí es porque los ítems establecen diferencias entre los sujetos *como si* se tratara de jueces cuyas opiniones coinciden (en este ejemplo, tabla 6, puede verse a simple vista).

Ya sabemos cómo cuantificar esta *fiabilidad* de los ítems (de las *columnas* en este caso, fórmula [4]):

$$r_{cc} = \frac{CM_{filas} - CM_{error}}{CM_{filas}} = \frac{13.375 - .675}{13.375} = .95$$

Hemos llegado al mismo valor del coeficiente α de Cronbach. Esta *fiabilidad de los ítems* es lo que denominamos *consistencia interna*. Si los ítems fueran unos jueces diríamos que los ítems juzgan de manera semejante a los sujetos; de esta consistencia interna deduciremos que *miden lo mismo,* aunque en esta deducción puede haber otras fuentes de error en las que no entramos aquí[20].

[20] Esta *consistencia interna* es puramente empírica; *de hecho* los ítems ordenan de manera semejante a los sujetos,

Habitualmente calculamos la fiabilidad de los tests y escalas mediante el cálculo directo del coeficiente α de Cronbach, pero posiblemente es *desde* el análisis de varianza como se entiende mejor qué es lo que *cuantifican* estos coeficientes.

6. Un ejemplo de análisis de varianza para muestras relacionadas

6.1. Análisis de varianza

En el ejemplo que nos ha servido para introducir el análisis de varianza para muestras relacionadas (o emparejadas) teníamos en las *columnas* (variable independiente) cuatro profesores que han calificado a los mismos alumnos (*filas*). El interés en este ejemplo era verificar que no había diferencias significativas entre los profesores y ver la fiabilidad o *grado de acuerdo* de los profesores al evaluar a los mismos alumnos. Ahora presentamos otro ejemplo en el que el interés va en otra dirección: verificar las diferencias entre las *columnas* (condiciones experimentales) y la fiabilidad o grado de acuerdo de los sujetos (*filas*) diferenciando unas columnas de otras. Este ejemplo es semejante al presentado en la situación 1 (apartado 5.2.2) y lo exponemos con cierto detalle y con un caso real porque puede sugerir otros planteamientos parecidos.

En una residencia de estudiantes (chicas) se pregunta a 10 residentes por sus preferencias a la hora de realizar 7 tareas distintas (*ayudar en la limpieza de los baños, la sala de estar, sala de estudio, atender el teléfono, barrer el pasillo, ayudar en lavadero y en la cocina*).

Las respuestas valorativas van desde 1 *(no me gusta nada)* a 6 *(me gusta mucho)*[21]. En la tabla 8 figuran las respuestas de las residentes, la media y desviación típica (de la muestra, dividiendo por N) de cada *columna* o tarea y el total de cada alumna al sumar todas sus respuestas.

residentes	1. *Baño*	2. *Sala de estar*	3. *Sala de estudio*	4. *Teléfono*	5. *Pasillo*	6. *Lavadero*	7. *Cocina*	Total filas
1	1	2	3	5	5	3	2	21
2	1	2	3	6	1	2	6	21
3	3	1	1	5	5	4	4	23
4	1	2	2	3	2	4	4	18
5	3	2	1	5	2	4	3	20
6	2	1	1	1	2	3	2	12
7	1	2	2	3	5	4	3	20
8	1	2	2	4	2	6	3	20
9	1	4	3	5	6	2	6	27
10	1	2	2	4	5	4	4	22
M *columnas*	1.5	2	2	4.1	3.5	3.6	3.7	
σ *columnas*	.806	.775	.775	1.375	1.113	1.345	1.345	

Tabla 8

pero esto no quiere decir *necesariamente* que *conceptualmente* midan un mismo rasgo *bien* definido; un grupo de niños de diversas edades pueden quedar ordenados de manera semejante en peso y altura, sin que esto quiera decir que peso y altura *midan lo mismo*.

[21] De un trabajo de Sara Lozano, alumna de 2º de Psicopedagogía, curso 1998-1999

La tabla de resultados del análisis de varianza (EXCEL) la tenemos en la tabla 9.

origen de la variación	SC *numerador*	gl *denominador*	CM = SC/gl (*varianza)*	F	p	F crítico (α = .05)
filas *(residentes)*	18.629	9	2.070	1.401	0.211	2.059
columnas *(tareas)*	65.086	6	10.848	7.343	0.000	2.272
interacción *(error)*	79.771	54	1.477			
total	163.486	69				

Tabla 9

Observamos que:

La F de las *filas* no es estadísticamente significativa; las diferencias entre las alumnas están dentro de lo aleatorio.

Aquí conviene tener claro en qué *no difieren* las alumnas: no difieren significativamente en sus *totales* (si sumamos a cada una todas sus respuestas) o en sus medias. A simple vista se detectan diferencias que parecen grandes (la más baja es 12 y la más alta es 27) pero con tan pocos sujetos cualquier diferencia entre dos sujetos tiene una probabilidad de ocurrir aleatoriamente superior al 5%. En cualquier caso esta F de las filas no tiene aquí interés interpretativo; nuestro interés está en verificar si hay diferencias entre las columnas (entre las tareas)..

La F de las *columnas* sí es estadísticamente significativa; entre las tareas hay diferencias superiores a lo que podemos esperar por azar.

Tenemos el problema de la condición de *esfericidad* que no hemos comprobado (sí se comprueba en el SPSS) y que no suele cumplirse. En este caso ya hemos visto que los grados de libertad más seguros y conservadores al consultar las tablas de la razón F son 1 y N-1 (en este caso 1 y 9). Con un nivel de confianza de α = .05 y grados de libertad 1 y 9 vemos en las tablas que el valor de F que necesitamos es de 5.12 y el nuestro (7.343) es superior por lo que podemos rechazar la Hipótesis Nula y afirmar que existen diferencias significativas entre las tareas.

6.2. Coeficientes de fiabilidad y η^2

Si las tareas difieren entre sí es porque hay *un cierto grado de acuerdo* entre las residentes al valorar de distinta manera las tareas que tienen que compartir; este grado de acuerdo es la *fiabilidad de las filas* (fórmula [6]):

$$r_{ff} = \frac{CM_{columnas} - CM_{error}}{CM_{columnas}} = \frac{10.848 - 1.477}{10.848} = .86$$

Podemos afirmar que el grado de acuerdo de las residentes al diferenciar unas tareas de otras es apreciablemente grande.

La fiabilidad de las columnas (tareas) diferenciando a las filas (sujetos) no tiene especial interés pero es ilustrativo calcular este coeficiente (fórmula [4])

$$r_{cc} = \frac{CM_{filas} - CM_{error}}{CM_{filas}} = \frac{2.070 - 1.477}{2.070} = .29$$

Obviamente es un coeficiente muy bajo; lo que tiene interés es caer en la cuenta de cuál hubiera sido nuestra interpretación si *la fiabilidad de las columnas al diferenciar a las filas* hubiera sido alta. Este coeficiente es el mismo coeficiente α de Cronbach; si hubiese sido relativamente alto estaríamos ante un test o escala de *actitud de servicio* o de *gusto* por este tipo de tareas *en general,* pero vemos que no es éste el caso.

Los coeficientes η^2 son coherentes con los resultados vistos (fórmulas [1] y [2] eliminando la raíz cuadrada):

Proporción de varianza en la variable dependiente (respuestas de los sujetos) explicada por diferencias:

entre las *tareas:* $$\eta^2_{columnas} = \frac{SC_{columnas}}{SC_{total}} = \frac{65.086}{163.486} = .398$$

entre los *sujetos:* $$\eta^2_{filas} = \frac{SC_{filas}}{SC_{total}} = \frac{65.086}{163.486} = .114$$

Casi el 40% de la varianza está explicado por diferencias entre las tareas y poco más del 11 % por diferencias entre los sujetos. El coeficiente $\eta^2_{parcial}$ (fórmula [3] sin la raíz cuadrada) es igual a .449: aproximadamente un 45% de la varianza (diferencias en las respuestas) queda explicado por las diferencias entre tareas teniendo en cuenta solamente estas diferencias entre tareas y prescindiendo de los errores de medición (que en este caso equivalen a las peculiaridades individuales manifestadas en las respuestas).

6.3. Contrastes posteriores y representación gráfica

En este caso hemos hecho los contrastes posteriores (de Tuckey, con el SPSS). Las cuatro tareas con medias mayores (*teléfono, cocina, lavadero y pasillo*) no se diferencian significativamente entre sí, pero estas cuatro tareas sí tienen diferencias estadísticamente significativas con las tres tareas con medias más bajas (*sala de estar, sala de estudio y baño*) que entre sí tampoco difieren. Tenemos dos bloques de tareas claramente diferenciados; con más sujetos sí podrían haber aparecido más diferencias significativas.

Una manera de presentar estos resultados es como aparecen en la figura 1: se ponen *por orden* las tareas y sus medias y se subrayan las que no difieren entre sí.

4. Teléfono	7. Cocina	6. Lavadero	5. Pasillo	2. Sala de estar	3. Sala de estudio	1.Baño
4.1	3.7	3.6	3.5	2.0	2.0	1.5

Figura 1

Aunque no se hagan los contrastes posteriores sí es útil presentar alguna figura semejante en la que aparezcan las tareas ordenadas según sus medias. La razón F de las tareas que es estadísticamente significativa y la alta fiabilidad o grado de acuerdo de los sujetos diferenciando unas tareas de otras ya es suficiente para hacer una interpretación de los resultados.

En este ejemplo los sujetos han valorado el gusto por una serie de tareas; es fácil intuir la utilidad de este modelo de análisis de varianza; en vez de tareas podrían ser otras variables de interés pertenecientes al mismo ámbito conceptual para que tenga sentido la comparación o establecer una jerarquía (experiencias, valores, motivaciones, etc.). También puede tratarse

de la misma variable en ocasiones sucesivas o medidas *después* de pasar por experiencias distintas (en el ejemplo expuesto los sujetos se limitan a responder a un cuestionario, no *después* de realizar una determinada tarea). En la introducción ya se han expuesto diversos planteamientos susceptibles de ser analizados con este modelo de análisis de varianza.

7. Análisis de varianza para muestras relacionadas en EXCEL y en el SPSS

EXCEL. Este análisis de varianza lo tenemos en *Herramientas - Análisis de datos - Análisis de varianza de dos factores con una muestra por grupo.*

Nos da solamente los datos descriptivos y la tabla de resultados; no incluye contrastes posteriores, ni los otros coeficientes (fiabilidad, η^2) ya expuestos.

Una cautela. Al señalar los datos (dispuestos en columnas) hay que seleccionar también una columna adicional a la izquierda de la tabla (que se puede intercalar y se deja en blanco) porque esta primera columna se interpreta como *rótulos* (que en las filas no suelen interesar).

SPSS. En el SPSS este modelo de análisis de varianza se encuentra en *analizar* y allí en *modelo lineal general*, en *medidas repetidas.* Este análisis de varianza es más complejo y hay que acudir a manuales específicos (como Pardo Merino y Ruíz Díaz, 2005; Ximénez y San Martín (2000). El SPSS no presenta la *tabla de resultados* convencional por lo que, si interesa presentar esta tabla de resultados, se puede hacer fácilmente con EXCEL. El SPSS calcula los contrastes posteriores pero no los coeficientes de fiabilidad (muy sencillos, como hemos visto, a partir de la información de la tabla de resultados).

8. El análisis de varianza para muestras relacionadas en Internet.

En Internet disponemos al menos de este programa:

LOWRY, RICHARD, VASSARSTATS: Web Site for Statistical Computation,Vassar College, Poughkeepsie, NY, USA; One-Way Analysis of Variance for Independent or Correlated Samples http://faculty.vassar.edu/lowry/anova1u.html;

Está en la misma página en la que está el análisis de varianza para muestras independientes. *Tiene allí mismo una versión en español.* Hay que introducir (o *copiar y pegar*) todos los datos. Una limitación de este programa es que no admite mas de cinco variables o columnas.

En el cuadro de diálogo hay que indicar el *número de muestras* (de *columnas*) y escoger *Correlated samples.* Este programa tiene los contrastes posteriores de Tukey (la *diferencia mínima* necesaria para afirmar que es estadísticamente significativa); en este sentido este programa es preferible a EXCEL, al menos si interesan los contrastes posteriores. No calcula los coeficientes de *fiabilidad,* pero con la fórmula a la vista se calculan muy fácilmente con una calculadora.

9. Referencias bibliográficas

DOWNIE, N.M. y HEATH, R.W., (1971). *Métodos estadísticos aplicados*. Madrid: Ediciones del Castillo

GARCÍA PÉREZ, J.F.; FRÍAS NAVARRO, M.D. y LLOBELL, J. PASCUAL (1999). *Los diseños de la experimentación experimental, comprobación de hipótesis***.** Valencia: C.S.V.

GIRDEN, ELLEN R., (1992). *Anova repeated measures.* Quantitative Applications in the Social Sciences. Newbury Park & London: Sage

GUILFORD, J. P. y FRUCHTER, B., (1984). *Estadística aplicada a la psicología y la educación,* México: McGraw-Hill. [En Inglés: *Fundamental Statistics in Psychology and Education,* 1973. New York: McGraw-Hill].

HOYT, C.J., (1941). Test Reliability Estimated by Analysis of Variance. *Psychometrika,* 3, 153-160.

HOYT, C.J., (1952). Estimation of Test Reliability for Un-Restricted Item Scoring Methods. *Educational and Psychological Measurement,* 12, 752-758.

JACCARD, JAMES (1998). *Interaction Effects in Factorial Analysis of Variance,* Sage University Paper Series on Quantitative Applications in the Social Sciences. Thousand Oaks: Sage

KIRK, ROGER E., (1995). *Experimental Design, Procedures for the Behavioral Sciences.* Boston: Brooks/Cole.

LLOVEL, J. PASCUAL; FRÍAS, DOLORES y GARCÍA, FERNANDO (1996). *Manual de psicología experimental*. Barcelona: Ariel.

MORALES VALLEJO, PEDRO (2008). *Estadística aplicada a las Ciencias Sociales.* Madrid: Universidad Pontificia Comillas.

NUNNALLY, JUM C. and BERNSTEIN, IRA H. (1994). *Psychometric Theory*, 3rd. ed., New York, McGraw-Hill.

PARDO MERINO, A. y RUÍZ DÍAZ, M.A. (2005). *Análisis de datos con SPSS 13 Base*. Madrid: McGraw Hill.

PERNEGER, THOMAS V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal* 1998; 316:1236-1238 http://www.bmj.com/cgi/content/full/316/7139/1236

ROSENTHAL, ROBERT and ROSNOW, RALPH L. (1991). *Essentials of Behavioral Research, Methods and Data Analysis.* Boston: McGraw-Hill.

TOOTHAKER, LARRY E. and MILLER, LISA (1995), *Introductory Statistics for the Behavioral Sciences,* 2nd edit., Pacific Grove, Brooks/Cole, 706pp.

WADDINGTON, CHRISTOPHER (2000). *Estudio comparativo de diferentes métodos de evaluación de traducción general (Inglés-Español).* Madrid: Universidad Pontificia Comillas.

XIMÉNEZ, CARMEN y SAN MARTÍN, RAFAEL (2000). *Análisis de Varianza con medidas repetidas*. Madrid: La Muralla.