Metodología de las Ciencias del Comportamiento 2(1), 85-103

Robustez y potencia de algunas pruebas no paramétricas para el efecto interactivo

Mª Eva Trigo Sánchez y José López Ruiz
Universidad de Sevilla[1]

Resumen

Se realizó un estudio de simulación Monte Carlo para analizar el comportamiento relativo de la *F* del AVAR paramétrico convencional frente a diversas pruebas estadísticas no paramétricas del efecto de interacción entre dos variables. En un diseño balanceado 2×2 con distribución normal en todas las condiciones se manipuló el grado de homogeneidad de las varianzas de error, el tamaño del efecto interactivo y el tamaño del efecto principal de una de las variables. Los resultados permiten descartar algunas de las pruebas no paramétricas incluidas en el estudio debido a su insuficiente control de la tasa de error tipo I, incluso en condiciones ideales de homogeneidad de varianzas. Otras en cambio resultan descartables debido a su fuerte dependencia del tamaño de efecto principal presente en los datos, disminuyendo la potencia de prueba a medida que éste aumenta. Entre las restantes, la *F* convencional resulta ser la prueba más potente cuando se cumple el supuesto de homogeneidad de las varianzas de error, pero se ve superada por otras alternativas no paramétricas en caso contrario.
PALABRAS CLAVE: *Simulación Monte Carlo, Pruebas no paramétricas, Efecto interactivo, Hhomogeneidad de varianzas, Error tipo I, Error tipo II.*

Abstract

ROBUSTNESS AND POWER OF SOME NONPARAMETRIC TESTS FOR INTERACTION. A Monte Carlo simulation study was carried out in order to analyze relative performance of the conventional parametric *F* test and different nonparametric tests for first-order interaction effect. The variance homogeneity assumption, the interaction effect size, and the main effect size of one variable were manipulated in a balanced factorial 2×2 design under population normality conditions. Some of the nonparametric tests show clear undesiderable properties related to type I error rate control, even when the variance homogeneity condition holds. Some other ones depend on the main effect size, decreasing their power as a function of the effect size increasing. Among the remaining procedures, the parametric *F* test shows the highest power under homogeneity of variances, but other nonparametric alternatives had better performance than the *F* test under violation of that assumption.
KEY WORDS: *Monte Carlo simulation, Nonparametric tests, Interaction effect, Variance homogeneity, Type I error rate, Type II error rate.*

La *F* del AVAR paramétrico sigue siendo la prueba más usada para contrastar un efecto de interacción entre dos variables, o su equivalente *t* de Student cuando dicho efecto tiene un grado de libertad. Entre los motivos para ello podrían destacarse principalmente dos: la creencia generalizada en su robustez y el desconocimiento de otras alternativas de análisis, especialmente en el caso de los diseños factoriales.

La supuesta robustez de la prueba *F* convencional tiene su origen en los estudios sobre tasas empíricas de error tipo I en condiciones de violación de los supuestos del modelo lineal de análisis. Así, la mayoría de las investigaciones sobre robustez de la prueba *F* muestran que la violación del supuesto de normalidad, ya sea por asimetría o curtosis, no tiene demasiadas consecuencias sobre la tasa de error tipo I, aun cuando ésta puede incrementarse ligeramente con distribuciones asimétricas y tamaño desigual de los grupos (Harwell, Rubinstein, Hayes y Olds, 1992). En cuanto al supuesto de homogeneidad de las varianzas de error, la mayoría de los investigadores están de acuerdo en que la prueba *F* resulta bastante robusta a la violación

[1]Dirección postal de los autores: Departamento de Psicología Experimental. Av. San Francisco Javier, s/n. 41005 SEVILLA. *E-mail:* trigo@psicoexp.us.es

de dicho supuesto cuando el diseño es balanceado. En cambio, si el tamaño de las muestras difiere la prueba se hace conservadora cuando se empareja el mayor tamaño muestral con la mayor varianza, y liberal cuando al mayor tamaño muestral le corresponde la varianza menor (Lix, Keselman y Keselman, 1996). Pero la importancia práctica de estas investigaciones sobre tasas de error tipo I con varianzas heterogéneas puede no ser muy elevada. Por un lado, en muchas ocasiones se puede conseguir un tamaño muestral constante sin un excesivo coste, especialmente si se trata de situaciones experimentales. Por otro lado, aunque no imposible (Wilcox, 1987), en la práctica resulta improbable encontrar distribuciones que difieran en sus varianzas y no en sus medias (Sawilowsky y Blair, 1992; Wilcox, 1995a).

Sin embargo, la robustez de la prueba *F* desaparece cuando nos referimos a las tasas de error tipo II, que sí pueden cambiar drásticamente en determinadas condiciones. La ausencia de normalidad puede influir de forma decisiva sobre la potencia, aumentando la tasa de error tipo II, cuando se comparan medias. El grado de curtosis es importante porque cuando las distribuciones tienen colas pronunciadas, los valores extremos, que afectan decisivamente a la estimación del error, son mucho más probables que cuando la distribución es normal; por su parte, las distribuciones asimétricas no pueden ser adecuadamente descritas a través de una media, sin otras medidas adicionales de posición (Wilcox, 1995b). En cuanto al supuesto de homocedasticidad, al igual que se ven afectadas por él las tasas de error tipo I, también pueden variar las de error tipo II, especialmente si el diseño no es balanceado. La cuestión fundamental no es por tanto que en algunos estudios los investigadores hayan rechazado erróneamente la hipótesis nula, sino que muchos descubrimientos pueden permanecer ocultos por haber utilizado sin más los métodos estadísticos convencionales (Wilcox, 1998b).

Para solucionar estos problemas se han propuesto y defendido múltiples alternativas, desde los ya clásicos métodos paramétricos heterocedásticos, como los propuestos por Welch (1951) o Yuen (1974), hasta los más modernos métodos robustos de estimación (Wilcox, 1995a; 1998a), pasando por las pruebas no paramétricas. Una primera ventaja de estas últimas sería (Hettmansperger, 1998; Sawilowsky, 1998) que no requieren criterios de aplicación adicionales, como el porcentaje de datos a considerar en el procedimiento de medias recortadas propuesto por Yuen. También se resaltan su simplicidad, sus propiedades de potencia ante el incumplimiento de los supuestos, así como la posibilidad de realizarlas con paquetes estadísticos sencillos (Hettmansperger, op.cit.). No obstante, algunas de las opciones anteriores también están implementadas en dichos paquetes, como el procedimiento de Welch, mientras que muchas pruebas no paramétricas, especialmente para diseños factoriales, aún no han sido programadas.

Para justificar nuestro interés por ellas podríamos resaltar además que los métodos ordinales pueden resultar más robustos y potentes cuando no se cumplen los supuestos del modelo lineal de análisis, que la mayoría de los datos conductuales sólo tienen carácter ordinal, y que la mayoría de las hipótesis hacen referencia a la pregunta general de si existe alguna condición donde la ejecución sea inferior o superior a la del resto (Cliff, 1993; 1996; Vargha y Delaney, 1998). Finalmente, un motivo adicional de interés en el caso de los diseños factoriales es la variedad de pruebas propuestas para el análisis del efecto interactivo, aún sometidas a estudio en la actualidad, a la vez que la escasa divulgación y utilización de las mismas (Harwell, 1991; Sawilowsky, 1990).

En el presente estudio hemos considerado diferentes pruebas no paramétricas para el contraste de la hipótesis nula de ausencia de efecto interactivo, $H_0 : AB = 0$. Dichas pruebas se describen a continuación.

Análisis ordinal de dominancia

Al igual que otros métodos ordinales, como la prueba U de Mann-Whitney, el análisis de dominancia realiza inferencias sobre el parámetro δ (Agresti, 1984; Hettmansperger, 1984). Dicho parámetro representa la probabilidad de que cada valor x_i extraído de una población X sea superior a cada valor y_j extraído de una población Y menos la probabilidad de que ocurra lo contrario:

$$\delta = p(x_i > y_j) - p(x_i < y_j)$$

y puede ser estimado a partir de la muestra, $\delta = E(d)$:

$$d = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{n_i n_j} = \frac{\sum_i \sum_j d_{ij}}{n_i n_j}$$

donde # indica el número de veces en que ocurre la expresión contenida entre paréntesis; d_{ij} denota el signo de la diferencia entre cada puntuación x_i de una de las muestras y cada puntuación x_j de la otra muestra; y $n_i n_j$ indica el número total de comparaciones entre cada una de las observaciones de una muestra con cada una de las observaciones de la otra (los emparejamientos de datos no contarán en el numerador, pero sí en el denominador de la ecuación).

Entre las características más destacables de d estarían el hecho de que su varianza muestral se estima fácilmente, dependiendo simplemente del grado de solapamiento entre las distribuciones, su invarianza ante transformaciones monotónicas de la escala de medida y sus propiedades descriptivas del grado en que dos distribuciones muestrales se solapan, lo que constituye en muchos casos la hipótesis directamente planteada por el investigador (Cliff, 1993; 1996). En este sentido, aunque d constituye una transformación de la U de Mann-Whitney:

$$d = \frac{2U}{n_1 n_2} - 1$$

resulta mucho más claramente interpretable que una diferencia de rangos promedio.

El estimador muestral de la varianza de d fue sugerido por Birnbaum (1956) y ha sido descrito por diversos autores. Un estimador insesgado de dicha varianza sería (Cliff, 1993):

$$\begin{aligned} S_d^2 &= \frac{n_i^2 \sum (d_{i.} - d)^2 + n_j^2 \sum (d_{.j} - d)^2 - \sum \sum (d_{ij} - d)^2}{n_i n_j (n_i - 1)(n_j - 1)} \\ &= \frac{n_i}{n_j} \frac{S_{d_{i.}}^2}{(n_i - 1)} + \frac{n_j}{n_i} \frac{S_{d_{.j}}^2}{(n_j - 1)} - \frac{S_{d_{ij}}}{n_i n_j} \end{aligned}$$

donde $d_{i.}$ representa el promedio de las diferencias de signo entre las distintas puntuaciones x_j de una de las muestras menos cada puntuación x_i de la otra; y $d_{.j}$ representa el promedio de las diferencias de signo entre las distintas puntuaciones x_i de una de las muestras menos cada puntuación x_j de la otra.

El cociente $Z = d/S_d$, que sigue una distribución $N\ (0,1)$ bajo el supuesto de ausencia de interacción en la población, nos permitirá llegar a una conclusión estadística sobre el grado en que la ejecución de un grupo es superior a la ejecución de otro. Cuando el análisis de dominancia se aplica al efecto interactivo en un diseño factorial lo que nos interesa es averiguar si el grado de solapamiento entre dos grupos, por ejemplo a_1 y a_2, en una de las condiciones de la otra variable, por ejemplo b_1, es igual o no al que ocurre en la otra condición de dicha variable, por ejemplo b_2. Los valores d y S_d^2 que nos interesarían en tal caso serían:

$$\begin{aligned} d &= d_{A|b_1} - d_{A|b_2} \\ S_d^2 &= S^2_{d\,A|b_1} + S^2_{d\,A|b_2} \end{aligned}$$

Esta aplicación del análisis de dominancia en los diseños factoriales, de la que no existe ningún estudio de simulación previamente publicado, tiene la particularidad de no resultar necesariamente equivalente cuando se aplica sobre los efectos simples de la variable A, como en nuestro ejemplo, o sobre los efectos simples de la variable B. Ambas formas de aplicación de la prueba fueron incluidas en nuestra simulación, resultando especialmente interesante analizar las posibles ventajas relativas de una u otra estrategia de análisis en función de la combinación de efectos presente en el diseño.

X^2 de Wilson

La prueba de Wilson (1956) comienza calculando la mediana de todos los datos y las frecuencias observadas de valores iguales o mayores a dicha mediana (O) por un lado, y menores a ella (O') por el otro, para cada una de las condiciones jk del diseño. Las frecuencias esperadas (E y E') dependerán del sumatorio de las frecuencias observadas y del número de condiciones experimentales del diseño ($ab = 4$ en nuestro diseño factorial 2×2):

$$E_{total} = \frac{\sum_j^a \sum_k^b O_{jk}}{ab}; \quad E'_{total} = \frac{\sum_j^a \sum_k^b O'_{jk}}{ab}$$

Una vez calculadas las frecuencias observadas y esperadas se procede a calcular los valores de chi-cuadrado para cada una de las tablas:

$$X^2_{(\geq mediana)} = \frac{\sum_j^a \sum_k^b (O_{jk} - E_{total})^2}{E_{total}}; \quad X^2_{(<mediana)} = \frac{\sum_j^a \sum_k^b (O'_{jk} - E'_{total})^2}{E'_{total}}$$

La chi-cuadrado total, para el conjunto de efectos del diseño, será el resultado de sumar ambos valores:

$$X^2_{total} = X^2_{(\geq mediana)} + X^2_{(<mediana)}$$

Por su parte, las frecuencias esperadas para los efectos principales de A y B se obtienen mediante el mismo procedimiento considerando únicamente las condiciones implicadas en cada efecto:

$$E_A = \frac{\sum_j^a O_{j.}}{a}; \quad E'_A = \frac{\sum_j^a O'_{j.}}{a}; \quad E_B = \frac{\sum_k^b O_{.k}}{b}; \quad E'_B = \frac{\sum_k^b O'_{.k}}{b}$$

y se calculan con ellas los valores de chi-cuadrado para ambos efectos principales:

$$X^2_A = \frac{\sum_j^a (O_{j.} - E_A)^2}{E_A} + \frac{\sum_j^a (O'_{j.} - E'_A)^2}{E'_A}; \quad X^2_B = \frac{\sum_j^a (O_{j.} - E_B)^2}{E_B} + \frac{\sum_j^a (O'_{j.} - E'_B)^2}{E'_B}$$

Finalmente, el valor correspondiente al efecto interactivo será el resultado de suprimir del total los valores obtenidos para los efectos principales de A y B:

$$X^2_{AB} = X^2_{total} - X^2_A - X^2_B$$

Dicho estadístico sigue una distribución chi-cuadrado con $(a-1)(b-1)$ grados de libertad bajo el supuesto de H_0 verdadera.

X^2 de Shoemaker

Ante los pobres resultados de la prueba de Wilson cuando existen efectos principales en el diseño, Shoemaker (1986) propuso un procedimiento para suprimir dichos efectos principales previamente a la estimación del efecto interactivo. Para ello agrupa los datos en función de una de las variables, por ejemplo B, calculando las medianas para cada nivel de dicha variable y las distancias de cada puntuación respecto a dichas medianas. A continuación se agrupan estas distancias en función de los valores de la otra variable, en nuestro ejemplo A, calculando las medianas de las distancias para cada nivel de A. Se calculan entonces por un lado las frecuencias observadas de distancias superiores a la mediana de las distancias, en función de los niveles de A, y por otro las frecuencias iguales o inferiores.

La frecuencia esperada depende del número de puntuaciones del diseño (N en un diseño transversal), del número de condiciones del mismo (ab en un diseño factorial) y del criterio de división de las puntuaciones (superiores/inferiores o iguales a la mediana):

El estadístico de contraste para el efecto interactivo se obtiene a través del sumatorio de todas las distancias cuadráticas entre frecuencias observadas y esperadas:

$$X^2_{AB} = \frac{\sum_j^a \sum_k^b (O_{jk} - E)^2 + \sum_j^a \sum_k^b (O'_{jk} - E)^2}{E}$$

y sigue una distribución chi-cuadrado con $(a-1)(b-1)$ grados de libertad bajo el supuesto de ausencia de efecto interactivo en la población.

Estas dos últimas pruebas incluidas en la simulación se basan por tanto en la distribución de frecuencias alrededor de la mediana, ya sea sin supresión previa de los efectos principales presentes en el diseño, la prueba de Wilson, o realizando previamente dicha supresión, la de Shoemaker.

En un tercer grupo se sitúan varias pruebas basadas todas ellas de forma muy directa en el modelo lineal de análisis y la transformación de las puntuaciones en rangos. Al igual que en el grupo anterior, algunas de ellas no suprimen previamente los efectos principales presentes en el diseño, el AVAR de rangos y la prueba de Puri y Sen, y otras realizan dicha supresión antes de transformar en rangos las puntuaciones originales, fundamentalmente las pruebas de Hettmansperger y McSweeney.

AVAR de rangos

Esta prueba se basa directamente en el concepto de transformación en rangos propuesto por Conover e Iman (1981). Consiste simplemente en transformar las puntuaciones originales en rangos y calcular las correspondientes $SSCC$ de rangos (SCr) para los efectos principales e interactivos y para el error:

$$SC_rA = n_{j.}\sum_{j}^{a}(\bar{r}_{j.} - \bar{r}_{..})^2; \quad SC_rB = n_{k.}\sum_{k}^{b}(\bar{r}_{k.} - \bar{r}_{..})^2$$

$$SC_rAB = n_{jk}\sum_{j}^{a}\sum_{k}^{b}(\bar{r}_{jk} + \bar{r}_{..} - \bar{r}_{j.} - \bar{r}_{.k})^2$$

$$SC_rError = \sum_{j}^{a}\sum_{k}^{b}\sum_{i}^{n_{jk}}(r_{ijk} - \bar{r}_{jk})^2$$

$$SC_rTotal = \sum_{j}^{a}\sum_{k}^{b}\sum_{i}^{n_{jk}}(r_{ijk} - \bar{r}_{..})^2 = \frac{N(N^2-1)}{12} \text{ (si no hay empates)}$$

donde r_{ijk} representa el rango asignado a cada sujeto i de cada condición experimental jk; $r_{j.}$ es el promedio de rangos correspondiente a cada condición j de la variable A; $\bar{r}_{.k}$ es el promedio de rangos correspondiente a cada condición k de la variable B; $\bar{r}_{jk}$ es el promedio de rangos correspondiente a cada combinación jk; y $\bar{r}_{..}$ es el promedio de todos los rangos.

Bajo el supuesto de H_0 verdadera, el cociente $F = CM_rAB/CM_rError$ seguirá una distribución F de Snedecor con $(a-1)(b-1)$ y $N-ab$ grados de libertad.

Esta prueba es una de las más estudiadas en diseños bicondicionales, donde resulta equivalente a la conocida U de Mann-Whitney, y en diseños multicondicionales, donde resulta equivalente a la prueba de Kruskal-Wallis (e.g. Vargha y Delaney, 1998; Zimmerman, 1992). Para diseños factoriales es considerada relevante por algunos autores (e.g. Zimmerman, 1994), o queda excluida por otros (e.g. Harwell, 1991) por haber obtenido unos pobres resultados en otros estudios previos (Akritas, 1990; Blair, Sawilowsky y Higgins, 1987; Sawilowsky, Blair y Higgins, 1989).

L de Puri y Sen

La prueba propuesta por Puri y Sen (1985) para el efecto interactivo resulta equivalente en un contexto de grupos independientes a la expresión:

$$L = (N-1)\frac{SC_{rAB}}{SC_{rTotal}}$$

que sigue una distribución chi-cuadrado con $(a-1)(b-1)$ grados de libertad bajo el supuesto de H_0 verdadera.

H de Hettmansperger

La prueba propuesta por Hettmansperger (1984) requiere la supresión de los efectos principales antes de realizar la conversión en rangos. Lo que se convierte en rangos son por tanto las estimaciones conjuntas de error y efecto interactivo:

$$y^*_{ijk} = y_{ijk} + \bar{y}_{..} - \bar{y}_{j.} - \bar{y}_{.k}$$

Una vez transformadas las puntuaciones originales para suprimir los efectos principales que pudieran estar presentes en el diseño, se convierten a rangos (rangos*) y se aplica la transformación adicional:

$$r'_{ijk} = \left(\frac{r_{ijk}}{N+1} - .5\right)\sqrt{12}$$

Finalmente se calcula la SC de rangos' de AB sobre estas puntuaciones, que sigue una distribución chi-cuadrado con $(a-1)(b-1)$ grados de libertad bajo el supuesto de ausencia de efecto interactivo en la población.

M de McSweeney

La prueba de McSweeney (1967) utiliza la supresión previa a la transformación en rangos de los efectos principales:

$$y^*_{ijk} = y_{ijk} + \bar{y}_{..} - \bar{y}_{j.} - \bar{y}_{.k}$$

y una vez realizada ésta aplica la misma formulación que la prueba de Puri y Sen:

$$M = (N-1)\frac{SC_{r^*}AB}{SC_{r^*}Total}$$

siguiendo igualmente una distribución chi-cuadrado con $(a-1)(b-1)$ grados de libertad bajo el supuesto de H_0 verdadera.

AVAR de rangos*

Adicionalmente decidimos realizar también sobre nuestros datos un AVAR de rangos*. Bajo el supuesto de H_0 verdadera, el cociente seguirá una distribución F de Snedecor con $(a-1)(b-1)$ y $N-ab$ grados de libertad. Aunque ningún autor ha

propuesto hasta la fecha esta prueba creemos que resulta una extensión lógica del principio de conversión en rangos de Conover e Iman cuando se quiere que al mismo tiempo el efecto interactivo no esté contaminado por los efectos principales.

En definitiva, el objetivo del presente trabajo es mostrar los resultados de un estudio de simulación donde se analiza el comportamiento relativo de estas pruebas no paramétricas de efecto interactivo tanto entre sí como respecto al AVAR paramétrico convencional. Dicha simulación se realizó en diferentes condiciones de homogeneidad de las varianzas de error, el supuesto más importante del modelo lineal de análisis en los diseños transversales, y en diferentes condiciones de tamaños de efecto principal e interactivo, ya que algunas de las pruebas incluidas no son independientes de los efectos principales que pudieran estar presentes en los datos.

Método

Se utilizó el programa MATLAB 5.0.0.4069 para generar valores muestrales de un diseño factorial 2×2 con 8 sujetos por combinación de niveles *AB* y distribución normal en todas las condiciones. Mientras que este tipo de distribución resulta con diferencia la más investigada en los estudios de simulación sobre alternativas de análisis a la *F* de Snedecor, merece la pena destacar que no ocurre lo mismo con los diseños factoriales (ver Harwell, 1992; Harwell et al., 1992; Lix et al., 1996 para revisiones cuantitativas sobre el tema). En cualquier caso, algunos de ellos incluyen entre sus condiciones de estudio la distribución normal y el mismo tamaño muestral constante de nuestro estudio (e.g. Harwell, 1991; Zimmerman, 1994), aunque difieren sustancialmente en las variables manipuladas y sus valores concretos. En nuestro estudio dichas variables fueron el grado de cumplimiento de la condición de homogeneidad de varianzas y el tamaño de los efectos principales e interactivo.

Las diferencias entre las medias de las combinaciones del diseño se produjeron sumando determinadas constantes a las combinaciones correspondientes, de forma que fuese posible estudiar tanto las probabilidades empíricas de error tipo I como de error tipo II con diferentes tamaños de efectos. Se combinaron 3 valores diferentes de diferencias de medias tipificadas de una de las variables, 0, 1 ó 2, con esos mismos 3 valores para la interacción. El efecto principal de la segunda variable del diseño siempre fue nulo. Los tamaños de efecto elegidos están dentro del rango utilizado en otros estudios de simulación con pruebas no paramétricas (e.g. Borges, Sánchez y Cañadas, 1996; Zimmerman y Zumbo, 1993). En cuanto a las varianzas de error, los valores utilizados fueron 1, 1, 1 y 1 para la condición de homogeneidad y 1, 1, 1 y 9 para la condición de heterogeneidad. Estos valores también fueron elegidos por Keselman, Kowalchuk y Lix (1998), por producir pruebas liberales en un diseño unifactorial, así como por no resultar nada infrecuentes a nivel práctico. Comprobamos además que dichos valores darían lugar al rechazo de la hipótesis nula de homogeneidad de varianzas con la prueba *F máxima* de Hartley: $F_{max.}(4,7) = 8.44, p = .05$. Las diferentes combinaciones de diferencias de medias se ilustran en la Tabla 1, donde σ puede tomar el valor 1 ó $\sqrt{3}(\sqrt{(1+1+1+9)/4})$ para las condiciones de homogeneidad y heterogeneidad de varianzas respectivamente.

Tabla 1. Combinaciones de efectos de tratamiento simuladas para un diseño factorial 2×2

		$B = 0$		$B = 1$		$B = 2$	
		b_1	b_2	b_1	b_2	b_1	b_2
$AB = 0$	a_1	μ_{11}	μ_{12}	$\mu_{11} + 1\sigma$	μ_{12}	$\mu_{11} + 2\sigma$	μ_{12}
	a_2	μ_{21}	μ_{22}	$\mu_{21} + 1\sigma$	μ_{22}	$\mu_{21} + 2\sigma$	μ_{22}
$AB = 1$	a_1	μ_{11}	$\mu_{12} + 1\sigma$	μ_{11}	$\mu_{12} + 2\sigma$	$\mu_{11} + 1\sigma$	μ_{12}
	a_2	$\mu_{21} + 1\sigma$	μ_{22}	$\mu_{21} + 1\sigma$	$\mu_{22} + 1\sigma$	$\mu_{21} + 2\sigma$	$\mu_{22} - 1\sigma$
$AB = 2$	a_1	μ_{11}	$\mu_{12} + 2\sigma$	μ_{11}	$\mu_{12} + 3\sigma$	μ_{11}	$\mu_{12} + 4\sigma$
	a_2	$\mu_{21} + 2\sigma$	μ_{22}	$\mu_{21} + 2\sigma$	$\mu_{22} + 1\sigma$	$\mu_{21} + 2\sigma$	$\mu_{22} + 2\sigma$

Se realizaron 5000 replicaciones de cada una de las 18 condiciones de estudio y se llevaron a cabo los análisis estadísticos correspondientes para los dos niveles de significación estadística convencionales, $\alpha = .05$ y $\alpha = .01$. De acuerdo con Robey y Barcilowski (1992), este número de iteraciones es superior al requerido para detectar con una potencia aceptable ($1 - \beta = .80$) y un nivel de significación pequeño ($\alpha = .01$), desviaciones moderadas ($\alpha \pm 1/4\alpha$) de la proporción empírica de error respecto a $\alpha = .05$. Dicho criterio está situado entre los extremos estricto y liberal propuestos por Bradley (1978). En cambio, tal número de iteraciones sólo permitiría detectar en condiciones de igual potencia y mayor nivel de significación ($\alpha = .05$) las desviaciones grandes respecto a $\alpha = .01$ ($\alpha \pm 1/2\alpha$). Estos serán por tanto los valores considerados a la hora de enjuiciar la robustez de las distintas pruebas para cada nivel de significación estadística.

Por último, debemos señalar que en algunas condiciones del diseño el número de iteraciones contabilizadas para el análisis de dominancia sobre los efectos simples de B no llegó a 5000, ya que en algunos casos resultó imposible aplicar dicha prueba. Cuando el efecto principal de B es fuerte puede ocurrir que el valor promedio de dominancia en las condiciones de A sea 1, $d = 1$, al ser todos los valores de una de las muestras superiores a todos los valores de la otra. En tal caso $S_d^2 = 0$, y el cociente d/S_d daría como resultado infinito. Aunque existen soluciones para este problema (Cliff, 1996; Feng y Cliff, manuscrito no publicado), derivado de la correlación negativa entre δ y σ_d^2, decidimos no realizar el ajuste correspondiente. Uno de los motivos de esta decisión fue recoger datos sobre el procedimiento original de análisis de dominancia en diseños factoriales, no estudiados hasta ahora; entre dichos datos incluimos la proporción de iteraciones en que resulta imposible calcular la prueba. Por otro lado, los propios autores (Feng y Cliff, op.cit.) destacan en sus conclusiones que probablemente deben realizarse modificaciones adicionales del procedimiento de ajuste propuesto.

Como puede observarse en la Tabla 2, cuando el efecto interactivo es fuerte ($AB = 2$) el tamaño de efecto de B prácticamente no afecta a las posibilidades de aplicar el análisis de dominancia original sobre los efectos simples de B. En cambio, con tamaños de efecto interactivo menores ($AB = 1$ y sobre todo $AB = 0$) dichas posibilidades disminuyen conforme aumenta el tamaño de efecto de la variable B. Este mismo patrón tiene lugar independientemente de que las varianzas sean o no homogéneas, aumentando en general el número de réplicas invalidadas cuando son heterogéneas.

Tabla 2. Proporción de réplicas ($N = 5000$) en que no puede aplicarse el análisis de dominancia sobre los efectos simples de B en función de las condiciones del estudio.

	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
	$B = 0$	$B = 1$	$B = 2$	$B = 0$	$B = 1$	$B = 2$
$AB = 0$	.0000	.0002	.0266	.0000	.0006	.0942
$AB = 1$	.0000	.0000	.0048	.0000	.0000	.0390
$AB = 2$	.0000	.0000	.0000	.0000	.0000	.0008

Resultados

En la Tabla 3 se muestran las probabilidades empíricas de error tipo I ($AB = 0$) obtenidas bajo las diferentes condiciones del estudio para $\alpha = .05$ y $\alpha = .01$.

Tabla 3. Probabilidades empíricas de error tipo I para las diferentes pruebas del estudio en función de la homogeneidad de varianzas, el tamaño del efecto principal y el nivel de significación.

	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
$\alpha = .05$	$B = 0$	$B = 1$	$B = 2$	$B = 0$	$B = 1$	$B = 2$
Dominancia $A/b_1 - Ab_2$	**.072***	**.070***	**.069***	**.077***	**.068***	**.078***
Dominancia $A/b_1 - Ab_2$	**.068***	.060	**.019**	**.074***	**.141***	**.037**
X^2 de Wilson	.059	**.033**	**.003**	**.068***	.051	**.013**
X^2 de Shoemaker	**.167***	**.154***	**.162***	**.326***	**.312***	**.322***
F AVAR paramétrico	.052	.051	.053	**.064***	.061	**.064***
L de Puri y Sen	.050	**.030**	**.004**	.055	.041	**.010**
F AVAR de rangos	.054	.052	.051	.061	**.076***	**.071***
H de Hettmansperger	.057	.055	.059	**.120***	**.118***	**.119***
M de McSweeney	.057	.055	.059	**.120***	**.118***	**.119***
F AVAR de rangos*	.054	.051	.054	**.117***	**.112***	**.115***
	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
$\alpha = .01$	$B = 0$	$B = 1$	$B = 2$	$B = 0$	$B = 1$	$B = 2$
Dominancia $A/b_1 - Ab_2$	**.024***	**.024***	**.026***	**.027***	**.022***	**.026***
Dominancia $A/b_1 - Ab_2$	**.024***	**.015***	**.001**	**.026***	**.039***	**.004**
X^2 de Wilson	.011	**.005**	**<.001**	.014	.010	**.002**
X^2 de Shoemaker	**.047***	**.052***	**.050***	**.160***	**.153***	**.152***
F AVAR paramétrico	.010	.011	.011	**.018***	**.016***	**.018***
L de Puri y Sen	.009	**.004**	**<.001**	.010	.008	**.001**
F AVAR de rangos	.011	.012	.011	**.016***	**.019***	**.017***
H de Hettmansperger	.010	.010	.011	**.035***	**.030***	**.033***
M de McSweeney	.010	.010	.012	**.038***	**.032***	**.036***
F AVAR de rangos*	.010	.010	.012	**.040***	**.035***	**.037***

Nota. En negrita se señalan los valores mayores o iguales (marcados con asterisco) y menores o iguales a los límites de los intervalos $.05 \pm .05/4$ y $.01 \pm .01/2$.

Estos resultados, muy similares para ambos niveles de significación estadística, revelan claramente cómo algunas de las pruebas incluidas en nuestro estudio no consiguen controlar las tasas de error tipo I, ni siquiera bajo la condición ideal de homogeneidad de las varianzas de error. Así ocurre con el análisis de dominancia, especialmente el realizado sobre los efectos simples de la variable B. Por su parte, el

análisis de dominancia sobre los efectos simples de A depende del tamaño de efecto de B, con probabilidades altas de error cuando éste es pequeño y muy bajas cuando es grande.

También presenta problemas de error tipo I la prueba de Shoemaker, ya que, si bien soluciona la dependencia respecto a los efectos principales de la prueba de Wilson, resulta excesivamente liberal. Dicha dependencia también aparece claramente en la prueba L de Puri y Sen, mucho más conservadora, pero no en la prueba F del AVAR de rangos, que tampoco suprime previamente los efectos principales presentes en el diseño.

Entre las restantes, todas ellas controlan adecuadamente la tasa de error tipo I cuando las varianzas son homogéneas. En cambio, cuando las varianzas de error son significativamente heterogéneas, todas estas pruebas se vuelven más liberales. Aquellas en que menos influye este aspecto son el análisis de dominancia de efectos simples de B, que ya resultaba prácticamente igual de liberal bajo la condición de homogeneidad, y las pruebas X^2 de Wilson y L de Puri y Sen, con una fuerte dependencia de los efectos principales presentes en el diseño.

Las probabilidades empíricas de error tipo II correspondientes a un tamaño de efecto interactivo igual a 1 se representan en la Tabla 4.

Tabla 4. Probabilidades empíricas de error tipo II para el análisis de un efecto interactivo $AB = 1$ en función de la homogeneidad de varianzas, el tamaño de efecto principal y el nivel de significación.

$AB = 1$	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
$\alpha = .05$	$B = 0$	$B = 1$	$B = 2$	$B = 0$	$B = 1$	$B = 2$
Dominancia $A/b_1 - Ab_2$	.216	.213	.207	.098	.101	.100
Dominancia $A/b_1 - Ab_2$	.211	.326	.497	.104	.182	.872
X^2 de Wilson	.363	.526	.827	.176	.313	.984
X^2 de Shoemaker	.218	.216	.221	.090	.085	.092
F AVAR paramétrico	.220	.222	.220	.214	.219	.223
L de Puri y Sen	.253	.342	.601	.122	.185	.746
F AVAR de rangos	.244	.248	.251	.106	.074	.293
H de Hettmansperger	.232	.227	.228	.146	.137	.143
M de McSweeney	.232	.227	.228	.146	.137	.143
F AVAR de rangos*	.244	.241	.237	.146	.139	.146
$AB = 1$	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
$\alpha = .01$	$B = 0$	$B = 1$	$B = 2$	$B = 0$	$B = 1$	$B = 2$
Dominancia $A/b_1 - Ab_2$	.374	.379	.348	.231	.232	.231
Dominancia $A/b_1 - Ab_2$	.372	.587	.817	.249	.451	.952
X^2 de Wilson	.642	.783	.962	.381	.547	.999
X^2 de Shoemaker	.443	.442	.437	.201	.189	.202
F AVAR paramétrico	.459	.470	.460	.446	.457	.456
L de Puri y Sen	.506	.645	.881	.329	.468	.968
F AVAR de rangos	.471	.493	.496	.289	.245	.516
H de Hettmansperger	.493	.502	.494	.328	.330	.332
M de McSweeney	.474	.487	.476	.312	.316	.317
F AVAR de rangos*	.473	.486	.474	.298	.303	.307

En primer lugar cabría destacar nuevamente el aumento de la probabilidad empírica de error que tiene lugar en algunas pruebas a medida que aumenta la cantidad

de efecto principal de B presente en el diseño: especialmente el análisis de dominancia sobre los efectos simples de B, X^2 de Wilson y L de Puri y Sen, siendo la prueba de Wilson la de menor potencia de todas incluso en ausencia de efecto principal. En relación con la F del AVAR de rangos, que no suprime previamente los efectos principales, no se aprecia esta dependencia respecto al tamaño de efecto de B cuando las varianzas son homogéneas, pero sí cuando no lo son.

Entre las restantes, las pruebas más potentes resultaron ser, para ambas condiciones de homogeneidad y ambos niveles de significación, el análisis de dominancia sobre los efectos simples de A y la X^2 de Shoemaker, si bien resulta conveniente recordar que estas pruebas no logran controlar adecuadamente el riesgo de error tipo I. Descartadas éstas, la F del AVAR paramétrico aparece como la prueba más potente cuando se cumple el supuesto de homogeneidad de varianzas. En cambio, cuando se viola dicho supuesto otras alternativas la superan en potencia: H de Hettmansperger, M de McSweeney y F del AVAR de rangos*, a pesar de que no se produce un incremento de la tasa de error tipo II del AVAR paramétrico. Todas estas alternativas presentan una potencia muy similar con $\alpha = .05$, mientras que la F del AVAR de rangos* resulta ligeramente más potente con un nivel de significación menor, $\alpha = .01$.

Tabla 5. Probabilidades empíricas de error tipo II para el análisis de un efecto interactivo $AB = 2$ en función de la homogeneidad de varianzas, el tamaño de efecto principal y el nivel de significación.

$AB = 1$	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
$\alpha = .05$	$B = 0$	$B = 1$	$B = 2$	$B = 0$	$B = 1$	$B = 2$
Dominancia $A/b_1 - Ab_2$	.0006	.0002	.0006	<.0001	.0002	<.0001
Dominancia $A/b_1 - Ab_2$	.0006	.0018	.1144	.0004	.0100	.1626
X^2 de Wilson	.0046	.0232	.2030	.0068	.0318	.2188
X^2 de Shoemaker	.0012	.0026	.0020	.0002	.0002	.0004
F AVAR paramétrico	.0002	<.0001	.0002	.0006	.0004	.0004
L de Puri y Sen	.0008	.0004	.0128	.0020	.0030	.0152
F AVAR de rangos	.0008	.0002	.0014	<.0001	.0004	.0002
H de Hettmansperger	.0004	<.0001	.0002	.0002	.0002	<.0001
M de McSweeney	.0004	<.0001	.0002	.0002	.0002	<.0001
F AVAR de rangos*	.0004	<.0001	.0002	.0002	.0002	<.0001
$AB = 1$	$\sigma^2 = 1:1:1:1$			$\sigma^2 = 1:1:1:9$		
$\alpha = .01$	B=0	B=1	B=2	B=0	B=1	B=2
Dominancia $A/b_1 - Ab_2$	.0020	.0014	.0028	.0016	.0014	.0018
Dominancia $A/b_1 - Ab_2$	.0018	.0198	.3666	.0020	.0662	.4448
X^2 de Wilson	.0424	.1178	.4774	.0176	.0874	.4334
X^2 de Shoemaker	.0144	.0142	.0168	.0006	.0020	.0018
F AVAR paramétrico	.0028	.0018	.0042	.0054	.0062	.0050
L de Puri y Sen	.0042	.0166	.1098	.0050	.0306	.1112
F AVAR de rangos	.0036	.0036	.0064	.0020	.0028	.0028
H de Hettmansperger	.0032	.0054	.0060	.0018	.0020	.0018
M de McSweeney	.0024	.0046	.0054	.0018	.0014	.0014
F AVAR de rangos*	.0024	.0044	.0052	.0016	.0012	.0010

La Tabla 5 muestra las probabilidades empíricas de error tipo II para un tamaño de efecto interactivo igual a 2. En términos generales, el patrón es muy similar al

que acabamos de describir, aunque lógicamente todas las pruebas muestran una mayor potencia, y como consecuencia, las diferencias entre ellas resultan más difíciles de apreciar debido a un efecto techo. Sólo algunos aspectos difieren respecto a la tabla anterior. Así, las pruebas de dominancia sobre los efectos simples de *A* y de Shoemaker dejan de ser las más potentes en todos los casos. Por su parte, la potencia de la *F* del AVAR de rangos en condiciones de heterogeneidad de varianzas no disminuye en función del tamaño de efecto de *B*. Y finalmente, la potencia del AVAR paramétrico sí parece disminuir ligeramente cuando se incumple el supuesto de homogeneidad de las varianzas de error.

Discusión y conclusiones

Nuestros resultados nos permiten en primer lugar descartar algunas de las pruebas no paramétricas estudiadas debido a su escaso control de la tasa de error tipo I en condiciones de homogeneidad de las varianzas de error. A este grupo pertenecen fundamentalmente el análisis de dominancia aplicado sobre los efectos simples de *A* y la X^2 de Shoemaker. En cambio, la problemática de la tasa de error tipo I en condiciones de heterocedasticidad recordemos que tiene menores implicaciones, ya que en la práctica si dos distribuciones difieren en sus varianzas lo más probable es que también difieran en sus medias. En cualquier caso, los datos ponen de manifiesto que cuando no se cumple el supuesto de homogeneidad la *F* convencional se vuelve liberal, pero en mucho menor medida que el resto de pruebas. Esto debe entenderse como una desventaja de éstas a la hora de validar la ausencia de efecto interactivo. No obstante, también podría ocurrir que dichas pruebas estuvieran detectando precisamente que las distribuciones difieren entre sí en otros aspectos como sus varianzas.

En segundo lugar podríamos descartar también otras alternativas no paramétricas a la *F* convencional para el efecto interactivo debido a su fuerte dependencia respecto a la existencia o no de efectos principales, así como respecto al tamaño de los mismos. En este grupo estarían fundamentalmente la X^2 de Wilson, la prueba *L* de Puri y Sen y el análisis de dominancia aplicado sobre los efectos simples de *B*. El análisis de dominancia por tanto resulta de difícil aplicación en los diseños factoriales. Aunque algunos ajustes posteriores del procedimiento consiguen un mejor comportamiento de la prueba en relación con la tasa de error tipo I en diseños unifactoriales (Feng y Cliff, manuscrito no publicado), quedaría por solucionar el problema de su falta de potencia para el efecto interactivo cuando también existen efectos principales. Adicionalmente, también podría formar parte de este grupo el AVAR de rangos en determinadas circunstancias; su dependencia respecto a la presencia de efectos principales se puso de manifiesto bajo condiciones de heterogeneidad de varianzas y un nivel de significación de .05. En cualquier caso, tampoco resulta ser la alternativa más potente a la *F* convencional.

Entre las restantes, la *F* del AVAR paramétrico resultó ser la mejor opción en términos de potencia en condiciones ideales de homogeneidad de varianzas. Sin embargo, cuando las varianzas de error diferían significativamente la alternativa no paramétrica más potente resultó ser la *F* del AVAR de rangos*, una vez suprimidos los efectos principales. Las pruebas de Hettmansperger y McSweeney resultaron tan potentes como ella con un nivel de significación de .05, pero perdieron potencia rela-

tiva con un nivel inferior. No obstante, conviene resaltar una vez más que no se trata de que el AVAR convencional pierda potencia de forma relevante; resultado lógico si tenemos en cuenta que el tamaño de efecto se mantuvo constante. Lo que ocurre es que el resto de alternativas al AVAR paramétrico ganan potencia relativa en condiciones de heterogeneidad de varianzas. En este sentido, podría resultar interesante de cara a un próximo estudio analizar si esta ganancia en potencia se produce igualmente a la hora de detectar una misma diferencia de medias con diferentes grados de heterogeneidad de varianzas.

En resumen, entre las alternativas no paramétricas analizadas, el análisis de dominancia, las pruebas basadas en la mediana y las basadas en rangos que no suprimen previamente los efectos principales serían difícilmente recomendables en la práctica. Las restantes resultan más potentes que la *F* del AVAR paramétrico en condiciones de heterocedasticidad, pero tampoco resultarían adecuadas si el investigador centra su preocupación en la posibilidad de cometer un error tipo I.

Finalmente, la mayoría de los estudios sobre pruebas no paramétricas comparan su comportamiento en situaciones de normalidad frente a múltiples distribuciones no normales (e.g. Sawilowsky y Blair, 1992; Zimmerman, 1994; ver también Lix et al., 1996 para una revisión sobre estudios de simulación). En cambio, cuando el problema que presentan los datos es el de heterocedasticidad se han estudiado más las alternativas paramétricas a la *F*, como los procedimientos de Welch, o el de medias recortadas de Yuen, denominados en general métodos heterocedásticos. Curiosamente en nuestro estudio algunas alternativas no paramétricas se han mostrado bastante eficientes en condiciones donde normalmente se opta por alguno de dichos procedimientos heterocedásticos. No obstante, sobre estos procedimientos suele resaltarse su control sobre las tasas de error tipo I, mientras que en nuestro caso estamos destacando la potencia; las tasas de error tipo I asociadas a estas pruebas siempre fueron superiores a las de la *F* convencional. En cualquier caso, convendría comparar ambos tipos de procedimientos respecto a ambas tasas de error, y puesto que resulta difícil encontrar en la literatura datos sobre ellas recogidos en circunstancias similares, esta comparación podría ser el objetivo de un próximo estudio de simulación.

Referencias

Agresti, A. (1984). *Analysis of ordinal categorical data.* New York: John Wiley & Sons.

Akritas, M.G. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association, 85,* 73-78.

Birnbaum, Z.W. (1956). On the use of the Mann-Whitney statistic. En J. Neyman (ed.). *Actas del Third Berkeley Symposium on Mathematical Statistics,* 13-17. Berkeley, Los Angeles: University of California Press.

Blair, R.C., Sawilowsky, S.S. y Higgins, J.J. (1987). Limitations of the rank transform in tests for interaction. *Communication in Statistics: Simulation and Computation, B16,* 1133-1145.

Borges, A., Sánchez, A. y Cañadas, I. (1996). El contraste de las diferencias de medias con grupos pequeños, con escalas ordinales y en ausencia de normalidad. *Psicológica, 17,* 455-466.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 147-150.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114,* 494-509.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Conover, W.J. y Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician,35,* 124-129.

Feng, D. y Cliff, N. *A Monte Carlo Exploration of Robustness and Power or Ordinal d and Welch's t.* Manuscrito no publicado.

Harwell, M.R. (1991). Completely randomized factorial analysis of variance using ranks. *British Journal of Mathematical and Statistical Psychology, 44,* 383-401.

Harwell, M.R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics, 17,* 297-313.

Harwell, M.R., Rubinstein, E.N., Hayes, W.S. y Olds, C.C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17,* 315-339.

Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks.* New York: John Wiley & Sons.

Hettmansperger, T.P. (1998). Comments (on The goal and strategies of robust methods by Rand R. Wilcox). *British Journal of Mathematical and Statistical Psychology, 51,* 41-42.

Keselman, H.J., Kowalchuk, R.K. y Lix, L.M. (1998). Robust nonorthogonal analyses revisited: an update based on trimmed means. *Psychometrika, 63,* 145-163.

Lix, L.M., Keselman, J.C. y Keselman, H.J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research, 66,* 579-619.

McSweeney, M. (1967). *An empirical study of two proposed nonparametric tests for main effects and interaction.* Tesis Doctoral. Universidad de Berkeley, California.

Puri, M.L. y Sen, P.K. (1985). *Nonparametric tests using a general linear model approach.* New York: John Wiley & Sons.

Robey, R.R. y Barcikowski, R.S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45,* 283-288.

Sawilowsky, S.S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60,* 91-126.

Sawilowsky, S.S. (1998). Comments (on The goal and strategies of robust methods by Rand R. Wilcox). *British Journal of Mathematical and Statistical Psychology, 51,* 49-52.

Sawilowsky, S.S. y Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111,* 352-360.

Sawilowski, S., Blair, R.C. y Higgins, J.J. (1989). An investigation of the type I error and power properties of the rank transformation procedure in factorial ANOVA. *Journal of Educational Statistics, 14,* 255-267.

Shoemaker, L.H. (1986). A nonparametric method for analysis of variance. *Communications in Statistics, 15,* 609-632.

Vargha, A. y Delaney, H.D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23,* 170-192.

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330-336.

Wilcox, R.R. (1987). New designs in analysis of variance. *Annual Review of Psychology, 61,* 165-170.

Wilcox, R.R. (1995a). ANOVA: The practical importance of heterocedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology, 48,* 99-114.

Wilcox, R.R. (1995b). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65,* 51-77.

Wilcox, R.R. (1998a). The goals and strategies of robust methods. British *Journal of Mathematical and Statistical Psychology, 51,* 1-39.

Wilcox, R.R. (1998b). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300-314.

Wilson, K.V. (1956). A distribution-free test of analysis of variance hypotheses. *Psychological Bulletin, 53,* 96-101.

Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika, 61,* 165-170.

Zimmerman, D.W. (1992). An extension of the rank transformation concept. *Journal of Experimental Education, 61,* 73-80.

Zimmerman, D.W. (1994). Simplified interaction tests for non-normal data in psychological research. *British Journal of Mathematical and Statistical Psychology, 47,* 327-335.

Zimmerman, D.W. y Zumbo, B.D. (1993). The relative power of parametric and nonparametric statistical methods. En G. Keren, y C. Lewis (eds.). *A Handbook for Data Analysis in the Behavioral Sciences: Methodological issues,* 481-517. Hillsdale, NJ: Lawrence Erlbaum Associates.

Original recibido: 11/3/00
Versión final aceptada: 15/5/00

ANEXO

datos hipotéticos	a_1	a_2
b_1	1, 2, 3, 3, 3, 3, 4, 5	3, 4, 5, 5, 5, 5, 6, 7
b_2	2, 3, 4, 4, 4, 4, 5, 6	1, 2, 3, 3, 3, 3, 4, 5

Análisis de Dominancia

		x_j								
	A/b_1	2	3	4	4	4	4	5	6	$d_{i.}$
x_i	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	2	0	-1	-1	-1	-1	-1	-1	-1	-.87
	3	1	0	-1	-1	-1	-1	-1	-1	-.62
	3	1	0	-1	-1	-1	-1	-1	-1	-.62
	3	1	0	-1	-1	-1	-1	-1	-1	-.62
	3	1	0	-1	-1	-1	-1	-1	-1	-.62
	4	1	1	0	0	0	0	-1	-1	0
	5	1	1	1	1	1	1	0	-1	-.62
	$d_{.j}$	.62	0	-.62	-.62	-.62	-.62	-.87	-1	-.47

$$S^2_{d_{ij}} = \frac{\sum\sum(d_{ij} - d)^2}{(n_i - 1)(n_j - 1)} = .82; \quad S^2_{d_{i.}} = \frac{\sum\sum(d_{i.} - d)^2}{(n_i - 1)} = .28; \quad S^2_{d_{.j}} = \frac{\sum\sum(d_{.j} - d)^2}{(n_j - 1)} = .28$$

$$S^2_{dA|b_1} = \frac{n_i}{n_j}\frac{S^2_{d_{i.}}}{(n_i - 1)} + \frac{n_j}{n_i}\frac{S^2_{d_{.j}}}{(n_j - 1)} - \frac{S^2_{d_{ij}}}{n_i n_j} = \frac{8}{8}\frac{.28}{7} + \frac{8}{8}\frac{.28}{7} - \frac{.82}{64} = .067$$

		x_j								
	A/b_2	1	2	3	3	3	3	4	5	$d_{i.}$
x_i	3	1	1	0	0	0	0	-1	-1	0
	4	1	1	1	1	1	1	0	-1	.62
	5	1	1	1	1	1	1	1	0	.87
	5	1	1	1	1	1	1	1	0	.87
	5	1	1	1	1	1	1	1	0	.87
	5	1	1	1	1	1	1	1	0	.87
	6	1	1	1	1	1	1	1	1	1
	7	1	1	1	1	1	1	1	1	1
	$d_{.j}$	1	1	.87	.87	.87	.87	.62	0	.77

$$S^2_{d_{ij}} = \frac{\sum\sum(d_{ij} - d)^2}{(n_i - 1)(n_j - 1)} = .36; \quad S^2_{d_{i.}} = \frac{\sum\sum(d_{i.} - d)^2}{(n_i - 1)} = .11; \quad S^2_{d_{.j}} = \frac{\sum\sum(d_{.j} - d)^2}{(n_j - 1)} = .11$$

$$S^2_{dA|b_2} = \frac{n_i}{n_j}\frac{S^2_{d_{i.}}}{(n_i - 1)} + \frac{n_j}{n_i}\frac{S^2_{d_{.j}}}{(n_j - 1)} - \frac{S^2_{d_{ij}}}{n_i n_j} = \frac{8}{8}\frac{.11}{7} + \frac{8}{8}\frac{.11}{7} - \frac{.36}{64} = .026$$

$$d_A = d_{A|b_1} - d_{A|b_2} = (-.47) - (.77) = -1.23; \quad S^2_{dA} = S_{dA|b_1} + S_{dA|b_2} = .067 + .026$$

$$Z_A = \frac{d_A}{\sqrt{S^2_{dA}}} = \frac{-1.23}{.30} = -4.05; p < .001$$

		x_j								
	B/b_1	3	4	5	5	5	5	6	7	$d_{i.}$
	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	2	-1	-1	-1	-1	-1	-1	-1	-1	-1
	3	0	-1	-1	-1	-1	-1	-1	-1	-.87
x_i	3	0	-1	-1	-1	-1	-1	-1	-1	-.87
	3	0	-1	-1	-1	-1	-1	-1	-1	-.87
	3	0	-1	-1	-1	-1	-1	-1	-1	-.87
	4	1	0	-1	-1	-1	-1	-1	-1	-.62
	5	1	1	0	0	0	-1	-1	-1	0
	$d_{.j}$	0	-.62	-.87	-.87	-.87	-.87	-1	-1	-.77

$$S^2_{d_{ij}} = \frac{\sum\sum(d_{ij} - d)^2}{(n_i - 1)(n_j - 1)} = .36; \quad S^2_{d_{i.}} = \frac{\sum\sum(d_{i.} - d)^2}{(n_i - 1)} = .11; \quad S^2_{d_{.j}} = \frac{\sum\sum(d_{.j} - d)^2}{(n_j - 1)} = .11$$

$$S^2_{dA|b_2} = \frac{n_i}{n_j}\frac{S^2_{d_{i.}}}{(n_i - 1)} + \frac{n_j}{n_i}\frac{S^2_{d_{.j}}}{(n_j - 1)} - \frac{S^2_{d_{ij}}}{n_i n_j} = \frac{8}{8}\frac{.11}{7} + \frac{8}{8}\frac{.11}{7} - \frac{.36}{64} = .026$$

		x_j								
	A/b_2	1	2	3	3	3	3	4	5	$d_{i.}$
	3	1	0	-1	-1	-1	-1	-1	-1	-.62
	4	1	1	0	0	0	0	-1	-1	0
	5	1	1	1	1	1	1	0	-1	.62
x_i	5	1	1	1	1	1	1	0	-1	.62
	5	1	1	1	1	1	1	0	-1	.62
	5	1	1	1	1	1	1	0	-1	.62
	6	1	1	1	1	1	1	1	0	.87
	7	1	1	1	1	1	1	1	1	1
	$d_{.j}$	1	.87	.62	.62	.62	.62	0	-.62	.47

$$S^2_{d_{ij}} = \frac{\sum\sum(d_{ij} - d)^2}{(n_i - 1)(n_j - 1)} = .36; \quad S^2_{d_{i.}} = \frac{\sum\sum(d_{i.} - d)^2}{(n_i - 1)} = .28; \quad S^2_{d_{.j}} = \frac{\sum\sum(d_{.j} - d)^2}{(n_j - 1)} = .28$$

$$S^2_{dB|a_1} = \frac{n_i}{n_j}\frac{S^2_{d_{i.}}}{(n_i - 1)} + \frac{n_j}{n_i}\frac{S^2_{d_{.j}}}{(n_j - 1)} - \frac{S^2_{d_{ij}}}{n_i n_j} = \frac{8}{8}\frac{.28}{7} + \frac{8}{8}\frac{.28}{7} - \frac{.36}{64} = .074$$

$$d_A = d_{A|b_1} - d_{A|b_2} = (-.77) - (.47) = -1.23; \quad S^2_{dB} = S_{dB|a_1} + S_{dB|a_2} = .026 + .074$$

$$Z_A = \frac{d_B}{\sqrt{S^2_{dB}}} = \frac{-1.23}{.32} = -3.90; p < .001$$

Prueba de Wilson

Mediana total = 3.5

$O_{jk}(\geq 3.5)$	a_1	a_2	
b_1	2	6	8
b_2	7	2	9
	9	8	17

$O'_{jk}(< 3.5)$	a_1	a_2	
b_2	6	2	8
b_2	1	6	7
	7	8	15

$$E_{total} = \frac{\sum_j^a \sum_k^b O_{jk}}{ab} = \frac{17}{4} = 4.25; \quad E'_{total} = \frac{\sum_j^a \sum_k^b O'_{jk}}{ab} = \frac{1}{4} = 3.75$$

$$X^2_{(\geq mediana)} = \frac{\sum_j^a \sum_k^b (O_{jk} - E_{total})^2}{E_{total}} = \frac{20.75}{4.25} = 4.88; \quad X^2_{(<mediana)} = \frac{\sum_j^a \sum_k^b (O'_{jk} - E'_{total})^2}{E'_{total}} = \frac{20.75}{3.75} = 5$$

$$X^2_{total} = X^2_{(\geq mediana)} + X^2_{(<mediana)} = 4.88 + 5.54 = 10.42$$

$$E_A = \frac{\sum_j^a O_{j.}}{a} = \frac{17}{2} = 8.5 \quad E'_A = \frac{\sum_j^a O'_{j.}}{a} = \frac{15}{2} = 7.5 \quad E_B = \frac{\sum_k^b O_{.k}}{b} = \frac{17}{2} = 8.5 \quad E'_B = \frac{\sum_k^b O'_{.k}}{b} = \frac{15}{2} = 7.5$$

$$X^2_A = \frac{\sum_j^a (O_{j.} - E_A)^2}{E_A} + \frac{\sum_j^a (O'_{j.} - E'_A)^2}{E'_A} = .13; \quad X^2_B = \frac{\sum_j^a (O_{.k} - E_B)^2}{E_B} + \frac{\sum_j^a (O'_{j.} - E'_B)^2}{E'_B} = .13$$

$$X^2_{AB} = X^2_{total} - X^2_A - X^2_B = 10.42 - .13 - .13 = 10.16; \quad X^2(1, N = 32) = 10.16; \; p = .001$$

Prueba de Shoemaker

Mediana de b_1 = 3.5 Mediana de b_2 = 4

	a_1	a_2
$y_{i.1} - 3.5$	$-2.5, -1.5, -.5, -.5, -.5, -.5, .5, 1.5$	$-1.5, -.5, .5, .5, .5, .5, 1.5, 2.5$
$y_{i.2} - 4$	$-1, 0, 1, 1, 1, 1, 2, 3$	-3,-2,-1,-1,-1,-1,0,1
Mediana	.25	-.25

	$O_{jk}(> .25)$	$O_{jk}(> .25)$
b_1	2	6
b_2	6	2

	$O'_{jk}(\leq .25)$	$O'_{jk}(\leq -.25)$
b_1	6	2
b_2	2	6

$$E = \frac{N}{ab} * 2 = \frac{32}{8} = 4$$

$$X^2_{AB} = \frac{\sum_j^a \sum_k^b (O_{jk} - E)^2 + \sum_j^a \sum_k^b (O'_{jk} - E)^2}{E} = \frac{16 + 16}{4} = 8; \quad X^2(1, N = 32) = 8; \; p = .005$$

AVAR paramétrico

$$F = \frac{\frac{SC_{AB}}{gl_{AB}}}{\frac{SC_{Error}}{gl_{Error}}} = \frac{18/1}{40/28} = \frac{18}{1.43}; \quad F(1, 28) = 12.6; \; p = .001$$

AVAR de rangos

rangos	a_1	a_2
b_1	1.5, 4, 10.5, 10.5, 10.5, 10.5, 19, 26	10.5, 19, 26, 26, 26, 26, 30.5, 32
b_2	4, 10.5, 19, 19, 19, 19, 26, 30.5	1.5, 4, 10.5, 10.5, 10.5, 10.5, 19, 26

$$F_r = \frac{\frac{SC_{rAB}}{gl_{rAB}}}{\frac{SC_{rError}}{gl_{rError}}} = \frac{780.125/1}{1656.313/28} = \frac{780.125}{9.154} = 13.19; \quad F(1,28) = 13.19; \quad p = .001$$

Prueba de Puri y Sen

$$L = (N-1)\frac{SC_{rAB}}{SC_{rTotal}} = (32-1)\frac{78.125}{2586.5} = 9.35; \quad X^2(1, N = 32) = 9.35; \quad p = .002$$

Prueba de Hettmansperger

rangos′	a_1	a_2
b_1	−1.57, −1.36, −.63, −.63, −.63, −.63, 10, 1.15	−1.15, −.10, .63, .63, .63, .63, 1.36, 1.57
b_2	−1.15, −.10, .63, .63, .63, .63, 1.36, 1.57	−1.57, −1.36, −.63, −63, −.63, −.63, .10, 1.15

$$H = SC_{r'AB} = 8.81; \quad X^2(1, N = 32) = 8.81; \quad p = .003$$

Prueba de McSweeney

rangos*	a_1	a_2
b_1	1.5,3.5,10.5,10.5,10.5,10.5,17.5,27.5	5.5,15.5,22.5,22.5,22.5,22.5,29.5,31.5
b_2	5.5,15.5,22.5,22.5,22.5,22.5,29.5,31.5	1.5,3.5,10.5,10.5,10.5,10.5,17.5,27.5

$$M = (N-1)\frac{SC_{r^*AB}}{SC_{r^*Total}} = (32-1)\frac{800}{2640} = 9.39; \quad X^2(1, N = 32) = 9.39; \quad p = .002$$

AVAR de rangos*

$$F_{r^*} = \frac{\frac{SC_{r^*AB}}{gl_{r^*AB}}}{\frac{SC_{r^*Error}}{gl_{r^*Error}}} = \frac{800/1}{1840/28} = \frac{800}{65.71} = 12.17; \quad F(1,28) = 12.17; \quad p = .002$$