anales de psicología
2010, vol. 26, nº 2 (julio), 400-409

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Análisis de medidas repetidas usando métodos de remuestreo

Guillermo Vallejo[1*], M. Paula Fernández[1], Ellián Tuero[1] y Pablo E. Livacic-Rojas[2]

[1] *Universidad de Oviedo (España)*
[2] *Universidad de Santiago (Chile)*

Resumen: Este artículo evalúa la robustez de varios enfoques para analizar diseños de medidas repetidas cuando los supuestos de normalidad y esfericidad multimuetral son separada y conjuntamente violados. Específicamente, el trabajo de los autores compara el desempeño de dos métodos de remuestreo, pruebas de permutación y de bootstrap, con el desempeño del usual modelo de análisis de varianza (ANOVA) y modelo lineal mixto con la solución Kenward-Roger implementada en SAS PROC MIXED. Los autores descubrieron que la prueba de permutación se comportaba mejor que las pruebas restantes cuando se incumplían los supuestos de normalidad y de esfericidad. Por el contrario, cuando se violaban los supuestos de normalidad y de esfericidad multimuestral los resultados pusieron de relieve que la prueba Bootstrap-F proporcionaba un control de las tasas de error superior al ofrecido por la prueba de permutación y por enfoque del modelo mixto. La ejecución del enfoque ANOVA se vio afectada considerablemente por la presencia de heterogeneidad y por la falta de esfericidad, pero escasamente por la ausencia de normalidad.
Palabras clave: Robustez; esfericidad multimuestral; valores críticos teóricos; valores críticos empíricos

Title: Analyzing repeated measures using resampling methods.
Abstract: This article evaluated the robustness of several approaches for analyzing repeated measures designs when the assumptions of normality and multisample sphericity are violated separately and jointly. Specifically, the authors' work compares the performance of two resampling methods, *bootstrapping* and *permutation tests*, with the performance of the usual analysis of variance (ANOVA) model and the mixed linear model procedure adjusted by the Kenward–Roger solution available in SAS PROC MIXED. The authors found that the permutation test outperformed the other three methods when normality and sphericity assumptions did not hold. In contrast, when normality and multisample sphericity assumptions were violated the results clearly revealed that the Bootstrap-F test provided generally better control of Type I error rates than the permutation test and mixed linear model approach. The execution of ANOVA approach was considerably influenced by the presence of heterogeneity and *lack of sphericity*, but scarcely affected by the absence of normality.
Key words: Robustness; multisample sphericity; theoretical critical values; empirical critical values.

Como ha sido puesto de relieve en diversos análisis de contenido metodológico (véase, p.e., Keselman, Lix & Keselman, 1996; Bono Arnau & Vallejo, 2008), los diseños de medidas repetidas juega un papel cada vez más destacado en la investigación actual, particularmente el centrado en comparar las respuestas de *N* participantes distribuidos en *J* grupos, ya sea en función de *K* tratamientos experimentales o de *K* mediciones temporales. Sirva de botón de muestra el trabajo publicado recientemente por Nuevo, Cabrera, Márquez-González y Montorio (2008) en la revista *Anales de Psicología*. Diversas pruebas son aplicables a los datos provenientes de estos estudios. Cuando las varianzas correspondientes a las diferencias entre pares de medidas repetidas son iguales entre sí (supuesto de esfericidad), resulta apropiado utilizar el usual modelo de análisis de la varianza (ANOVA) propuesto por Scheffé. Si la propiedad de esfericidad no se mantiene, pero las matrices de dispersión son homogéneas, se suele optar por utilizar el enfoque ANOVA con los grados de libertad corregidos mediante alguno de los múltiples correctores tipo Box existentes (para detalles véase, p.e., Blanca, 2004 o Fernández, Livacic-Rojas & Vallejo, 2007) o el enfoque del análisis multivariado de la varianza (MANOVA). Sin embargo, ambos enfoques son típicamente inválidos cuando las matrices de dispersión son heterogéneas y/o los datos se desvían de la normalidad, especialmente cuando el tamaño de los grupos no esta convenientemente equilibrado (Algina & Oshima, 1994; Olson, 1974).

* **Dirección para correspondencia [Correspondence address]:** Guillermo Vallejo. Facultad de Psicología. Universidad de Oviedo. Plaza de Benito Feijóo, s/n. 33003 Oviedo (España).
E-mail: gvallejo@uniovi.es.

Para vencer el impacto negativo que la heterogeneidad de las matrices de dispersión ejerce sobre la robustez de los enfoques referidos, se han desarrollado diversas soluciones alternativas, incluyendo la versiones multivariadas de los enfoques Welch-James y Brown-Forsythe desarrolladas por Johansen (1980) y por Vallejo y Ato (2006), respectivamente, y la metodología del modelo lineal mixto (MLM). Las dos primeras usan métodos de estimación basados en el principio de los mínimos cuadrados ordinarios (MCO) y se centran en corrigen los grados de libertad correspondientes a la matriz del error y de la hipótesis, mientras que la tercera utiliza métodos de estimación basados en el principio de máxima verosimilitud y se centra en modelar la matriz de dispersión. Mediante el último enfoque, más que asumir una matriz complemente general o excesivamente simple, se trata de buscar un equilibrio entre los criterios de flexibilidad y parsimonia. Hay que advertir, no obstante, que bajo algunas condiciones de desviación de la normalidad los problemas de inferencia no desaparecen con estas soluciones (Kowalchuk, Keselman, Algina & Wolfinger, 2004; Vallejo y Ato, 2006). Por esta razón, es también es importante apoyarse en soluciones que estén ***destinadas*** a combatir la falta de normalidad de los datos.

En orden a mantener las tasas de error de Tipo I a un nivel aceptable cuando se incumplen los supuestos distribucionales, diversas soluciones se hallan disponibles hoy en día. Entre las alternativas propuestas destacan las cuatro que siguen: (a) las encaminadas a lograr la normalidad de los datos utilizando alguna transformación de la familia Box-Cox; (b) las basadas en utilizar procedimientos no paramétricos (p.e., Akritas, Arnold & Brunner, 1997; Beasley, 2002); (c) las orientadas a sustituir los usuales estimadores de ten-

dencia central y variabilidad por estimadores robustos, tales como medias recortadas y varianzas winsorizadas (Lix, Algina & Keselman, 2003) y; (d) las caracterizadas por emplear la metodología del modelo mixto generalizado, la cual permite especificar explícitamente una estructura de error no normal. Se excusa decir que la viabilidad de estas soluciones no es la misma, ya que unas adolecen de mayores defectos que otras.

Específicamente, el remedio más comúnmente usado para normalizar los datos consiste en transformar la escala de los mismos. Sin embargo, puede ocurrir que como resultado de una transformación monótona (aquella que preserva el ordenamiento de las medias pero no la relativa distancia entre ellas), no sólo se elimine el sesgo de la distribución sino también alguno de los efectos factoriales definidos en el modelo lineal. Por ejemplo, si las medias de los niveles del factor *A* tienen el mismo rango para todos los niveles del factor *B,* entonces una transformación puede remover la interacción *AB* (véase p.e, Quinn & Keough, 2002). También ha sido descubierto que la utilidad de las pruebas no paramétricas basadas en rangos puede ser limitada, sobre todo, cuando se viola el supuesto de homogeneidad de las matrices de covarianza (Kowakchuk, Keselman y Algina, 2003; Lei, Holt y Beasley, 2004). Con respecto a las pruebas que utilizan medias recortadas y matrices de dispersión winsorizadas, conviene tener presente que las hipótesis probadas con los estimadores MCO difieren de las probadas con los estimadores robustos cuando la distribución es asimétrica (Vallejo, Fernández & Livacic, 2009). Por último, aunque se sabe poco de las características operantes del enfoque del modelo mixto generalizado, lo cierto es que requiere especificar la distribución analítica que siguen los datos, lo cual puede ser tan complicado como conocer *a priori* la forma de la matriz de covarianza.

Afortunadamente, diversos trabajos, incluyendo los de Berkovits, Hancock y Nevitt (2000) y Vallejo, Cuesta, Fernández y Herrero (2006), sugieren que el enfoque basado en el remuestreo bootstrap puede constituir una alternativa viable para abordar los problemas referidos. En el contexto específico de los diseños de medidas repetidas sencillos, Berkovits et al. (2000) muestran que el desempeño del método bootstrap-F era generalmente satisfactorio cuando los datos incumplían los supuesto de normalidad y de esfericidad. Resultados similares fueron obtenidos por Vallejo et al. (2006) con diseños de medidas repetidas en ausencia de esfericidad multimuestral. También cabe la opción de generar la distribución de probabilidad empírica de un estadístico de contraste usando las pruebas de permutación introducidas por Fisher (1935). En el contexto de los diseños experimentales con factores no repetidos, se ha encontrado que esta técnica estadística limita el número de errores al valor nominal (Anderson & ter Braak, 2003; Jung, Jhun & Song, 2006).

Hoy en día los términos pruebas de aleatorización y pruebas de permutación suelen usarse indistintamente, sin embargo, aún persiste cierto grado de confusión entre ambos. Algunos autores, incluyendo Fisher (1936) y Kempthorne y Doerfler (1969), emplean el término *pruebas de permutación* en situaciones donde el muestreo aleatorio justifica los cálculos realizados para determinar el grado de significación de un test estadístico en ausencia de asignación aleatoria de las unidades experimentales a los tratamientos y el término *pruebas de aleatorización* en contextos donde dichos cálculos se justifican apoyándose en el diseño experimental. Otros emplean el término permutación para referirse a las pruebas basadas en todos los posibles ordenamientos de los datos *y* el término pruebas de aleatorización lo reservan para el subconjunto elegido al azar de todas las permutaciones posibles (véase, Edgington, 1995; Manly, 2007). En nuestro caso, ambos términos serán usados para denotar la estrategia que implica comparar el valor del estadístico de prueba con la distribución muestral que resulta de aplicar la misma prueba a múltiples ordenamientos de los datos originales.

Aunque las pruebas de permutación requieren menos supuestos que sus homólogas paramétricas, especialmente en lo referido a la forma de la distribución (no se necesita que los datos sigan una distribución analítica conocida), a la extracción aleatoria de la muestra desde la población y a la naturaleza métrica de los datos, esto no significa que no requieran ninguno. Estas pruebas asumen la independencia de los datos, de modo que los distintos reordenamientos que se efectúen a partir de la muestra original deben ser igual de probables y, por ende, intercambiables (Good, 2002). Sin embargo, es improbable que el supuesto de independencia se satisfaga cuando los tratamientos se aplican de forma sistemática. Por ejemplo, una prueba de aleatorización basada reordenar los datos de todas las maneras posibles, no sería apropiada para comparar el promedio de la tasa de cambio de *N* participantes distribuidos al azar en *J* grupos a lo largo de *K* mediciones temporales. ***Sea por la razón que sea,*** cuando no se utilizan secuencias aleatorias de recepción de los tratamientos, se debe respetar la estructura de los datos usando esquemas de aleatorización restringida en lugar de completa (Manly, 2007).

Con lo anterior en mente, el objetivo fundamental del presente trabajo es examinar la robustez de la prueba *F* tradicional cuando los valores críticos se obtienen mediante de permutación estocástica aproximada y remuestreo bootstrap, en vez de calcularlos analíticamente desde la teoría normal. Además, dada la demanda casi exclusiva que los usuarios de los diseños de medidas repetidas hacen del modelo lineal general (MLG), y de un tiempo a esta parte del MLM, con fines comparativos también examinaremos el comportamiento de estos enfoques. La robustez de los procedimientos reseñados será estudiada usando un diseño de medidas parcialmente repetidas no equilibrado carente de homogeneidad y/o de normalidad. Hasta la fecha, no se ha evaluado conjuntamente el desempeño de las técnicas analíticas reseñadas.

Definición de los procedimientos estadísticos usados en la investigación

El diseño discutido en este trabajo postula dos factores de tratamiento manipulables experimentalmente, pero con la condición que a cada sujeto le sean administrados todos los niveles del factor *B* en combinación con un solo nivel del factor *A*. Los efectos del diseño se analizan mediante las cuatro técnicas que se describen debajo.

Modelo lineal general (MLG)

Para el caso de contar con *K* respuestas de *N* (= n_1 + n_2+...+ n_J) unidades experimentales distribuidas al azar en *J* grupos con n_j unidades por grupo, el modelo puede ser escrito

$$y_{ijk} = \mu + \alpha_j + s_{ij} + \beta_k + (\alpha\beta)_{jk} + e_{ijk}, \quad (i = 1,...,n_j; j = 1,...,J; k = 1,...,K) \qquad (1)$$

donde y_{ijk} es el valor de la respuesta dada por el *i*-ésima unidad asignada al *j*-ésimo nivel del factor *A* en el *k*-ésimo nivel del factor *B*, μ es una constante común para todas las observaciones, α_j es el efecto del *j*-ésimo nivel del factor *A*, s_{ij} es el efecto aleatorio asociado con la *i*-ésima unidad dentro del *j*-ésimo nivel del factor *A*, β_k es el efecto del *k*-ésimo nivel del factor *B*, $(\alpha\beta)_{jk}$ es el efecto de la interacción entre el *j*-ésimo nivel del factor *A* y el *k*-ésimo nivel del factor *B* y e_{ijk} es el error aleatorio asociado con la *i*-ésima unidad asignado al *j*-ésimo nivel del factor *A* en el *k*-ésimo nivel del factor *B*. Asumiendo que los efectos aleatorios del modelo son normal e independientemente distribuidos, $s_{ij} \sim N(0,\sigma_s^2)$ y $e_{ijk} \sim N(0,\sigma_e^2)$, se tiene que la varianza de una observación es $\sigma_s^2 + \sigma_e^2$, mientras que la covarianza entre dos observaciones del mismo sujeto en diferentes tiempos es σ_s^2. Lo anterior presupone que tanto las varianzas como las covarianzas son constantes.

Bajo la hipótesis de normalidad de los términos de error, los efectos diferenciales de *A, B* y *AB* son medidos mediante las razones *F* que resultan de aplicar las usuales definiciones de las medias cuadráticas, las cuales son, respectivamente

$$F_A = \frac{MC_A}{MC[S(A)]}, \quad F_B = \frac{MC_B}{MC[BS(A)]}, \quad F_{AB} = \frac{MC_{AB}}{MC[BS(A)]}. \qquad (2)$$

Para que los estadísticos definidos en (2) se distribuyan según la distribución *F* de Fisher cuando la H_0 es verdadera $(H_0^A: \alpha_j = 0,\ H_0^B: \beta_k = 0$ y $H_0^{AB}: (\alpha\beta)_{jk} = 0)$, se requiere que las matrices de dispersión sean homogéneas y que satisfagan la condición de esfericidad.

Modelo lineal mixto (MLM)

El modelo de efectos mixtos extiende el modelo clásico a situaciones donde los supuestos de homogeneidad e independencia no son requeridos, y a situaciones donde las variables son tanto fijas como aleatorias. El modelo mixto propuesto por Laird y Ware (1982) para analizar datos longitudinales puede ser escrito en forma matricial como

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \qquad (3)$$

donde $\mathbf{y}_i$ un vector $(k_i \times 1)$ respuesta para las medidas repetidas $1, 2,...,k$ del *i*-ésimo participante perteneciente al *j*-ésimo grupo (si no existen datos faltantes, entonces $k_i = k$), $\mathbf{X}_i$ es una matriz $(k_i \times p)$ de diseño conocida que puede incorporar tanto efectos de tratamiento como tendencias temporales, $\boldsymbol{\beta}$ es un vector $(p \times 1)$ de parámetros desconocidos de efectos fijos, $\mathbf{Z}_i$ es una matriz $(k_i \times q)$ de diseño conocida para los efectos aleatorios, $\mathbf{b}_i$ es un vector $(q \times 1)$ de efectos aleatorios desconocidos específico para cada participante y $\mathbf{e}_i$ es un vector $(k_i \times 1)$ de parámetros desconocidos cuyos elementos no necesitan ser homogéneos ni tampoco independientes. Los supuestos distribucionales acerca de los vectores aleatorios del modelo (3) implican que $\mathbf{b}_i$ y $\mathbf{e}_i$ son mutuamente independientes y que están distribuidos, respectivamente, como $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$ y $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{R}_i)$. Aquí $\mathbf{G}$ es una matriz $(q \times q)$ de covarianza para los efectos aleatorios específicos del participante incluidos en el modelo y $\mathbf{R}_i$ es una matriz $(k_i \times k_i)$ de covarianza para los errores dentro de los participantes. Como señalan Gurka y Edwards (2008), estos supuestos implican que, marginalmente, el vector respuesta $\mathbf{y}_i$ se distribuye normal e independientemente con media $\mathbf{X}_i\boldsymbol{\beta}$ y matriz de covarianza $\mathbf{V}_i(\boldsymbol{\theta}) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. Las matrices $\mathbf{G}$ y $\mathbf{R}_i$ son convenientemente caracterizadas por los componentes de varianza (CV) únicos contenidos en el vector $(r \times 1)$ $\boldsymbol{\theta}$, es decir, por todos los parámetros desconocidos de la matriz de covarianza marginal del vector respuesta $\mathbf{y}_i$.

Los CV constituyen la piedra angular sobre la que se sustenta la metodología del modelo mixto, tanto en lo referido a la estimación e inferencia acerca de los parámetros de efec-

tos fijos como de los efectos aleatorios. Si $\mathbf{V}_i(\boldsymbol{\theta})$ es conocida, entonces los estimadores estándar para $\boldsymbol{\beta}$ y $\mathbf{b}_i$ son el estimador de mínimos cuadrados generalizados (GLS)

$$\hat{\boldsymbol{\beta}}=\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{y}_i \quad (4)$$

y el predictor $\hat{\mathbf{b}}_i=\mathbf{G}\mathbf{Z}_i'\mathbf{V}_i(\boldsymbol{\theta})^{-1}(\mathbf{y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}})$, donde N es el número de participantes. En la practica, los componentes de la matriz $\mathbf{V}_i(\boldsymbol{\theta})$ son desconocidos y se debe proceder a su estimación usando todos los datos disponibles. Existen múltiples métodos de estimación (véase, p.e., Searle, Casella y McCulloch, 1992 ó Vonesh y Chinchilli, 1997), pero los que ofrecen las propiedades más deseables en términos de robustez, eficiencia asintótica y consistencia (Robinson, 1991), son los basados en maximizar la verosimilitud de la muestra como una función de los parámetros del modelo vía FML o REML.

Después de que la matriz de covarianza ha sido satisfactoriamente modelada, se está en condiciones de proceder a realizar las inferencias correspondientes a los efectos del modelo. Aunque en ocasiones puede resultar de interés comprobar si existen diferencias entre los perfiles individuales y los promedios grupales, en la práctica, lo verdaderamente prioritario es obtener inferencias válidas acerca de los efectos fijos del modelo usando pruebas tipo Wald y contrastes de razón de verosimilitud (LRT).

La prueba estadística más usada para contrastar hipótesis de la forma H_0: $\mathbf{C}'\boldsymbol{\beta}=\mathbf{0}$ versus H_A: $\mathbf{C}'\boldsymbol{\beta}\neq\mathbf{0}$, es el estadístico F_W de Wald que sigue

$$F_W=\upsilon_1^{-1}(\mathbf{C}'\tilde{\boldsymbol{\beta}})'[\mathbf{C}'\,V(\tilde{\boldsymbol{\beta}})\mathbf{C}]^{-1}(\mathbf{C}'\tilde{\boldsymbol{\beta}}), \quad (5)$$

donde ν_1 es el rango (***número de filas linealmente independientes***) de la matriz de contrastes $\mathbf{C}$, $\tilde{\boldsymbol{\beta}}$ es el estimador GLS empírico del vector $\boldsymbol{\beta}$ y $V(\tilde{\boldsymbol{\beta}})=[\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{V}}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i^{-1}]^{-1}$ es la matriz de covarianza estimada. Bajo hipótesis nula, el estadístico F_W de Wald se distribuye aproximadamente como una F con ν_1 grados de libertad en el numerador y ν_2 en el denominador. Por regla general, el valor de ν_2 hay que computarlo desde los datos usando algún método tipo Satterthwaite. Cuando el tamaño de muestra sea reducido, la inferencias basadas en la verosimilitud deben ser interpretadas con suma cautela, pues $V(\tilde{\boldsymbol{\beta}})$ subestima la variabilidad muestral de $\tilde{\boldsymbol{\beta}}$ (Wolfinger, 1996).

Para corregir los efectos del sesgo en la estimación asintótica de los errores estándar, Kenward y Roger (KR; 1997) proponen modificar ν_2, inflar la matriz de covarianza estimada ($V(\tilde{\boldsymbol{\beta}})$) y ajustar el estadístico definido en (5). Recientemente, Kenward y Roger (2009) han corregido la fórmula utilizada en su trabajo original para inflar la matriz de covarianza. El módulo *Proc Glimmix* implementado en la última versión del programa SAS (SAS® 9.2 Institute Inc., 2008), permite una aproximación intuitiva a dicha mejora añadiendo la subopción FIRSTORDER en la opción DDFM=KR. Otra solución alternativa para calcular ν_2 consiste en usar una aproximación Satterthwaite generalizada dada por Fai y Cornelius (FC; 1996). Aunque ambas soluciones se hallan implementadas en el módulo *Proc Mixed*, los trabajos de Schaalje, McBride y Fellingham (2002), Livacic-Rojas, Vallejo y Fernández (2006) y Arnau, Bono y Vallejo (2009) muestran la relativa superioridad de la solución KR sobre la solución FC. Con todo, salvo que se conozca el verdadero proceso generador de los datos, el desempeño de *Proc Mixed* con la solución KR basado en los criterios información depende en exceso del tamaño de muestra, de la complejidad de la matriz de covarianza y de la forma de la distribución (Kowalchuk, Keselman, Algina & Wolfinger, 2004; Gómez, Schaalje y Fellingham, 2005; Vallejo y Livacic-Rojas, 2005; Vallejo y Ato, 2006; Vallejo, Ato y Valdés, 2008).

Prueba de aleatorización

Para determinar el grado de significación estadística de una hipótesis mediante esta técnica de análisis es preciso, definir con claridad las hipótesis a contrastar, seleccionar la prueba estadística que refleje adecuadamente la diferencia entre los datos observados y la situación nula, elegir el esquema de permutación que genere la distribución de probabilidad del test de acuerdo a lo sucedido en el estudio y concretar el tipo de datos usados para realizan las permutaciones. A continuación, se resumen los pasos a seguir:

1. Se especifica el patrón de aleatorización adoptado en el experimento para asignar los tratamientos a las unidades de muestreo. En nuestro caso, n_j unidades diferentes fueron asignadas al azar a cada uno de los J niveles del factor A. Después, dentro de cada nivel del factor A, los K niveles del factor repetido B se asignaron al azar a cada una de las unidades. Las acciones referidas son críticas, pues el esquema de permutación utilizado para generar distribución probabilidad empírica se debe corresponder con el patrón seguido en la planificación del estudio. El esquema de permutación usado será uno u otro, dependiendo de si la naturaleza del factor B es activa o asignada. Las respuestas se asumen intercambiables si el factor B es experimental y dependientes cuando no lo es. Adviértase que bajo los esquemas de aleatorización comple-

ta y restringida resultan $[N!/(n_1!n_2!...n_J!)](K!)^N$ y $[N!/(n_1!n_2!...n_J!)](1!)^N$ permutaciones, respectivamente.

2. Se lleva a cabo un muestro aleatorio sin reposición desde la población de permutaciones obtenida tras aplicar la fórmula correspondiente al esquema de aleatorización completa y se eligen *M* matrices independientes de dimensión *N*×*K*, $\mathbf{Y}^*(1),...,\mathbf{Y}^*(M)$. En vista de la ingente cantidad de cálculo que implicaría trabajar con cada uno de los posibles ordenamientos de los datos originales, las pruebas de aleatorización están basadas en *M* replicaciones. Para evitar que se puedan obtener valores *p* diferentes usando los mismos datos, algunos autores (Crowley, 1992; Manly, 2007) recomiendan seleccionar un número de permutaciones que no sea inferior a 1000.
3. Se especifica la prueba estadística que se va a utilizar y se obtiene su valor para los *M* conjuntos de datos permutados. En concreto, para comprobar que los posibles efectos diferenciales de *A*, *B* y *AB* no están determinados por los tratamientos y, por ende, sería razonable que ocurriesen bajo cualquier reordenamiento de los datos originales, se halla *M* veces $F_c^{*m}(c=A,B,o\,AB)$, donde F_c es el valor del estadístico definido en (2).
4. Se determina la excepcionalidad del estadístico F_c utilizando los valores F_c^{*m} de las *M* permutaciones. De acuerdo con Efron y Tibshiriani (1993), el nivel de significación alcanzado por el test F_c se calcula mediante la expresión: $p_{alea}=M^{-1}\sum_{m=1}^{M} I[F_c^* > F_c]$, donde $I[F_c^* > F_c]$, la usual función indicador, vale 1 si $F_c^* > F_c$ y 0 si es menor. La proporción de valores F_c^* que superan al valor F_c representa el valor *p*.

Enfoque bootstrap-F

Los pasos seguidos para determinar el grado de significación de las hipótesis del diseño mediante el enfoque bootstrap-F son los siguientes:

1. Se define la unidad de muestreo a utilizar y se desplazan las distribuciones empíricas de manera que la hipótesis nula sea verdadera. En nuestro caso, dicha unidad será los vectores de errores asociados con los valores observados para cada una de las poblaciones de muestreo existentes. Esto supone que los datos obtenidos por esta vía y los datos originales tienen la misma forma y dispersión, pero diferente localización. La operación de centrado garantiza que en cada uno de los *J* niveles del factor *A*, las medias de los *K* niveles del factor *B* no difieran entre sí.
2. Se genera una muestra bootstrap, $\mathbf{C}_j^* = (\mathbf{C}_{1j}^*,...,\mathbf{C}_{n_j j}^*)$, remuestreando aleatoriamente con reposición n_j filas a partir de la matriz de datos centrados $\mathbf{C}_j = (\mathbf{C}_{1j}',...,\mathbf{C}_{n_j j}')$, $\mathbf{C}_{ij} = (C_{ij1},...,C_{ijK})'$, donde $C_{ijk} = Y_{ijk} - \bar{Y}_{jk}$ e $\bar{Y}_{jk} = \sum_{i=1} Y_{ijk}/n_j$. Debido a la existencia de matrices de dispersión heterogéneas y a la falta de equilibrio que presentan los diferentes grupos que configuran el diseño, el procedimiento de remuestreo se efectúa en cada uno de los *J* niveles del factor *A*. De este modo, cada muestra bootstrap es obtenida desde una distribución para la cual las hipótesis de nulidad son verdaderas.
3. Se calcula $F_c^*(c=A,B,o\,AB)$, el valor del estadístico de contraste F_c obtenido a partir de una muestra bootstrap $\mathbf{C}_j^*$ definido en (2).
4. Se genera la distribución bootstrap repitiendo B veces los pasos 2-3. De acuerdo con Wilcox (2001), el valor de B puede ser convenientemente fijado en 599 remuestreos. Hall (1986) proporciona la justificación teórica para dicha elección.
5. Se determina la excepcionalidad del estadístico F_c utilizando los valores F_c^{*b} de las B remuestras. Siguiendo el trabajo de Efron y Tibshiriani (1993), el nivel de significación alcanzado por el test F_c se calcula mediante la expresión: $p_{boot}=B^{-1}\sum_{b=1}^{B} I[F_c^* > F_c]$, donde $I[F_c^* > F_c]$, la usual función indicador, vale 1 si $F_c^* > F_c$ y 0 si es menor. La proporción de valores F_c^* que superan a F_c representa el valor *p* bootstrap.

Por consiguiente, bajo los últimos métodos, a diferencia de lo que sucede con los enfoques MLG y MLM, los valores críticos derivados desde la teoría normal son innecesarios.

Método de la simulación

En orden a evaluar la robustez de los enfoques definidos en el apartado anterior cuando los valores críticos se obtienen mediante valores teóricos y mediante técnicas de computación intensiva, llevamos a cabo un estudio de simulación usando un diseño de medidas parcialmente repetidas no equilibrado carente de homogeneidad y/o de normalidad con *J* = 3 y *K*= 4. Para ello fueron manipuladas las cinco variables siguientes:

1. *Tamaño de muestra total*. El desempeño fue investigado usando dos tamaños de muestra distintos: *n* = 30 y *n* = 45. Estos tamaños grupales son representativos de los encontrados frecuentemente en las investigaciones psicológicas. Dentro de cada tamaño de muestra, el valor del coeficiente de variación muestral (Δ) se fijó en 0.33, donde $\Delta = \frac{1}{\bar{n}}[\Sigma_j (n_j - \bar{n})^2 / J]^{1/2}$, siendo $\bar{n}$ el tamaño promedio de los grupos. Cuando el diseño está equilibrado, $\Delta = 0$. Para *n* = 30, los tamaños grupales fueron:

6, 10 y 14; mientras que para n = 45, los tamaños grupales fueron: 9, 15 y 21

2. *Patrones de covarianza empleados para generar los datos.* Los patrones utilizados para generar los datos fueron dos, a saber: una matriz de simetría compuesta (CS) y otra no estruturada (UN). El primer patrón se caracteriza por tener idénticas varianzas y covarianzas, es decir, se asume que la correlación entre cualquier par de medidas repetidas es la misma, independientemente de la distancia existente entre ellas. Se trata de una estructura bastante parca, ya que sólo requiere estimar dos parámetros. El segundo patrón representa la estructura de covarianza que mejor se ajusta a los datos, además no exige que las observaciones se encuentren igualmente espaciadas. No obstante, requiere estimar $K(K+1)/2$ parámetros. En nuestro trabajo manipulamos una matriz UN cuya desviación del patrón de esfericidad se puede tildar de moderado tirando a severo ($\varepsilon = .50$). El valor de ε oscila entre 1 y 1/(*k*-1), donde 1 corresponde a una matriz de covarianza esférica.
3. *Igualdad de las matrices de dispersión.* El desempeño de los procedimientos fue evaluado cuando las matrices de covarianza grupales eran homogéneas y también cuando eran heterogéneas. En el primer caso, los elementos de las tres matrices de dispersión eran iguales entre sí, mientras que en el segundo caso los elementos de las matrices mantenían entre sí las relaciones siguientes: 1: 3: 5 y 1: 5: 9.
4. *Emparejamiento de las matrices de covarianza y el tamaño de los grupos.* La forma de relacionar el tamaño de los grupos y el tamaño de las matrices de dispersión pueden tener diferentes efectos en las pruebas estadísticas. Cuando el diseño está equilibrado, la relación entre el tamaño de las matrices de dispersión y el tamaño de los grupos es nula. Cuando el diseño está desequilibrado, la relación puede ser positiva o negativa. Una relación positiva implica que el grupo de menor tamaño se asocia con la matriz de dispersión menor, mientras que una relación negativa implica que el grupo de menor tamaño se asocia con la matriz de dispersión mayor.
5. *Forma de la distribución de la variable de medida.* Aunque los enfoques MLG y MLM están basados en el cumplimiento del supuesto de normalidad, cuando se trabaja con datos reales es común que los índices de asimetría (γ_1) y curtosis (γ_2) se desvíen de cero (Micceri, 1989), lo cual puede inducirnos a interpretar incorrectamente los resultados. Para investigar el efecto que ejerce la forma de la distribución en el desempeño de las técnicas analíticas, generamos datos desde distribuciones normales y no normales mediante las distribuciones *g* y *h* introducidas por Tukey (1977). Además de la distribución normal $(g = h = 0; \gamma_1 = \gamma_2 = 0)$, también investigamos otras tres: (a) g = 0 y h = .109, una distribución que tiene el mismo grado de sesgo y de curtosis que la exponencial doble o de Laplace $(\gamma_1 = 0 \ \& \ \gamma_2 = 3)$; (b) g = .76 y h = -.098, una distribución que tiene el mismo grado de sesgo y de curtosis que la exponencial $(\gamma_1 = 2 \ \& \ \gamma_2 = 6)$; y (c) g = 1 y h = 0, una distribución que tiene el mismo grado de sesgo y de curtosis que la distribución lognormal $(\gamma_1 = 6.18 \ \& \ \gamma_2 = 110.94)$. Mediante función RANNOR del SAS generamos variables aleatorias normales estándar (Z_{ijk}) y transformamos cada una de ellas como $Z^*_{ijk} = g^{-1}[\exp(gZ_{ijk}) - 1]\exp(hZ^2_{ijk}/2)$, donde g y h son números reales que controlan el grado sesgo y de curtosis. Para obtener una distribución con desviación estándar σ_{jk}, cada una de las puntuaciones que conforman la variable dependiente fue creada utilizando el modelo lineal $Y_{ijk} = \sigma_{jk} \times (Z^*_{ijk} - \mu_{gh})$, donde $\mu_{gh} = \{\exp[g^2/(2-2h)] - 1\}/[g(1-h)^{1/2}]$ es la media de la de la distribución g y h (para detalles véase Kowalchuk y Headrick, 2009).

Resultados del estudio numérico

El procedimiento más directo para decidir si un determinado enfoque es o no robusto consiste en identificar todas aquellas tasas que excedan significativamente el valor nominal de alfa (α) en más/menos dos errores estándar. No obstante, utilizamos el criterio liberal de Bradley (1978) para facilitar la comparación entre nuestros resultados y los obtenidos por otros investigadores en estudios similares. De acuerdo con este criterio, aquellas pruebas cuya tasa de error empírica $(\hat{\alpha})$ se encuentre en el intervalo $.5\alpha \le \hat{\alpha} \le 1.5\alpha$, serán consideradas robustas. Por consiguiente, para el nivel de significación nominal usado en esta investigación $(\alpha = 5\%)$, el intervalo utilizado para definir la robustez de las pruebas fue $2.5 \le \hat{\alpha} \le 7.5$ Se excusa decir que de haber utilizado otros criterios, diferentes interpretaciones de los resultados son posibles.

La Tabla 1 contiene las tasas de error empíricas correspondientes a la interacción entre los grupos y tratamientos cuando las matrices de dispersión eran homogéneas. Básicamente, el patrón de resultados hallado para los efectos principales entre y dentro de los grupos era cualitativa y cuantitativamente similar al descrito. Por consiguiente, en orden a evitar redundancias tan sólo nos centraremos en la fuente que usualmente acapara el interés de los investigadores en este tipo de estudios. Globalmente, los resultados de la Tabla 1 indican lo siguiente:

1. Cuando los datos fueron generados usando un modelo excesivamente parco (i.e., matriz SC), tanto los enfoques basados en derivar los valores críticos analíticamente a partir de la teoría normal (MLG y MLM), como los centrados en obtenerlos numéricamente mediante permutación y remuestreo (prueba de aleatorización y bootstrap-F) controlaban adecuadamente las tasas de error Tipo I.

Aunque, enfoque bootstrap-F tendía a producir tasas de error inferiores al nivel de significación nominal bajo las distribuciones asimétricas (i.e., exponencial y lognormal).

2. Cuando los datos fueron generados usando un modelo completamente general (i.e., matriz UN), el patrón de resultados hallado con los enfoques MLM, de aleatorización y bootstrap-F fue muy similar al obtenido bajo la condición de esfericidad. Sin embargo, el enfoque MLG evidenció un comportamiento excesivamente liberal.

Observando detenidamente la Tabla 1, se aprecia que el patrón de resultados obtenido con el enfoque MLG no se ve afectado por la forma distribución. Como cabía esperar, la variable crítica resultó ser la ausencia de esfericidad. Por consiguiente, cuando las matrices de dispersión son homogéneas, resulta factible corregir la liberalidad del enfoque MLG ajustando los grados de libertad mediante alguno de los múltiples correctores tipo Box existentes. 3

Tabla 1. Porcentaje de veces que los procedimientos rechazaban la hipótesis nula referida a la interacción *AB* cuando las matrices de dispersión eran homogéneas (NS = 5%).

			Matriz CS (ε = 1.00)				Matriz UN (ε = 0.50)			
N	n_j/Σ_j	GH	GLM	Mixto	Aleat	Bootst	GLM	Mixto	Aleat	Bootst
Distribución Normal ($\gamma_1 = 0$ & $\gamma_2 = 0$)										
30	=	1:1:1	5.28	5.96	5.42	3.22	**10.12**	4.96	5.02	4.34
30	≠	1:1:1	5.36	5.74	5.32	3.88	**9.80**	5.86	5.00	5.32
45	=	1:1:1	4.92	5.94	4.88	4.06	**10.00**	5.90	5.00	4.44
45	≠	1:1:1	4.94	5.16	5.08	4.26	**9.30**	4.88	4.98	5.10
Distribución tipo Laplace ($\gamma_1 = 0$ & $\gamma_2 = 3$)										
30	=	1:1:1	4.59	5.60	4.78	2.66	**8.48**	5.86	4.62	4.12
30	≠	1:1:1	5.00	6.62	4.96	3.34	**8.94**	5.14	4.82	4.36
45	=	1:1:1	5.16	5.14	5.32	3.64	**9.54**	4.66	5.00	4.86
45	≠	1:1:1	5.10	5.06	5.00	3.96	**9.56**	5.30	4.84	4.58
Distribución tipo Exponencial ($\gamma_1 = 2$ & $\gamma_2 = 6$)										
30	=	1:1:1	3.82	4.54	4.40	**1.54**	**9.34**	3.82	5.00	3.00
30	≠	1:1:1	4.76	4.58	5.04	**1.76**	**8.80**	4.32	4.78	3.42
45	=	1:1:1	4.52	4.90	4.86	**2.18**	**9.74**	3.96	5.04	3.58
45	≠	1:1:1	4.98	5.60	4.86	**2.12**	**9.58**	4.64	4.90	3.40
Distribución tipo Lognormal ($\gamma_1 = 6.18$ & $\gamma_2 = 114.94$)										
30	=	1:1:1	4.06	3.66	4.74	**1.16**	**8.88**	3.70	4.98	**2.20**
30	≠	1:1:1	4.94	4.34	4.66	**1.16**	**9.16**	4.16	5.02	2.50
45	=	1:1:1	3.98	4.32	4.42	**1.52**	**9.54**	3.96	5.02	2.74
45	≠	1:1:1	5.10	4.54	4.86	**1.78**	**9.18**	4.58	5.00	2.94

Leyenda: NS = nivel de significación; ε = índice de ausencia de esfericidad; CS = matriz de simetría compuesta; UN = matriz no estructurada; N = tamaño de muestra; n_j/Σ_j = relación entre tamaño de los grupos y el tamaño de las matrices de covarianza; GH = grado de heterogeneidad; GLM = modelo lineal general; Mixto = modelo lineal mixto; Aleat = Prueba de aleatorización; Bootst = enfoque Bootstrap-F; γ_1 = índice de asimetría; γ_2 = índice de curtosis.

Por otra parte, la Tabla 2 contiene las tasas de error empíricas correspondientes a la interacción entre los grupos y las ocasiones de medida cuando las matrices de dispersión eran heterogéneas. Al igual que en el caso anterior, el patrón de resultados correspondiente a los efectos principales no aparece recogido por ser muy similar al descrito a continuación. Globalmente, los resultados de la Tabla 2 indican lo siguiente:

1. Cuando los datos fueron generados a partir de una matriz SC, sólo la prueba de aleatorización limitaba el número de errores al valor nominal establecido bajo todas las condiciones manipuladas. El enfoque MLG evidenció la mayor sensibilidad a la violación del supuesto de homogeneidad. Con independencia de la forma de la distribución, el comportamiento de este enfoque era conservador cuando la relación entre el tamaño de los grupos y el tamaño de las matrices de covarianza era positiva, liberal cuando era nula y sustancialmente liberal cuando era negativa. Las tasas de error obtenidas con el enfoque MLM se encontraban dentro de las bandas de robustez de Bradley cuando la distribución era simétrica, pero se volvían liberales cuando era asimétrica y el tipo de apareamiento negativo. El enfoque bootstrap-F, por su parte, se volvía conservador conforme se incrementa el sesgo y la curtosis de la distribución, prescindiendo del tipo de apareamiento manipulado.
2. Cuando los datos fueron generados a partir del modelo UN (ε = 0.50), sólo el enfoque bootstrap-F mantuvo controladas las tasas de error al nivel nominal establecido bajo todas las condiciones manipuladas. El tamaño de las tasas de error obtenido con la prueba de aleatorización duplicaba el permitido. Aunque no aparece recogido en la Tabla 2, cuando los datos fueron generados a partir del modelo UN con ε = 0.75, las tasas de error empíricas obtenidas con esta prueba excedían el valor nominal en aproximadamente 1.35 veces. También duplicaban su valor nominal las tasas de error obtenidas con el enfoque MLG cuando el apareamiento era nulo, mientras que le llegaban a quintuplicar cuando el apareamiento era negativo. Cuando el tipo de apareamiento era positivo, las ta-

sas de error obtenidas con este enfoque se encontraban dentro de las bandas de robustez de Bradley. El enfoque MLM, por su parte, evidenció un comportamiento similar al descrito bajo el modelo SC.

Tabla 2: Porcentaje de veces que los procedimientos rechazaban la hipótesis nula referida a la interacción AB cuando las matrices de dispersión eran heterogéneas (NS=5%).

			Matriz CS (ε = 1.00)				Matriz UN (ε = 0.50)			
N	$n_j/\boldsymbol{\Sigma}_j$	GH	GLM	Mixto	Aleat	Bootst	GLM	Mixto	Aleat	Bootst
Distribución Normal ($\gamma_1 = 0$ & $\gamma_2 = 0$)										
30	≠	1:3:5	5.60	5.70	4.46	3.68	**10.62**	6.16	**9.84**	4.64
30	≠	1:5:9	**7.72**	5.52	5.70	3.42	**11.32**	6.40	**9.54**	4.72
30	+	1:3:5	**2.12**	5.16	5.04	3.52	5.50	4.80	**9.76**	4.18
30	+	1:5:9	**1.36**	5.34	4.98	3.76	4.52	5.44	**8.99**	4.20
30	−	1:3:5	**16.00**	**8.96**	5.44	4.34	**18.54**	**9.64**	**9.58**	5.78
30	−	1:5:9	**22.34**	**7.90**	4.36	4.10	**24.44**	**9.60**	**11.22**	6.20
45	≠	1:3:5	**9.88**	5.76	5.00	4.22	**10.18**	5.96	**9.36**	4.92
45	≠	1:5:9	**13.96**	4.74	4.88	3.94	**10.28**	5.30	**8.78**	4.28
45	+	1:3:5	**1.82**	6.18	4.81	4.20	5.72	5.74	**9.62**	4.94
45	+	1:5:9	**1.66**	6.30	4.92	4.44	4.64	4.92	**8.84**	5.22
45	−	1:3:5	**16.64**	5.70	5.04	4.32	**18.48**	7.16	**9.66**	4.64
30	−	1:5:9	**21.96**	5.32	4.52	3.86	**22.32**	5.84	**8.84**	5.56
Distribución tipo Laplace ($\gamma_1 = 0$ & $\gamma_2 = 3$)										
30	≠	1:3:5	5.78	4.72	4.98	**2.30**	**9.72**	6.34	**8.90**	4.40
30	≠	1:5:9	**7.66**	4.76	5.48	2.90	**10.00**	5.70	**8.86**	4.64
30	+	1:3:5	**2.06**	5.08	5.06	3.14	5.06	4.74	**8.78**	4.42
30	+	1:5:9	**1.34**	5.40	5.48	3.22	3.86	4.42	**8.20**	4.60
30	−	1:3:5	**15.22**	7.02	5.04	3.36	**18.24**	**8.22**	**10.00**	5.08
30	−	1:5:9	**21.19**	6.34	5.00	3.77	**23.19**	**7.89**	**9.42**	6.20
45	≠	1:3:5	**9.56**	5.42	5.06	3.00	**9.58**	5.86	**8.94**	4.08
45	≠	1:5:9	**13.06**	4.86	4.76	3.54	**10.06**	5.67	**8.68**	4.58
45	+	1:3:5	**1.84**	5.42	4.96	3.49	5.07	4.72	**9.14**	4.72
45	+	1:5:9	**1.54**	6.34	4.80	3.70	4.66	5.84	**8.58**	4.72
45	−	1:3:5	**15.74**	5.46	5.22	3.82	**17.84**	6.78	**8.86**	5.16
45	−	1:5:9	**22.58**	5.14	5.08	3.54	**22.50**	6.04	**9.00**	5.82
Distribución tipo Exponencial ($\gamma_1 = 2$ & $\gamma_2 = 6$)										
30	≠	1:3:5	5.88	5.56	5.36	**1.74**	**10.62**	6.72	**10.00**	3.32
30	≠	1:5:9	7.04	6.88	6.24	**2.08**	**10.30**	7.48	**9.86**	4.16
30	+	1:3:5	**1.82**	4.46	4.74	**2.00**	5.42	4.06	**9.90**	3.46
30	+	1:5:9	**1.70**	4.94	5.02	**2.14**	4.74	4.80	**9.12**	3.34
30	−	1:3:5	**13.92**	**9.06**	5.00	**2.14**	**19.50**	**9.62**	**10.06**	3.88
30	−	1:5:9	**20.34**	**9.98**	5.34	**2.90**	**22.70**	**9.60**	**11.10**	5.20
45	≠	1:3:5	**8.16**	5.96	5.00	**2.26**	**9.86**	5.94	**10.02**	3.52
45	≠	1:5:9	**12.56**	7.10	5.78	**2.12**	**11.24**	5.30	**9.76**	4.12
45	+	1:3:5	**2.00**	5.20	5.16	**2.64**	5.52	5.74	**10.04**	3.84
45	+	1:5:9	**1.56**	6.04	5.32	**2.30**	4.78	4.96	**9.42**	3.42
45	−	1:3:5	**13.38**	**8.82**	4.94	**2.24**	**18.22**	7.14	**10.50**	4.20
45	−	1:5:9	**20.98**	**10.06**	5.20	**2.66**	**22.26**	5.86	**10.08**	4.80
Distribución tipo Lognormal ($\gamma_1 = 6.18$ & $\gamma_2 = 114.94$)										
30	≠	1:3:5	4.80	6.86	5.08	**1.12**	**10.32**	**8.90**	**10.07**	2.92
30	≠	1:5:9	6.54	**8.74**	6.50	**1.96**	**11.60**	**10.42**	**11.29**	3.80
30	+	1:3:5	2.58	4.30	5.82	**1.42**	5.94	4.20	**10.03**	2.90
30	+	1:5:9	**1.92**	5.34	5.14	**1.44**	5.36	6.54	**10.01**	3.24
30	−	1:3:5	**13.06**	**9.38**	5.52	**2.10**	**19.00**	**11.60**	**11.05**	3.90
30	−	1:5:9	**18.40**	**14.84**	5.64	**2.28**	**23.08**	**15.90**	**12.03**	5.16
45	≠	1:3:5	**7.62**	**8.22**	5.46	**1.84**	**10.22**	**8.12**	**10.06**	3.28
45	≠	1:5:9	**11.64**	**9.14**	6.00	**2.26**	**11.82**	**10.50**	**10.08**	4.74
45	+	1:3:5	**2.10**	4.90	4.92	**1.46**	5.74	4.78	**10.04**	3.48
45	+	1:5:9	**1.76**	5.68	5.04	**2.00**	6.00	5.86	**9.99**	3.72
45	−	1:3:5	**13.32**	**10.46**	5.62	**2.22**	**18.26**	**12.42**	**11.00**	4.20
45	−	1:5:9	**19.00**	**14.10**	6.04	**2.68**	**23.70**	**15.38**	**11.89**	4.66

Leyenda: ver Tabla 1

Discusión y conclusiones

El propósito de la presente investigación se ha centrado en proponer una prueba de aleatorización para analizar diseños de medidas parcialmente repetidas carentes normalidad y esfericidad multimuestral. También hemos examinado su desempeño en términos de robustez y lo hemos comparado con los enfoque clásico (MLG), mixto (MLM) y bootstrap-F. Las tasas empíricas de error de Tipo I fueron recopiladas cuando los datos habían sido extraídos desde distribuciones

simétricas (normal y exponencial doble) y asimétricas (exponencial y lognormal), las matrices de covarianza eran iguales y distintas, el grado de heterogeneidad de las matrices era moderado y severo, y el supuesto de esfericidad se cumplía o se incumplía severamente.

Estudios previos han puesto de relieve que el tradicional enfoque MLG, basado en obtener los valores críticos desde la teoría normal, sólo resulta apropiado cuando se cumplen los supuestos de esfericidad y de homogeneidad. Cuando se viola el supuesto de esfericidad, pero se satisface el de homogeneidad se requiere ajustar los grados de libertad para controlar razonablemente las tasas de error (Fernández et al., 2007; Keselman & Keselman, 1990). El enfoque MLM, por su parte, se ha mostrado robusto cuando ha sido usado para analizar datos repetidos heterogéneos, tanto normales como moderadamente sesgados (Chen & Wei, 2003; Kowalchuck et al., 2004; Vallejo & Livacic-Rojas, 2005). Mientras que distintos enfoques basados en el remuestreo bootstrap también limitan el número de errores Tipo I al valor nominal cuando las matrices dispersión son heterogéneas, aunque se comportan de un modo conservador bajo ciertas condiciones de no normalidad (Berkovits et al., 2000; Keselman et al., 2000; Kowakchuk et al., 2003; Vallejo et al., 2006). Nuestros resultados, además de ser consistentes con los obtenidos en los estudios reseñados, ofrecen nuevos hallazgos para ayudar a los investigadores a elegir alternativas analíticas viables.

Globalmente, nuestros resultados ponen de relieve que ninguno de los enfoques evaluados se ha mostrado efectivo bajo todas las condiciones manipuladas. En línea con lo que acabamos de expresar en el apartado anterior, el enfoque MLM tendía a comportarse de manera liberal cuando se violaban conjuntamente los supuestos de normalidad y de homogeneidad, especialmente cuando los datos eran extraídos desde la distribución lognormal. Mientras que el enfoque MLG con valores críticos obtenidos mediante remuestreo bootstrap tendía a comportarse de manera conservadora a medida que se incrementaba el sesgo de los datos, excepto cuando las matrices de dispersión eran heterogéneas y no estructuradas; situación esta última donde este enfoque se comportaba correctamente. Estos resultados no hacen más que confirmar y generalizar los encontrados en otros estudios similares (Vallejo & Livacic-Rojas, 2005; Vallejo et al., 2006).

Por lo que respecta al desempeño de la prueba de F con los valores críticos determinados mediante permutación aleatoria, hemos descubierto que esta prueba controlaba las tasas de error al nivel nominal elegido bajo todas las condiciones examinadas, excepto cuando las matrices de dispersión eran heterogéneas y no estructuradas. Así pues, la prueba resulta completamente robusta a la falta de normalidad, pero requiere del supuesto de homogeneidad (Hayes, 1996). El problema sería menor si, como ocurría en nuestra investigación, las matrices de dispersión fuesen proporcionales unas de otras. Una inspección detallada de las tasas empíricas de error Tipo I recogidas en las Tabla 2, evidencia que el tamaño de éstas coincide estrechamente con el resultado de multiplicar el valor nominal por el valor inverso de la desviación de esfericidad. Por consiguiente, una forma rápida de corregir el problema consiste dividir el valor crítico por el recíproco de ε (índice de ausencia de esfericidad). Sin embargo, no estamos seguros que la solución recomendada sea igual de efectiva cuando las matrices de dispersión no sean proporcionales entre sí, lo cual parece sea lo más factible de acontecer en la práctica. En estos casos habría que ajustar los valores críticos por el valor promediado de ε.

Para concluir queremos efectuar tres breves comentarios. En primer lugar, los resultados son limitados a las condiciones examinadas en el estudio, aunque intuimos que pueden ser generalizadas a un rango de condiciones más amplio. En segundo lugar, cuando las varianzas sean heterogéneas y la forma de la distribución simétrica, recomendamos utilizar el enfoque MLM. Cuando las varianzas sean homogéneas y la forma de la distribución asimétrica, consideramos más justificado utilizar las pruebas de aleatorización. Por último, cuando las varianzas sean heterogéneas y la forma de la distribución asimétrica, sugerimos usar la prueba de aleatorización ajustada y/o el enfoque bootstrap.

Reconocimientos: Este trabajo ha sido financiado mediante el proyecto de investigación concedido a los autores del mismo por el MCI (Ref.: PSI-2008-03624).

Referencias

Arnau, J., Bono, R. & Vallejo, G. (2009). Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics - Simulation and Computation, 38*, 1083-1103.

Akritas, M. G., Arnold, S .F., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Associacion, 92*(437), 258-265.

Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. *British Journal of Mathematical and Statistical Psychology, 47*, 151-165.

Anderson, M. J. & ter Braak, C. J. F. (2003). Permutation tests for multifactorial analysis of variance. *Journal of Statistical Computation and Simulation*, *73*, 85-113.

Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research, 37*(2), 197-226.

Berkovits, I., Hancok, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60,* 877-892.

Blanca-Mena, M. J. (2004). Alternativas de análisis estadístico en los diseños de medidas repetidas. *Psicothema, 16,*509-518.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Bono, R., Arnau, J. & Vallejo, G. (2008). Técnicas de análisis aplicadas a datos longitudinales en psicología y ciencias de la salud: Período 1985-2005. *Papeles del Psicólogo, 29,* 1-15.

Chen, X., & Wei, L. (2003). A comparison of recent methods for the analysis of small-sample cross-over studies. *Statistics in Medicine, 22*, 2821-2833.

Crowley, P.H. (1992). Resampling methods for computation-intensive data analysis in ecology and evolution, *Annual Review of Ecological Systems 23*, 405–447.

Edgington, E. S. (1995). *Randomization Tests*. 3rd Edition. New York: Marcel Dekker.

Edgington, E. S. & Onghena, P. (2007). *Randomization Tests*. 4th Edition, London: Chapman & Hall/CRC.

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fai, A. y Cornelius, P. (1996). Approximate F-test of multiple degree of freedom hypotheses in generalized least squares analyzes of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation, 54,* 363-378.

Fernández, P., Livacic-Rojas, P. & Vallejo, G. (2007). Cómo elegir la mejor prueba estadística para analizar un diseño de medidas repetidas. *International Journal of Clinical and Health Psychology* 7, 153-175.

Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

Gomez, V. E., Schaalje, G. B. & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics - Simulation and Computation, 34*, 377–392.

Gonzales, L. & Manly, B. F. J. (1998), Analysis of variance by randomization with small data sets. *Environmetrics* 9, 53-65.

Good, P. (2002), Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods 1*, 243-247.

Good, P. I. (2005). *Resampling Methods*: *a Practical Guide Data Analysis*. 3rd Edition, Boston: Birkhäuser.

Gurka, M. J. & Edwards, L. J. (2008). Mixed models. En C. R. Rao, J. P. Miller & D. C. Rao (Eds.): *Handbook of Statistics, Vol 27, Epidemiological and Medical Statistics* (pp. 253-280). New York: Elsevier Science.

Johansen, S. (1980). The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. *Biometrika, 67*, 85-92.

Jung, B. C., Jhun, M. & Song, S. H. (2006). A new random permutation test in ANOVA models. *Statistical Papers, 48,* 47-62.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics, 14*, 1431-1452.

Hayes, A. F. (1996). Permutation test is not distribution free. *Psychologial Methods, 1*, 184-198.

Kempthorne, O. & T. E. Doerfler (1969). The behaviour of some significance tests under experimental randomization. *Biometrika* 56, 231-48.

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Kenward, M. G. & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis, 53*, 2583-2595.

Keselman, J. C. & Keselman, H. J. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology, 43*, 265-282.

Keselman, J. C., Lix, L. M. & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology, 49*, 275-298.

Kowalchuk, R. K. & Headrick, T. C. (2009). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*. DOI:10.1348/000711009X42 3067.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M. & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology, 53*, 175-191.

Kowalchuk, R. K., Keselman, H. J. & Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research, 38*(4), 433-461.

Kowalchuk, R. K., Keselman, H. J., Algina, J. & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement, 64,* 224-242.

Laird, N. M. & Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics,* 38, 963-974.

Lei, X., Holt, J. K., & Beasley, T. M. (2004). Aligned rank tests as robust alternatives for testing interactions in multiple group repeated measures designs with heterogeneous covariances. *Journal of Modern Applied Statistical Methods, 3*(2), 462-475.

Lix, L. M., Algina, J., & Keselman, H. J. (2003). Analysing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research, 38*, 403-431.

Livacic-Rojas, P. E., Fernández, M. P. & Vallejo, G. (2006). Procedimientos estadísticos alternativos para evaluar la robustez mediante diseños de medidas repetidas. *Revista Latinoamericana de Psicología, 38,* 579-598.

Manly, B.F.J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 3rd Edition, London: Chapman& Hall/CRC.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 92*, 778-785.

Nuevo, R., Cabrera, I., Márquez-González, M. & Montorio, I. (2008). Comparación de dos procedimientos de inducción colectiva de ansiedad. *Anales de Psicología, 24*, 106-114.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association, 69*, 894-908.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science, 6,* 15-51.

SAS Institute Inc (2008). *SAS/STAT*® 9.2 *User's Guide*. Cary, NC: SAS Institute Inc.

Schaalje, G. B., McBride, J. B. & Fellingham, G. F. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 512-524.

Schabenberger, O. (2005). Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models. *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute.

Schabenberger, O. (2007). Growing Up Fast: SAS® 9.2 Enhancements to the GLIMMIX Procedure. *Proceedings of the 2007 SAS Global Forum*. Cary, NC: SAS Institute.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley and Sons, Inc.

ter Braak, C. J .F. (1992), Permutation versus bootstrap significance tests in multiple regression and ANOVA. In Jockel, K. H. (ed.). *Bootstrapping and Related Techniques, pp. 79-86*. Berlin: Springer.

Tukey, J.W. (1977). Modern techniques in data analysis. NSF-sponsored regional research conference at Southern Massachusetts University (North Dartmouth, MA).

Vallejo, G. & Ato, M. (2006). Modified Brown-Forsythe procedure for testing interaction effects in split-plot designs. *Multivariate Behavioral Research*, *41,* 549-578.

Vallejo, G., Ato, M. & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology. European Journal of Research for the Behavioral and Social Sciences, 4,* 10-21.

Vallejo, G. & Livacic-Rojas, P. E. (2005). A comparison of two procedures for analyzing small sets of repeated measures data. *Multivariate Behavioral Research, 40,* 179-205.

Vonesh, E. F. & Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wilcox, R. R. (2001). *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. New York: Springer.

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*, 205-230.

(Artículo recibido: 13-9-2009; revisado: 12-3-2010; aceptado: 12-3-2010)