Psicológica (1999) 20, 211-226.

Puntuaciones atípicas y potencia estadística con diferentes procedimientos de análisis de los tiempos de reacción: Un estudio de simulación

Manuel Perea* y Salvador Algarabel

Universitat de València

Se analizó el efecto de las puntuaciones atípicas en el análisis de los tiempos de reacción. Se efectuaron simulaciones Monte Carlo para investigar la influencia de diversos procedimientos usuales de recorte y transformación de datos y su impacto en diseños intra-sujeto en los que se recogen múltiples observaciones para cada participante por condición experimental. Las simulaciones mostraron que el empleo de medias truncadas (esto es, medias obtenidas tras la exclusión de datos que estén más allá de cierto rango, v.g., 300-1500 ms) es un procedimiento estadísticamente potente. En todo caso, se debería efectuar un análisis de las puntuaciones atípicas para observar si los porcentajes de puntuaciones excluidas son similares a través de las condiciones. Adicionalmente, se analiza cómo procedimientos gráficos (a partir de promedios «vincentizados») pueden complementar los análisis de los tiempos de reacción.

Palabras clave: simulación, potencia, tiempos-reacción, puntuaciones-atípicas.

El tiempo de reacción es, sin duda, la variable dependiente más empleada a la hora de estudiar los procesos cognitivos, y los investigadores han ideado un buen número de tareas cronométricas a tal efecto. Un fenómeno habitual en los experimentos que emplean el tiempo de reacción como variable dependiente es la existencia de cierto número de puntuaciones atípicas (es decir, puntuaciones que se alejan del resto de los datos), habitualmente producidas por factores externos a los procesos de interés, tales como distracciones de los participantes o fallos en el instrumental. Resulta evidente que, dado que usualmente se recogen múltiples observaciones por participante por condición experimental, el empleo de la media aritmética de los tiempos de reacción por condición experimental se verá influida muy

* Esta investigación fue subvencionada por una beca de la Dirección General de Investigación Científica y Técnica (PB/97-1379). La correspondencia sobre este trabajo deberá ser enviada a Manuel Perea. Departament de Metodologia. Facultat de Psicologia. Av. Blasco Ibáñez, 21. 46010-València (Spain). (e-mail: mperea@uv.es)

fácilmente por la existencia de tales puntuaciones atípicas, que añadirán variabilidad espuria a las estimaciones y, en consecuencia, provocarán un descenso en la potencia estadística.

Por ello, a modo de «póliza de seguros» (Anscombe, 1960) los psicólogos cognitivos suelen emplear diferentes técnicas de recorte (o transformación) de aquellos datos provenientes de distribuciones de los tiempos de reacción (para una revisión histórica sobre los procedimientos de recorte de datos, véase Beckman y Cook, 1983). La justificación de tales procedimientos de recorte se basa en que tales observaciones atípicas no son meramente observaciones alejadas del grueso de datos, sino que son puntuaciones provenientes de una distribución que no es la de interés (esto es, son datos contaminantes, según la nomenclatura de Barnett y Lewis, 1995). Lógicamente, el problema radica en que no existe un procedimiento estandarizado que permita distinguir las observaciones que aun siendo atípicas siguen la distribución de interés de aquellas puntuaciones atípicas que no provienen de tal distribución (véase Barnett y Lewis, 1995, para una revisión de diversos procedimientos para tratar con las puntuaciones extremas). Cuando el tiempo de reacción es muy breve (v.g., 122 ms), es claro que dicha observación atípica es contaminante porque difícilmente un participante, incluso con un esfuerzo hercúleo, puede lograr un tiempo de reacción legítimo que sea tan bajo. Sin embargo, con tiempos de reacción altos, siempre queda la duda de que la puntuación atípica obtenida sea representativa de los procesos bajo estudio (que, aunque poco probable, es siempre una posibilidad) o bien sea una puntuación contaminante. Como indicó Ratcliff (1993), tales puntuaciones contaminantes se esconden en el extremo superior de la distribución y son difícilmente distinguibles de datos legítimos (pero elevados) que provengan de la distribución de interés.

En la literatura de los estudios que emplean el tiempo de reacción puede observarse que los investigadores emplean diferentes técnicas de recorte o transformación de datos: medianas (que pueden ser consideradas como un caso extremo de media recortada en que se mantiene sólo la observación central), transformaciones de datos (v.g., medias logarítmicas, en las que las puntuaciones extremas positivas tiene un menor peso), medias truncadas (medias obtenidas tras la exclusión de aquellas observaciones que se encuentren fuera de ciertos límites prefijados, v.g., 300-1500 ms), medias restringidas (medias obtenidas tras la exclusión de aquellos datos que se encuentren más allá de dos desviaciones típicas de la media de cada participante, véase Miller, 1991), medias semi-restringidas (o windsorizadas; similares a las medias restringidas, pero en las que se sustituyen las observaciones que se encuentren más allá de la media ± dos desviaciones típicas de la media de cada participante por los puntos de corte, véase Ratcliff, 1993), entre otros procedimientos. Esta es, sin duda, una situación poco deseable porque, como señalaron Ulrich y Miller (1994), se podría sospechar que los resultados están afectados por la técnica de recorte/transformación de datos empleada. De este modo, es posible que los investigadores pudieran estar incrementando la probabilidad de cometer un error de tipo I o bien un error de tipo II como consecuencia del tipo de procedimiento de análisis de

los tiempos de reacción. Ello hace necesario un estudio detallado de los procedimientos más aconsejables de recorte de datos.

Antes de examinar los trabajos previos sobre este ámbito de estudio, es importante señalar que las distribuciones de los tiempos de reacción no sólo suelen contener cierto número de observaciones atípicas, sino que suelen mostrar cierta asimetría positiva. Una distribución teórica que tiene un buen ajuste con la distribución de los tiempos de reacción es la distribución ex-Gaussiana (v.g., Balota y Spieler, 1999; Heathcote, 1996; Heathcote, Popiel y Mewhort, 1991; Luce, 1986; Plourde y Besner, 1997; Ratcliff, 1978, 1979; Ratcliff y Murdock, 1976). La distribución 'ex-Gaussiana' se define como la suma de dos variables aleatorias independientes, una que sigue la distribución exponencial (responsable de la asimetría positiva, con un parámetro único τ, que hace a la vez de media y varianza) y otra que sigue la distribución Gaussiana (o normal, con los parámetros μ y σ). De esta manera, estos tres parámetros: μ, σ y τ especifican la forma de la distribución ex-Gaussiana y permiten una amplia flexibilidad respecto a su forma. Aunque la distribución de los tiempos de reacción también puede ser simulada con otras distribuciones teóricas, como la lognormal, gamma, la distribución de Wald, entre otras (véase Luce, 1986; Ulrich y Miller, 1994), hemos preferido centrarnos en la distribución ex-Gaussiana por diversas razones (véase Balota y Spieler, 1999): 1) la media de una condición experimental se puede estimar fácilmente mediante la suma de μ y τ; 2) los efectos pueden ser tratados mediante un componente relativo al grueso de la distribución y otro relativo a la asimetría de la distribución; y 3) se ha observado en diferentes estudios empíricos que las características de la distribución ex-Gaussiana son estables a través de los participantes.

Recientemente ha habido varios trabajos que han analizado la potencia de diversos procedimientos de recorte y transformación de datos con distribuciones de los tiempos de reacción (v.g., Ratcliff, 1993; Ulrich y Miller, 1994). El trabajo de Ratcliff (1993) se ocupaba de diversos procedimientos de recorte (empleo de medias truncadas, empleo de medias restringidas a 1 y 1’5 desviaciones típicas de la media del participante, empleo de medias semi-restringidas a 2 desviaciones típicas de la media del participante) y de transformación de datos (medias armónica y logarítmica). A partir de las simulaciones efectuadas sobre la distribución ex-Gaussiana, Ratcliff (1993) aboga por el empleo de medias truncadas, siempre y cuando los efectos sean significativos a través de diferentes puntos de corte (es decir, los resultados serían realmente fiables cuando el efecto fuera significativo no sólo con los puntos de corte fijo de 300 y 1000 ms, sino también que el efecto fuera significativo con otros puntos de corte como, por ejemplo, los de 300 y 1250 ms, o 300 y 1500 ms). Además, dos procedimientos que también se mostraron con bastante potencia estadística como las medias armónicas y las medias restringidas podrían emplearse para confirmar los análisis obtenidos con el empleo de puntos de corte fijo. En caso de divergencia, o en caso de que el efecto obtenido fuera nuevo o inesperado, Ratcliff (1993) sugirió que se debería optar por la replicación del experimento (ya sea replicación directa o, mejor, sistemática). Por su parte, Ulrich y Miller (1994) sólo analizaron una

técnica de recorte de datos (el empleo de medias truncadas), indicando mediante un complejo análisis que los límites deberían eliminar sólo una mínima cantidad de los datos (no más de 0'5% de los datos). En todo caso, Ulrich y Miller (1994) reconocen la pérdida de potencia de su sugerencia. Cabe señalar que, en otro trabajo de simulación, Bush, Hess y Wolford (1993) también efectúan un análisis de potencia de diversos procedimientos, aunque lo aplican a una distribución asimétrica menos realista para los tiempos de reacción que la distribución ex-Gaussiana, como es la distribución chi-cuadrado. Además, Bush et al. (1993) se centran más en procedimientos de transformación de datos (v.g., mediante logaritmos, raíces cuadradas, puntuaciones z, etc.) de variables fisiológicas que en procedimientos de recorte de datos (v.g., medianas, medias truncadas) habituales en psicología cognitiva.

En el presente trabajo se analizarán los procedimientos de recorte y transformación de datos más empleados en la actualidad ¾ con un mayor número de procedimientos que los empleados por Ratcliff (1993)¾ ante situaciones en que se varía el grado de asimetría y la posible presencia/ausencia de observaciones contaminantes. El objetivo fue observar los efectos de tales manipulaciones sobre la tasa de errores de tipo I y la potencia de los diversos procedimientos. En el trabajo de Ratcliff (1993) el grado de asimetría se mantuvo constante en el marco de una distribución marcadamente asimétrica (μ=400 y τ=200) y resulta de interés averiguar si los procedimientos más potentes bajo unas condiciones pueden no serlo tanto, comparativamente, bajo otras. Otro aspecto de interés de este estudio es que, a diferencia del trabajo de Ratcliff (1993), se compararon directamente las medias restringidas con las semi-restringidas con un mismo criterio de exclusión (o sustitución, según se trate de medias restringidas o semi-restringidas): media ± 2 desviaciones típicas de la media del participante, que es el más habitual en la literatura psicológica (v.g., Forster y Veres, 1998). En las simulaciones efectuadas por Ratcliff (1993) el criterio era de media ± 1 (o 1'5) desviaciones típicas para las medias restringidas, y de media ± 1 desviaciones típicas de la media del participante para las medias semi-restringidas, lo que hacía imposible comparar la potencia de las medias restringidas y semi-restringidas (ni tampoco comparar las medias restringidas con media ± 2 desviaciones típicas del participante respecto al empleo de medias truncadas). Adicionalmente, el cómputo de la media y desviación típica de las medias restringidas se calculó tanto sobre la media global del participante en todas las condiciones (como se suele efectuar habitualmente) como sobre la media del participante en cada condición (véase Miller, 1991), con el objeto de examinar cuál de ambos procedimientos es el más potente.

Los procedimientos analizados en el presente trabajo, junto al empleo de la media aritmética con todos los datos (que hará las funciones de línea base), son de dos tipos, por una parte, procedimientos de transformación de datos (medias armónicas y medias logarítmicas; sobre la base de que los datos extremos positivos tienen menos peso y así la distribución resultante se parece más a la normal) y, por otra, siete procedimientos de recorte de datos: 1) medias restringidas sobre la media de cada participante, en las que se eliminan

las observaciones que están más allá de 2 desviaciones típicas de la media de cada participante; 2) medias semi-restringidas sobre la media de cada participante, en las que las observaciones que están más allá de 2 desviaciones típicas de la media de cada participante no se eliminan, sino que se sustituyen por los puntos de corte; 3) medias restringidas sobre cada condición, en las que se eliminan las observaciones que están más allá de 2 desviaciones típicas de la media de cada condición y participante; 4) medias semi-restringidas sobre cada condición, en las que las observaciones que están más allá de 2 desviaciones típicas de la media de cada condición y participante no se eliminan, sino que se sustituyen por los puntos de corte; 5) medias truncadas con unos puntos de corte fijos de 250-1000 ms; 6) medias truncadas con unos puntos de corte fijos de 250-1500 ms; y 7) mediana, que es la observación central dentro del conjunto de datos ordenados por condición.

Sin duda, la potencia de una prueba estadística se incrementa a medida que aumenta el tamaño muestral, de manera que es más fácil detectar una diferencia de medias que sea real (en tiempos de reacción o en otra variable) cuando las muestran se extraen de muchas observaciones que cuando se extraen de pocas. Dado que los procedimientos de recorte reducen el número de observaciones de las que se extraen los promedios de cada participante por condición se podría pensar que estos procedimientos producen un descenso en potencia estadística al compararlos con la potencia obtenida con la media aritmética de todos los datos. No obstante, en los procedimientos de recorte no se eliminan (o alteran) datos al azar, sino los datos extremos (que varían de acuerdo con el procedimiento), de manera que los promedios obtenidos tras el recorte estimarán un valor central con una variabilidad menor que los obtenidos al aplicar directamente la media aritmética sobre todos los datos. La cuestión a analizar es cuál de los dos componentes tiene un peso mayor en el descenso de potencia (Ulrich y Miller, 1994). Además, puede pensarse que la reducción en la variación muestral que ocurre en las medias obtenidas con los procedimientos de recorte afecta a la probabilidad de error de tipo I, de manera que sea más sencillo rechazar la hipótesis nula (siendo ésta cierta) cuando se efectúan procedimientos de recorte que cuando no, por lo que parece conveniente analizar primeramente dicho aspecto, antes de ocuparnos de la potencia estadística.

SIMULACIÓN

MÉTODO

Fueron consideradas tres situaciones experimentales básicas a partir de variaciones de la distribución ex-Gaussiana, en la que se variaban los parámetros de dicha distribución. Tales situaciones son paralelas a las empleadas en simulaciones anteriores (v.g., Hockley, 1984; Miller, 1988, 1991; Perea, 1999; Ponsoda y Alcázar, 1996; Van Selst y Jolicoeur, 1993) y, como en tales trabajos, la media poblacional fue de 600 ms. Como se indicó en la Introducción, la distribución ex-Gaussiana se define como la suma de dos variables aleatorias independientes: 1) un componente que sigue la

distribución normal (Gaussiana), con media μ y desviación típica σ, y 2) un componente que sigue la distribución exponencial con media y varianza τ. He aquí la función de densidad de la distribución ex-Gaussiana:

$$f(x) = \frac{e^{-[(x-\mu)/\tau] + \sigma^2/2\tau^2}}{\tau(2\pi)^{1/2}} \int_{-\infty}^{[(x-\mu)/\sigma] - \sigma/\tau} e^{-y^2/2} dy$$

En el programa de simulación realizado al efecto, la generación de los números pseudoaleatorios, distribuidos uniformemente en el intervalo (0,1), fue realizada mediante el generador de números pseudoaleatorios de Wichmann y Hill (1984). Las muestras aleatorias de la distribución normal estandariza fueron generadas a partir de números pseudoaleatorios uniformes mediante el método propuesto por Box y Müller (1958). Lógicamente, las medias y las varianzas de la distribución normal estandarizada se modificaron para obtener las distribuciones de las diferentes condiciones experimentales mediante transformaciones aritméticas simples. Finalmente, las muestras aleatorias de la distribución exponencial (con media τ) fueron obtenidas —a partir de números pseudoaleatorios uniformes— mediante el método de la transformación inversa.

Las tres situaciones básicas presentaban una misma variabilidad en el componente de la distribución normal (σ=40) y variaban en el grado de asimetría, si bien todas ellas tenían una media de 600 ms. Dicha asimetría podía ser pequeña (μ=575, τ=25), moderada (μ=475, τ=125) o alta (μ=375, τ=225). En la figura 1 se representa la función de densidad de las tres distribuciones ex-Gaussianas consideradas. Para simular la variación de los tiempos de reacción de los participantes, los diferentes componentes de la distribución variaban aleatoriamente, de acuerdo con una distribución uniforme (véase Ratcliff, 1993). En concreto, y de manera similar al trabajo de Ratcliff (1993), los componentes μ y τ podían variar ± 25 ms respecto al indicado en la situación experimental para cada participante, mientras que el componente σ variaba ± 10 ms respecto al indicado en la situación experimental.

Estas tres situaciones básicas se presentan bien sin la inclusión de puntuaciones contaminantes, o bien con la inclusión de puntuaciones contaminantes. Para simular las condiciones en que la distribución de datos presentaba observaciones "contaminadas" se efectuó lo siguiente. Cada dato tenía una probabilidad de 0’05 de ser una observación contaminante. En tal caso se sumaba a dicha observación extraída de la distribución ex-Gaussiana un valor aleatorio uniforme entre 1 y 2000 ms. Estas condiciones de «contaminación» son similares a las simuladas por Ratcliff (1993), si bien en dicho trabajo la probabilidad de tener una observación contaminada era algo mayor (0’10). En nuestro caso, hemos preferido el valor de 0’05 porque pensamos que ante participantes cooperativos y con un instrumental preciso, un valor del 10% de puntuaciones contaminantes es posiblemente demasiado

elevado. Por otra parte, el hecho de emplear una distribución uniforme para la distribución contaminante se debe a que no hay razón definida que permite saber si los datos espurios (contaminantes) son más habituales en unos valores que en otros, por lo que parece razonable optar por una distribución uniforme. Además, parece razonable suponer que, en cierta proporción de casos, la existencia de observaciones contaminantes no implica necesariamente un valor atípico. Por ejemplo, un error producido por la activación de la llave vocálica en un experimento de pronunciación ha podido dar lugar a un tiempo de reacción "razonable" de 500-700 ms (véase Barnett y Lewis, 1995).

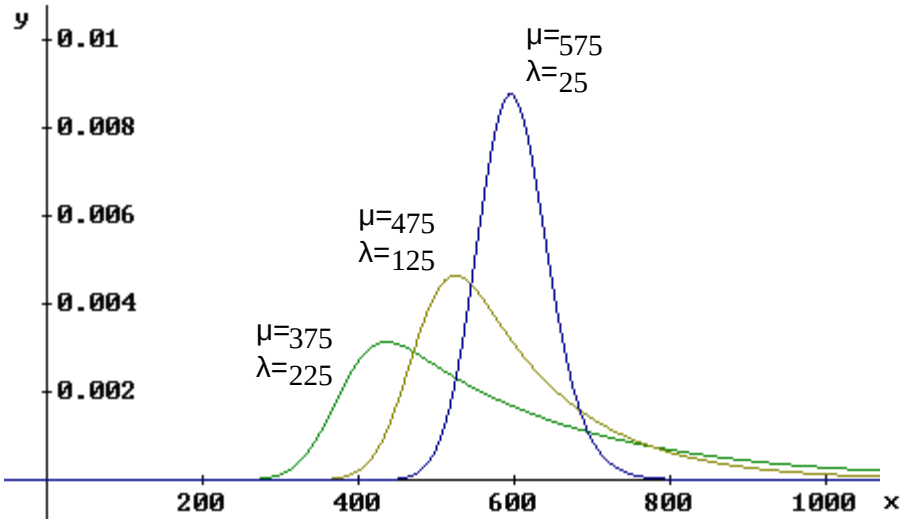

Figura 1. Representación gráfica de las tres situaciones experimentales consideradas en la distribución ex-Gaussiana, en las que se variaba el grado de asimetría, que podía ser baja (μ=575, τ=25), moderada (μ=475, τ=125) o alta (μ=375, τ=225).

En las simulaciones consideradas el número de participantes fue de 10 en un diseño intra-sujeto, que es el diseño más frecuente en los experimentos en que se mide el tiempo de reacción. El número de condiciones experimentales era de dos y el número de ensayos por condición experimental era de 10. Es importante recalcar que el patrón de resultados de las diferentes pruebas de recorte y transformación de datos con un número de participantes diferente a 10 (o con un número de ensayos diferente a 10) es análogo al indicado en las simulaciones presentadas por lo que, por brevedad, sólo indicamos el caso de 10 participantes y 10 ensayos por condición.

En las simulaciones presentadas se efectuaron 10.000 réplicas de cada situación, tanto para evaluar la situación de análisis de errores de tipo I (en la que todos los parámetros eran los mismos para los dos grupos) como para evaluar la potencia estadística de los diferentes procedimientos. Para decidir si los efectos fueron estadísticamente significativos se empleó la prueba t de

Student de dos colas con un nivel nominal α=0’05. Para obtener la potencia de los diferentes procedimientos de recorte/transformación de datos, el procedimiento seguido fue el mismo que el efectuado para obtener las tasas de errores de tipo I, con la salvedad de que se añadía cierto tamaño del efecto (25 ms.) a las puntuaciones de uno de los dos grupos. La potencia viene expresada como la proporción de casos en que se han encontrado diferencias significativas en el sentido esperado. Es decir, no se contabilizan las diferencias aparecidas en la dirección no esperada.

RESULTADOS

Error de tipo I

De acuerdo con el criterio más exigente propuesto por Bradley (1978), un procedimiento es considerado robusto cuando el valor absoluto de diferencia entre la tasa empírica del error tipo I y la tasa nominal del error de tipo I es menor o igual que dicha tasa nominal de error de tipo I dividida por diez. Ello implica que para una tasa nominal de error tipo I fuera 0’05, dicho criterio requeriría que la tasa empírica del error de tipo I se situase entre 0’045 y 0’055. No obstante, Bradley (1978) también expuso un criterio algo más liberal en el que se requería que, para una tasa nominal de 0’05, la tasa empírica del error de tipo I se situase entre 0’04 y 0’06.

Como se aprecia en la tabla 1, las tasas empíricas de error de tipo I en los diferentes procedimientos de recorte y transformación de datos se hallan muy cercanas a la probabilidad nominal de 0’05, ya sea en la distribución sin puntuaciones contaminantes (véase la tabla 1, izquierda) como en la distribución con puntuaciones contaminantes (véase la tabla 1, derecha) y se pueden enmarcar en los criterios definidos anteriormente por Bradley (1978). En todo caso, ello no es sorprendente dado que, aunque pudiera parecer extraño que a partir de una distribución asimétrica, como la ex-Gaussiana se consiga una distribución aparentemente normal cuando el número de elementos muestreados es pequeño, debemos señalar que, por el teorema del límite central, la distribución muestral de las medias obtenidas a través de los diferentes procedimientos tiende a la normalidad. Aunque Ratcliff (1993) no presentó detalladamente los resultados obtenidos en una tabla los valores obtenidos en sus simulaciones, sí que indicó que las tasas empíricas de error de tipo I en los diferentes procedimientos estaban en los márgenes de 0'04 y 0'06.

En consecuencia, la evaluación respecto a cuál es el procedimiento más deseable se basará la mayor o menor potencia estadística.

Tabla 1. Errores de tipo I para las diferentes condiciones experimentales en los diferentes procedimientos de recorte-transformación de datos.

	Sin Datos Contam.			Con Datos Contam.		
	μ=375 τ=225	μ=475 τ=125	μ=575 τ=25	μ=375 τ=225	μ=475 τ=125	μ=575 τ=25
Error de tipo I						
Media	0,046	0,051	0,050	0,048	0,049	0,044
Media Logarítmica	0,043	0,052	0,052	0,047	0,050	0,046
Media Armónica	0,046	0,050	0,052	0,047	0,050	0,048
Mediana	0,046	0,048	0,049	0,048	0,048	0,050
Media truncada (250-1000 ms)	0,050	0,051	0,050	0,049	0,052	0,048
Media truncada (250-1500 ms)	0,048	0,051	0,050	0,048	0,048	0,042
Media restringida (2 dt, cond.)	0,044	0,048	0,049	0,047	0,044	0,031
Media semi-restringida (2 dt, cond.	0,045	0,051	0,051	0,047	0,048	0,039
Media restringida (2 dt, part.)	0,048	0,047	0,048	0,047	0,048	0,036
Media semi-restringida (2 dt, part.)	0,044	0,051	0,052	0,045	0,047	0,040

Potencia

Los resultados para las distribuciones «no contaminadas» se presentan en la tabla 2 (izquierda). Como se puede apreciar, no hay grandes diferencias entre los diversos procedimientos de recorte y transformación de datos dentro de las condiciones de asimetría. Únicamente se puede remarcar que la mediana es el procedimiento menos potente. Por su parte, las transformaciones de datos (media armónica y media logarítmica) son los procedimientos que se comportan mejor, posiblemente por el menor peso asignado a las puntuaciones extremas positivas. Evidentemente, dado que τ es, en buena parte, responsable de la variabilidad de la distribución ex-Gaussiana (véase la figura 1), hay mayor potencia en los casos que τ es menor. Aunque el lector pueda pensar que en el caso de distribuciones con elevada asimetría la potencia es muy pequeña, queremos indicar que los procedimientos con más potencia estadística con un tamaño del efecto de 25 ms siguen siendo las más potentes en el caso de efectuar las simulaciones con un tamaño del efecto superior (v.g., 70 ms), por lo que, por brevedad, sólo hemos indicado el caso de un efecto de 25 ms. Hay que tener en cuenta que nuestro interés al efectuar las simulaciones se centra más en la comparación respecto a la potencia de los diferentes procedimientos bajo una condición dada más que en comparaciones de la potencia a través de poblaciones con diferente variabilidad.

Por lo que respecta a las distribuciones «contaminadas» (tabla 2, derecha), el procedimiento más potente es la media truncada con un punto de corte fijo en 1 segundo. Ello no es sorprendente dado que con este procedimiento se evita el influjo de todas aquellas observaciones mayores de 1 segundo. La mediana también se comporta bien bajo estas circunstancias (especialmente con distribuciones casi simétricas), así como las medias restringidas basadas en la media de los participantes (más que las medias

restringidas basadas en la media por condición y participante). Adicionalmente, las medias restringidas ofrecen mayor potencia que las medias semi-restringidas.

Tabla 2. Potencia para las diferentes condiciones experimentales en los diferentes procedimientos de recorte/transformación de datos con un tamaño del efecto de 25 ms.

	Sin Datos Contam.			Con Datos Contam.		
	μ=375 τ=225	μ=475 τ=125	μ=575 τ=25	μ=375 τ=225	μ=475 τ=125	μ=575 τ=25
Potencia						
Media	0,103	0,225	0,872	0,068	0,089	0,102
Media Logarítmica	0,142	0,280	0,884	0,108	0,151	0,218
Media Armónica	0,177	0,323	0,887	0,147	0,216	0,371
Mediana	0,107	0,241	0,789	0,098	0,206	0,733
Media truncada (250-1000 ms)	0,140	0,264	0,873	0,132	0,246	0,787
Media truncada (250-1500 ms)	0,112	0,227	0,872	0,099	0,177	0,430
Media restringida (2 dt, cond.)	0,108	0,228	0,843	0,083	0,155	0,454
Media semi-restringida (2 dt, cond.)	0,104	0,230	0,874	0,071	0,100	0,130
Media restringida (2 dt, part.)	0,126	0,263	0,875	0,102	0,195	0,553
Media semi-restringida (2 dt, part.)	0,113	0,238	0,873	0,079	0,121	0,185

DISCUSIÓN

Los resultados de las simulaciones muestran, en primer lugar, que los diferentes procedimientos de recorte y transformación de datos empleados en psicología cognitiva guardan unos niveles de probabilidad de tipo I muy próximos a los nominales (véase también Ratcliff, 1993). Ello es así incluso cuando la distribución teórica de los datos a partir de los cuales se computan los promedios para la prueba estadística se halla relativamente lejos de la distribución normal (con clara asimetría positiva y con la inclusión de cierto número de observaciones contaminantes).

Por lo que respecta a la potencia estadística, la situación es compleja. La elección de la prueba a emplear depende, por una parte, de la asimetría de la distribución de datos y, por otra, de la presencia o no de observaciones contaminantes. El mejor ejemplo viene dado por la mediana, que es una prueba muy potente en las diferentes distribuciones «contaminadas», pero que es poco potente en las distribuciones «no contaminadas» (especialmente en las casi simétricas), mientras que la mediana resulta muy potente cuando la distribución «contaminada» es casi simétrica, pero no lo es tanto cuando la distribución es marcadamente asimétrica. Por otra parte, la potencia obtenida con una técnica de transformación de datos como la media armónica (que es preferible a la media logarítmica) es elevada, especialmente en las distribuciones marcadamente asimétricas, y es una prueba recomendable en diseños unifactoriales. Sin embargo, no recomendamos su empleo en el caso de diseños factoriales, dado que, al ser una técnica de transformación de datos,

podría hacer aparecer o desaparecer una posible interacción entre los factores (véase Heathcote et al., 1991).

Es importante remarcar que una técnica muy empleada en la literatura psicológica como el uso de medias truncadas se revela como el procedimiento más potente a la hora de evitar la influencia de las puntuaciones atípicas (véase también Ratcliff, 1993) y que es aconsejable en todas las condiciones de asimetría analizadas. No obstante, debe señalarse que la potencia depende de los puntos de corte empleados, y que la potencia disminuye claramente a medida que empleamos puntos de corte más extremos. Además, un posible inconveniente de este procedimiento es que los diferentes participantes podrían variar en su rapidez en responder, de manera que algunos tengan promedios de 500 ms. mientras que otros tengan promedios de 850 ms. En tales casos, una observación de 1281 ms. es posiblemente una observación atípica para el primer participante, pero no tanto para el segundo. Para evitar tal problema cabe emplear bien un punto de corte más alto, con la consiguiente pérdida en potencia como se observa en las simulaciones, o bien emplear un punto de corte que sea dependiente de la rapidez de cada sujeto: es decir, el empleo de medias restringidas.

De hecho, las simulaciones realizadas han mostrado que el empleo de las medias restringidas basadas en la media de los participantes es un procedimiento potente en presencia de observaciones contaminantes. A este respecto, las medias restringidas basadas en las medias de los participantes se muestran más potentes que las medias restringidas basadas en las medias de los participantes por condición, por lo que han de preferirse a éstas. Además, cabe indicar que las medias restringidas (ya sean por participante, o por participante y condición) se muestran claramente más potentes que las medias semi-restringidas (es decir, aquellas en las que las observaciones que están más allá de los puntos de corte se reemplazan por los propios puntos de corte). Un posible inconveniente del empleo de las medias restringidas reside en que se eliminarán aproximadamente el 5% de los datos con lo que es muy posible que estemos eliminando datos extremos que no sean propiamente datos contaminantes (es decir, procedentes de otra distribución de datos). Ello es, lógicamente, una posibilidad en el caso del empleo de medias truncadas, si bien cuando los puntos de corte están claramente alejados de la posible media poblacional, por ejemplo 1’5 segundos en una tarea de decisión léxica, es muy poco probable que las puntuaciones que rebasen tales puntos de corte se deban a los procesos bajo estudio en el experimento.

Con el fin de ilustrar las diferencias en potencia entre las diferentes técnicas de recorte y transformación de datos, hemos procedido a efectuar un análisis con los diferentes procedimientos de recorte y transformación de datos sobre un experimento reciente en nuestro laboratorio. En dicho experimento participaron 26 estudiantes, con 2 condiciones (20 observaciones por condición). La tarea era la decisión léxica (en la que los participantes habían de indicar si una serie de letras era una palabra o no) y se recogía el tiempo de reacción (Perea, Gotor y Nácher, 1999). Dado que los participantes habían de contestar en valenciano, cuando su lengua de escolarización en enseñanza primaria había sido el castellano, era de esperar que los

participantes muestren cierta proporción de puntuaciones atípicas. Efectivamente, como se indica en la tabla 3, si el análisis se efectúa con la media aritmética incluyendo todos los datos, las diferencias no son estadísticamente significativas. Sin embargo, efectuar un análisis de tal manera conduce muy probablemente a un error de tipo II, al incluir cierto número de observaciones atípicas en el análisis. De hecho, las diferentes técnicas de recorte/transformación de datos muestran que hay una diferencia entre condiciones (aproximadamente 40 ms), que es bastante elevada para una tarea de decisión léxica. Cabe señalar varios aspectos. Por una parte, las diferencias de medias son muy similares con los diversos procedimientos, pero los valores del estadístico de contraste difieren claramente entre los diferentes procedimientos. Por otra parte, las dos técnicas más empleadas por los investigadores, esto es, las medias truncadas y las medias restringidas por participantes, son las que ofrecen los valores menores de p. Estas técnicas también se comportaban con una elevada potencia en las simulaciones. Finalmente, cabe señalar que, al igual que ocurría en las simulaciones, las medias restringidas por participantes han sido más potentes que las medias restringidas por participante y condición, así como que las medias restringidas han sido más potentes que las medias semi-restringidas.

Tabla 3. Valores empíricos observados (t), valores p obtenidos (en prueba de dos colas) y diferencia de medias (en ms) entre condiciones en el experimento de Perea, Gotor y Nácher (1999).

	t	p	Dif. medias
Media	1'45	0'1596	43
Media Logarítmica	2.33	0'0283	38
Media Armónica	2'90	0'0077	33
Mediana	2'58	0'0161	36
Media truncada (250-1500 ms)	4'22	0'0003	40
Media restringida (2 dt, cond.)	2'92	0'0072	46
Media semi-restringida (2 dt, cond.)	1'96	0'0605	46
Media restringida (2 dt, part.)	3'53	0'0016	41
Media semi-restringida (2 dt, part.)	2'93	0'0071	47

En todo caso, debemos recalcar que el estudio de las puntuaciones atípicas puede ser importante por sí mismo. El hecho de que haya más puntuaciones atípicas en una condición que en otra es una señal de que los procesos cognitivos pueden diferir entre ambas condiciones, lo que implica que tales tiempos de reacción no parecen ser debidos a procesos aleatorios externos a los procesos bajo estudio, sino que pueden deberse a una distribución que pudiera ser de interés para el investigador. Por ello, creemos que los investigadores deberían indicar el porcentaje de casos de datos eliminados en cada condición experimental cuando emplean las medias restringidas o las medias truncadas. Por ejemplo, la distribución subyacente para una condición en que se midan los tiempos de reacción ante palabras de

alta frecuencia tiene una asimetría menor que la distribución para las palabras de baja frecuencia en la tarea de decisión léxica (véase Balota y Spieler, 1999). Ello hace que ambas condiciones no sólo se distingan por la situación del grueso de los datos (en donde la distribución de las palabras de alta frecuencia se halla más a la izquierda que la de las palabras de baja frecuencia), sino que se distinguen por el grado de asimetría, el cual es debido posiblemente a procesos de tipo decisional que ocurren en palabras de baja frecuencia. En consecuencia, habrá más observaciones eliminadas (empleando medias truncadas con un límite superior bastante alto, v.g., 2 segundos) ante palabras de baja frecuencia que ante palabras de alta frecuencia.

Por todo ello, una opción que puede complementar los anteriores análisis sobre la tendencia central es la realización de un análisis de la forma de la distribución de datos. Desgraciadamente, hace falta un número bastante elevado de datos por condición experimental para efectuar tales análisis, lo que no suele ser el caso en muchos trabajos en psicología cognitiva (v.g., en psicología del lenguaje, por impedimentos derivados de la selección del material estimular). No obstante, en el caso de tener un número escaso de datos por condición experimental (v.g, 15-20 datos) una opción muy deseable es la «vincentización» (véase Balota y Spieler, 1999; Heathcote, 1996; Ratcliff, 1979), en la que a partir de cierto número de cuantiles de cada participante (v.g., los 9 deciles) y de sus promedios obtenidos a través de los participantes, se puede obtener un histograma que guarda la forma de la distribución de los participantes.

Una ilustración de tal punto lo tenemos en el reciente estudio sobre los efectos de secuencia efectuado por Perea y Carreiras (1999), en el que se hipotetizaba que los participantes variaban el criterio de respuesta para las respuestas a las palabras según las características del ensayo previo, de manera que los criterios de respuesta para el ensayo N eran más altos cuando el ensayo anterior era una pseudopalabra (v.g., CALINO) que cuando el ensayo anterior era una palabra muy frecuente (v.g., CIUDAD). El análisis estadístico "estándar" sobre las medias por condición experimental reveló la existencia de un efecto de secuencia del estatus léxico del ensayo previo. No obstante, para observar tal posible cambio de criterio, se puede acudir a un histograma "vincentizado" que, bajo la hipótesis de los autores, debería mostrar que la distribución de datos en la condición de "palabra en el ensayo previo" estuviera más a la izquierda que la distribución de datos en la condición de "pseudopalabra en el ensayo previo". De hecho, esto fue lo que ocurrió (véase la figura 2). El comienzo de la distribución de "pseudopalabra en el ensayo previo" está más allá de los 500 ms, mientras que el comienzo de la distribución es claramente menor en el caso de que hubiera una palabra en el ensayo previo. Un excelente ejemplo de cómo efectuar un análisis distribucional, mediante el procedimiento de «vincentización» y el posterior ajuste con los parámetros de la distribución ex-Gaussiana, puede verse en el reciente trabajo de Balota y Spieler (1999).

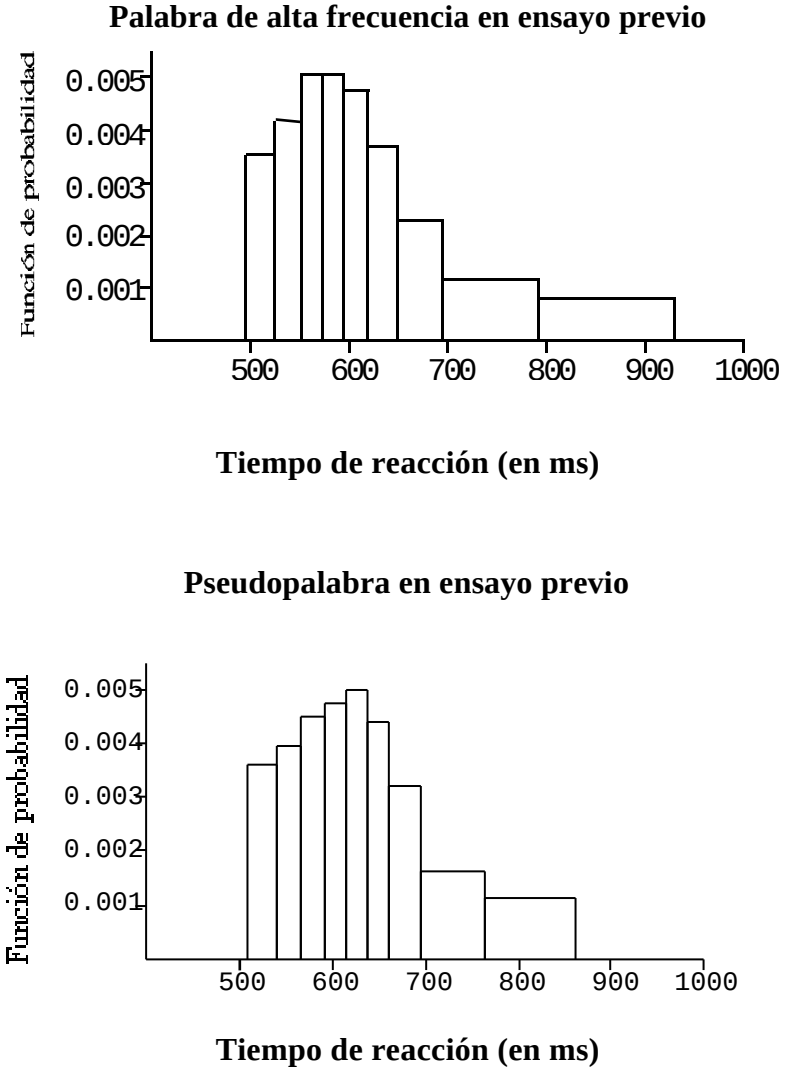

Figura 2. Histogramas «vincentizados» a partir de los datos del experimento primero de Perea y Carreiras (1999).

En definitiva, la elección de cuál es el mejor procedimiento para analizar los datos procedentes de un experimento en que se mida el tiempo de reacción no es una cuestión fácil de resolver. En todo caso, dos técnicas muy empleadas por los investigadores en la actualidad (el empleo de medias truncadas y medias restringidas) son, como se ha visto en las simulaciones y en el análisis del experimento, procedimientos estadísticamente potentes. Ello no obsta para que, además de analizar si los porcentajes de datos excluidos son similares en cada una de las condiciones experimentales, los

investigadores efectúen un análisis más completo sobre las distribuciones de datos, más allá de indicar una medida de tendencia central y una medida de variabilidad. En todo caso, creemos que los investigadores deben ser especialmente cuidadosos en los análisis de los tiempos de reacción en sus experimentos, más cuando los criterios empleados para recortar/transformar sus distribuciones son criterios *ad hoc* y, por consiguiente, la efectividad de los procedimientos de recorte/transformación dependerá de las características de la distribución bajo estudio y de esa otra distribución contaminante que el investigador habitualmente supone que es causada por procesos externos a los procesos bajo estudio. Finalmente, pensamos que futuros trabajos deberían ocuparse más a fondo del análisis de técnicas más sofisticadas para la detección de puntuaciones contaminantes en la experimentación con datos psicológicos, dado que los análisis de datos que reflejan relaciones modestas entre variables (como suelen ser las relaciones entre variables en psicología) son los más vulnerables a la influencia de las puntuaciones atípicas (véase Judd, McClelland y Culhane, 1995 para una revisión de tales procedimientos).

ABSTRACT

Statistical power with different methods for dealing with reaction time outliers: A Monte Carlo investigation. We explored the effect of outliers on reaction time analyses. Monte Carlo simulations were used to investigate the influence of a number of common trimming and transformation techniques to improve power in within-subject designs in which multiple observations are collected for each participant in each condition. Correction for outliers was useful in increasing power, especially with truncated means (i.e., exclusion of the data outside a given range). We propose that an analysis of outliers should be conducted to find out whether the percentage of excluded observations is similar across conditions. In addition, graphical methods (e.g., Vincent averaging) appear to complement the analyses based on mean reaction times.

Key words: simulation, power, reaction-time analysis, outliers.

REFERENCIAS

Anscombe, F.J. (1960). Rejection of outliers. *Technometrics, 2*, 123-147.

Balota, D. A. y Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition. *Journal of Experimental Psychology: General, 128*, 32-55.

Barnett, V. y Lewis, T. (1995). *Outliers in statistical data*. New York: Wiley.

Box, G. E. P. y Müller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29*, 610-611.

Bush, L. K., Hess, U. y Wolford, G. (1993), Transformation for within-subject designs: A Monte Carlo investigation. *Psychological Bulletin, 113*, 566-579.

Bradley, J. V. (1978). Robustness*? British Journal of Mathematical and Statistical Psychology, 34*, 144-152.

Forster, K. I., & Veres, C. (1998). The prime lexicality effect: Form-priming as a function of prime awareness, lexical status, and discrimination difficulty. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 498-514.

Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments, and Computers, 28*, 427-445.

Heathcote, A., Popiel, S.J. y Mewhort, D.J.K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin, 109*, 340-347.

Hockley, W.E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6*, 598-615.

Judd, Ch. M., McClelland, G. H. y Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology, 46*, 433-465.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 539-543.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *Quarterly Journal of Experimental Psychology, 43A*, 907-912.

Perea, M. (1999). Tiempos de reacción y psicología cognitiva: Dos procedimientos para evitar el sesgo debido al tamaño muestral. *Psicológica, 20*, 13-21.

Perea, M. y Carreiras, M. (1999). *Sequential effects in the lexical decision task: The role of the item-frequency of the previous trial*. Enviado para publicación.

Perea, M., Gotor, A. y Nácher, M. J. (1999). Efectos de facilitación asociativa entre lenguajes con pares no cognaticios: Evidencia empleando una breve asincronía estimular. *Revista de Psicología Universitas Tarraconensis*. En prensa.

Ponsoda, V. y Alcázar, M. A. (1996). Reaction time analysis with outlier exclusion: A simple method providing bias nearly independent of sample size. *Psicológica, 17*, 31-40.

Plourde, C. E. y Besner, D. (1997). On the locus of the word frequency in visual word recognition. *Canadian Journal of Experimental Psychology, 51*, 181-194.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86*, 446-461.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510-532.

Ulrich, R. y Miller, J. (1994). Effects of outlier exclusion on reaction time analysis. *Journal of Experimental Psychology: General, 123*, 34-80.

Wichmann, B.A. y Hill, J.D. (1984). An efficient and portable pseudo random number generator. *Applied Statistics, 33*, 123.

Zumbo, B. D. y Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51*, 139-149.

(Revisión aceptada: 6/7/99)