Documentos CEDE
ISSN 1657-7191 edición electrónica

Detección de copia en pruebas del Estado

Diego Jara
Álvaro Riascos
Mauricio Romero

15
MAYO DE 2010

Universidad de los Andes
Facultad de Economía

CEDE
Centro de Estudios
sobre Desarrollo Económico

Serie Documentos Cede, 2010-15
ISSN 1657-7191

Mayo de 2010

Calle 19A No. 1 – 37, Bloque W.
Bogotá, D. C., Colombia
Teléfonos: 3394949- 3394999, extensiones 2400, 2049, 3233
infocede@uniandes.edu.co
http://economia.uniandes.edu.co

Ediciones Uniandes
Carrera 1ª Este No. 19 – 27, edificio Aulas 6, A. A. 4976
Bogotá, D. C., Colombia
Teléfonos: 3394949- 3394999, extensión 2133, Fax: extensión 2158
infeduni@uniandes.edu.co

Edición, diseño de cubierta, preprensa y prensa digital:
Proceditor ltda.
Calle 1ª C No. 27 A – 01
Bogotá, D. C., Colombia
Teléfonos: 2204275, 220 4276, Fax: extensión 102
proceditor@etb.net.co

Impreso en Colombia – *Printed in Colombia*

Universidad de los Andes
Facultad de Economía

CEDE
Centro de Estudios
sobre Desarrollo Económico

DETECCIÓN DE COPIA EN PRUEBAS DEL ESTADO

Diego Jara[*] Álvaro Riascos*[†] Mauricio Romero*

Resumen[‡]

Este trabajo implementa la metodología de detección de copia en exámenes de múltiples opciones expuesta en Sotaridona, van der Linden y Meijer [2006] basada en el índice Kappa de Cohen. Presentamos los resultados de aplicar esta metodología a las pruebas del Estado ICFES 2009 II. Este es, en el conocimiento de los autores, el primer estudio en la literatura colombiana que busca proveer de herramientas cuantitativas de carácter estadístico a los administradores de las pruebas del Estado (y posiblemente muchas otras) para detectar casos sospechosos de fraude. Con este apoyo se pueden generar alertas y priorizar investigaciones detalladas caso por caso. Desde el punto de vista de la literatura académica especializada, el aporte del trabajo es evaluar, con datos reales, las bondades del índice Kappa propuesto por los autores y que ellos validan con datos artificiales generados mediante la simulación de un modelo de respuesta nominal. En particular, los resultados ponen en evidencia que, si bien el índice es muy valioso en la práctica, éste adolece de problemas de especificación difíciles de superar utilizando datos reales. Esto alerta sobre la necesidad de progresar en la implementación de metodologías más apropiadas para complementar la labor de supervisión y monitoreo desarrollada por los administradores de tales exámenes. Adicionalmente, este trabajo tiene como objetivo parcial servir de motivación y dar un primer paso para ahondar en esta área del conocimiento.

Palabras clave: Índice Kappa, copia de respuestas, ICFES, trampa, exámenes de opción múltiple.

Clasificación JEL: C19, I20.

[*] Quantil | Matemáticas Aplicadas.

[†] Facultad de Economía, Universidad de los Andes.

[‡] Autor corresponsal: A. Riascos (alvaro.riascos@quantil.com.co). Agradecemos los comentarios y sugerencias de Julian Mariño y el equipo de estadística de la subdirección académica del ICFES. Por su excelente colaboración en las etapas iniciales de este estudio agradecemos a Camilo Pabon. Este trabajo fue posible gracias la financiación del Instituto Colombiano para la Educación Superior - ICFES.

DETECTING ANSWER COPYING IN STANDARDIZED TESTS

Diego Jara Álvaro Riascos Mauricio Romero

Abstract

This paper implements the answer-copying detection methodology for multiple-choice tests developed in Sotaridona, van der Linden and Meijer [2006], based on Cohen's Kappa index. We present results of this methodology applied to the standardized test for college admissions in Colombia that took place in the second semester of 2009. This is, to the authors' knowledge, the first study in Colombia that seeks to provide quantitative tools, of statistical nature, to detect fraud. This paper also evaluates, with real data, the performance of the Kappa index proposed in the referenced paper, which had only been evaluated using simulated data generated with a nominal item response model. In particular, we find evidence that there are misspecification problems that are difficult to overcome when using real data. We hope this paper motivates further applied research in this field in Colombia

Key words: Kappa Index, answer copying, ICFES, cheating, multiple-choice test.

JEL Classification: C19, I20.

1. Introducción

Los exámenes de Estado de Colombia, diseñados, ofrecidos y administrados por el Instituto Colombiano de Fomento a la Educación Superior (ICFES), tienen un formato de selección múltiple. En efecto, la mayoría de los módulos son compuestos por preguntas[1] para las cuales existen cuatro posibles respuestas, una de las cuales es la correcta. Las ventajas de este formato son : mayor eficiencia en la evaluación de los conocimientos medidos en los estudiantes, reducción de elementos subjetivos en la calificación, y simplificación en la logística administrativa. Sin embargo, existen desventajas también: incapacidad de medir conocimientos en algunos temas, exclusión de conocimientos parciales (debido a la naturaleza binaria de la respuesta), inclusión de un elemento aleatorio y especialmente, en casos en que el examen se ofrece a muchos estudiantes de forma simultánea en un mismo salón, como es el caso de los exámenes de Estado en Colombia, la posibilidad de copia. Todas excepto la última son características estructurales de los exámenes de selección múltiple; sin embargo, el elemento de copia puede ser atacado de forma preventiva, o investigado posterior a la presentación del examen. Este trabajo estudia formas en que la copia puede investigarse con base en los exámenes presentados.

Existen varias formas en que la copia puede darse en exámenes de selección múltiple, pero la más sencilla y frecuente es la que ocurre al nivel de una pareja de estudiantes sentados contiguamente, en donde uno (el "sospechoso de copia") escribe una o varias respuestas de forma idéntica al otro (la "fuente"). Es precisamente este tipo de copia que se analiza en el presente trabajo, para lo cual se cuenta con información de las respuestas de todos los estudiantes en cada examen, y de la partición del conjunto de estudiantes en salones distintos. En la actualidad no se cuenta con información más precisa de la ubicación exacta de cada estudiante en cada salón, por lo cual se supone que cada pareja en un mismo salón es candidata para presentar la copia descrita, pero puede esperarse que en un futuro se tenga esta información, lo cual incrementaría la eficiencia en el análisis de los datos.

A pesar de que la motivación para construir una metodología que alerte sobre la posibilidad de copia es clara para el administrador del examen, y para sus usuarios, tales

[1] En realidad, en muchos casos cada elemento no constituye como tal una pregunta, pero se usará este término de forma genérica.

como los centros educativos que incorporan los resultados en sus decisiones de aceptación de alumnos, las siguientes observaciones permiten sospechar que la presencia de copia en estos exámenes es real, y merece un algoritmo de detección confiable.

1.1. Observaciones de casos de pruebas idénticamente contestadas

Se analizaron 36 formas distintas (algunas materias tienen más de una forma, y éstas se consideran exámenes separados). En los datos se observan cerca de 10,000 parejas que tenían todas las respuestas iguales (si tienen todas las respuestas iguales en dos formas o más, se contabiliza el total de formas, luego una pareja pudo haber sido contada más de una vez; asimismo, se considera el orden de la pareja: es distinta una pareja con los mismos individuos, si el "sospechoso" y la "fuente" cambian de una pareja a la otra); vale la pena notar que no se consideran como casos de copia aquéllos en donde todas las respuestas contestadas eran correctas. Entre los involucrados en estas parejas, existen más de 1787 individuos que tenían al menos dos formas idénticas a alguien más (o la misma forma idéntica a al menos otros dos individuos). Se presentan casos interesantes como los siguientes; cabe notar que respuestas omitidas o contestadas mal se incluyen en los casos de respuestas idénticas, luego algunos resultados pueden sobreestimar los números reales si se ajustara por este tipo de respuestas, aunque este efecto es relativamente pequeño.

- Hay casos de aparente copia masiva, en las que muchos estudiantes coinciden en todas las respuestas de una forma, en un mismo salón. Por ejemplo, hay casos de 17 estudiantes de un mismo salón con respuestas idénticas para la forma 100 (6 bien de 15); 15 en la forma 43 (6 bien de 15); 13 en las formas 42 (15 bien de 24), 43 (6 bien de 15) y 100 (6 bien de 15).
- Hay un caso de cuatro personas con las mismas respuestas en tres formas. Asimismo, se observan dos casos de tres personas con respuestas idénticas en cuatro formas.
- Un grupo de 13 personas (del salón 437 788) tuvieron respuestas idénticas en dos formas (la 43 y la 100). Adicionalmente, hubo otras cuatro personas con idénticas respuestas en la forma 100, y otras dos en la forma 43. El número de respuestas correctas en las dos formas fue de 6 entre 15 preguntas.
- Hay 14 personas involucradas en 30 casos de respuestas idénticas. 13 de éstas forman parte del caso masivo mencionado en el punto anterior.

- Hay más de 35 parejas que tuvieron respuestas idénticas en 4 formas, nueve con respuestas idénticas en 5 formas, y hubo incluso dos con respuestas idénticas en 6 formas.

Específicamente, este trabajo describe un algoritmo de detección de copia en exámenes estandarizados de múltiples opciones basado en el índice Kappa de (Cohen, 1960). La adaptación, descripción e implementación de éste como un algoritmo computacional programado en R[2], siguen de cerca a (Sotaridona, van der Linden, & Meijer, Detecting Answer Copying uisng Kappa Statistic, 2006).

El documento hace una rápida revisión de la literatura y describe con detalle la versión más sencilla del índice Kappa. En ésta, la probabilidad conjunta de las respuestas de dos individuos por pregunta es invariante entre preguntas. Es decir, la hipótesis es que la distribución multinomial conjunta de cualquier dos individuos es la misma para cada pregunta. Posteriormente explicamos cómo modificar el algoritmo utilizando una técnica de recodificación y cómo condicionar a la habilidad de los individuos. En ausencia de una medida de habilidad adecuada de los individuos, una alternativa es aproximar ésta por el número de respuestas correctas.

El trabajo está dividido en cinco secciones, siendo esta introducción la primera. En la segunda sección se hace una breve revisión de la literatura para poner en contexto la metodología utilizada en este trabajo. La tercera sección presenta tres versiones del índice Kappa. La primera versión explica la idea básica del índice y sus limitaciones, dadas las hipótesis fuertes que se hacen. La segunda describe una técnica (recodificación) para mitigar estas limitaciones prácticas, y la última es la versión completa que incorpora los elementos de las dos primeras y condiciona la detección de copia a la habilidad del individuo de quien se sospecha. La cuarta sección muestra esquemáticamente los pasos que se siguieron para la programación y sirve como una introducción a la estructura de la programación del algoritmo. La quinta sección presenta algunos resultados y conclusiones de la aplicación del algoritmo a las pruebas del Estado 2009-2.[3]

[2] Véase *The R Project for Statistical Computing*: http://www.r-project.org/
[3] El lector puede consultar: www.quantil.com.co/investigacion/copia/ para mayor información sobre la implementación de la metodología en R.

2. Revisión de la literatura

La mayoría de los modelos de detección de copia existentes (van der Linden & Sotaridona, 2006) se basan en calcular diferentes estadísticos que cuentan el número de respuestas comunes entre el individuo sospechoso de hacer copia y aquel de quien copia (la fuente). Estos estadísticos difieren principalmente en: (1) El conjunto de respuestas sobre las que se calcula (todas las respuestas, correctas e incorrectas, solo las respuestas incorrectas, solo las respuestas incorrectas condicional a la habilidad de los dos individuos, etc.); (2) La distribución que se supone tiene este estadístico y si se usa o no alguna estandarización del estadístico relevante.

Por ejemplo, el índice K se basa en el cálculo del estadístico para las respuestas incorrectas comunes entre dos exámenes condicional al número de respuestas incorrectas del individuo sospechoso. Este mismo tipo de estadístico, basado en las respuestas incorrectas, es usado en la gran mayoría de las propuestas que existen en la literatura como el índice B en (Angoff, 1974), el índice propuesto por (Belleza & Belleza, 1989) y los índices S_1, S_2 de (Sotadoridona & Meijer, 2003).

Otra alternativa interesante se basa en la comparación de la totalidad de las respuesta comunes (correctas o no). Este es el caso del índice ω (Sotaridona & Mejier, Statistical Properties of the K-Index for Detecting Answer Copying, 2001) y varios de los índices estudiados por (Angoff, 1974) así como el el índice g_2 (Frary, Tideman, & Watts, 2009).

Sin embargo, la diferencia más importante entre todas estas metodologías está en la distribución que se supone tienen estos estadísticos. El índice K (o más precisamente, la aproximación del índice K propuesta por (Holland, 1996) y sus diferentes variaciones propuestas por (Sotaridona & Mejier, Statistical Properties of the K-Index for Detecting Answer Copying, 2001), así como la metodología utilizada por el ICFES, están relacionadas con la distribución binomial. El índice ω esta basado en una distribución que se deduce de un modelo de respuesta nominal. Apelando a alguna versión del teorema central del límite se deduce una distribución normal para el estadístico. Los índices S_1, S_2 de (Sotadoridona & Meijer, 2003) están basados en una distribución de Poisson. La versión del índice Kappa que aquí se presenta se basa en una distribución no paramétrica de la

probabilidad con la que los individuos responden una pregunta específica y es la base de la recodificación de las respuestas de los individuos que se describe con detalle más adelante. Sin embargo, la metodología apela al teorema centra del límite para deducir la distribución asintótica del estadístico de interés.

En conclusión, las características mencionadas en el primer párrafo sirven como un principio unificador y clasificador de gran parte de la literatura teórica y aplicada existente.

3. Índice de detección de copia cuando la función de respuesta es desconocida (el índice Kappa)

3.1. Introducción

Una característica importante del índice Kappa es que no se basa en un modelo de respuesta nominal. Es decir, para su aplicación no es necesario primero estimar la probabilidad de que los individuos respondan con cierta probabilidad a una pregunta. Las hipótesis básicas son:

1. La función de respuesta a una pregunta es probabilística.
2. Es posible seleccionar una muestra en la cual las respuestas son independientes (en particular, no hay copia).

En la primera versión que presentamos del índice Kappa se hace una hipótesis muy fuerte. Ésta es que la probabilidad conjunta de dos individuos de sus respuestas es igual para todas las preguntas. Esta hipótesis es fundamental para deducir la probabilidad de las coincidencias observadas entre dos individuos y la distribución asintótica bajo la hipótesis nula del índice Kappa. Esta es una hipótesis muy fuerte y por esta razón se utiliza una técnica de recodificación de las respuestas para mitigar el problema[4]. Las bondades de esta metodología de recodificación han sido validadas mediante simulaciones en la cuales los individuos se suponen que responden con base en un modelo de respuesta nominal como el de Bock (1972)[5]. Específicamente, (Sotaridona, van der Linden, & Meijer, Detecting

[4] Denominada *Unconditional Recoding of the Responses.* Véase página 420 de (Sotaridona, van der Linden, & Meijer, Detecting Answer Copying uisng Kappa Statistic, 2006)
[5] Véase (van der Linden & Hambleton, 1997)

Answer Copying uisng Kappa Statistic, 2006) demuestran mediante un análisis de simulaciones, que la metodología que se presenta en este documento comete errores de tipo I comparables a los de un modelo en donde se conoce con exactitud el modelo probabilístico de respuesta (modelo de respuesta nominal de Bock). Adicionalmente, muestran que la prueba basada en el estadístico Kappa tiene una buena potencia. Esto sugiere que las diferentes variantes de la metodología, que aquí se dejan para un desarrollo posterior, van a ser, en el mejor de los casos, marginalmente superiores a la metodología base que a continuación se discute.

Como en toda prueba estadística, existe la posibilidad de cometer un error de Tipo I (la hipótesis nula siendo que los dos individuos no se copiaron). Esto es, con un nivel de confianza del 99%, el 1% de la población "inocente" va a ser típicamente sospechoso de copia. Adicionalmente, el rechazo de la hipótesis nula no excluye otras razones distintas a la copia entre los individuos; por ejemplo, ambos individuos pudieron responder de la misma forma debido a que han estado expuestos a una enseñanza común. En cualquier caso, como todo modelo, éste presenta limitaciones similares a los modelos existentes en la literatura.

3.2. Índice Kappa: versión simplificada

Para comenzar, se supone que tenemos un cuestionario con N preguntas, n posibles respuestas a cada pregunta y dos estudiantes c (sospechoso de copia) y s (sospechoso de ser la fuente) que responden de forma independiente cada pregunta (no se copian). Los resultados de las respuestas de cada estudiante se pueden representar por una matriz nxn de la siguiente forma. Supongamos que las filas corresponden a respuestas de c y las columnas a respuestas de s. La fila i se llena de la siguiente forma. Se seleccionan todas las preguntas en las que c respondió la opción i (i es una de las posibles respuestas a una pregunta). En esas preguntas se determina el número de veces que s respondió cada una de las n alternativas y se llenan las respectivas celdas. Luego, en la celda (i,i) (en la diagonal) van aparecer el número de respuestas coincidentes entre c y s con repuesta común i y en una celda (i,j) (por fuera de la diagonal), va aparecer el número de veces que s respondió la opción j en las preguntas en que c respondió la opción i.

Sea π_{vv} la probabilidad de que una pregunta sea clasificada en la celda (v, v) y π_{v+}, π_{+v} la probabilidad de que la pregunta sea clasificada en la fila o columna v respectivamente. La probabilidad observada de que los individuos coincidan en por lo menos una pregunta, π_o es:

$$\pi_0 = \sum_{\mathrm{v}} \pi_{\mathrm{vv}}$$

Ahora, bajo la hipótesis de que la distribución conjunta de c y s de responder las diferentes opciones es igual para todas las preguntas se puede demostrar que la probabilidad de que coincidan en por lo menos una pregunta π_e es:

$$\pi_e = \sum_{v} \pi_{v+}\, \pi_{+v}$$

El estadístico κ (Kappa) se define como:

$$\kappa = \frac{\pi_{\mathrm{o}} - \pi_{\mathrm{e}}}{1 - \pi_{\mathrm{e}}},$$

que es un indicador del grado de coincidencia en las respuestas de dos individuos controlando por lo que sería normal por azar. Obsérvese que el índice κ es creciente en ambas variables. Entre más cercano a cero, menor es la probabilidad de que c se haya copiado de s.[6]

La prueba estadística para detectar sospechosos de copia se basa en la siguiente hipótesis:

$$H_0: \pi_{\mathrm{o}} = \pi_{\mathrm{e}}$$

$$H_1: \pi_{\mathrm{o}} \neq \pi_{\mathrm{e}},$$

o de forma equivalente:

$$H_0: \kappa = 0$$
$$H_1: \kappa \neq 0.$$

[6] Con las hipótesis hasta ahora mencionadas el índice de la pareja ordenada (c,s) es igual al índice de la pareja ordenada (s,c). La diferencia entre los dos se hace evidente más adelante que se condicione el cálculo del índice a la habilidad de c.

Con las hipótesis mencionadas hasta el momento se puede obtener la distribución asintótica de κ, en el número de preguntas N: sea $\hat{\kappa}$ el estadístico que se obtiene de sustituir las probabilidades por las frecuencias observadas (proporciones en la diagonal, fila y columna). Entonces asintóticamente $\hat{\kappa}$ se distribuye $N(\mu, \sigma^2)$ donde:

$$\mu = \kappa,$$
$$\sigma^2 = \frac{1}{N}\left(\frac{\pi_0(1-\pi_0)}{(1-\pi_e)^2} + a + b\right),$$

y a y b son las siguientes constantes:

$$a = \frac{2(1-\pi_0)(2\pi_0\pi_e - \sum_{\mathrm{v}} \pi_{\mathrm{vv}}\,(\pi_{\mathrm{v}+} + \pi_{+\mathrm{v}}))}{(1-\pi_{\mathrm{e}})^3}$$
$$b = \frac{(1-\pi_0)^2(\sum_v \sum_{\mathrm{v}'} \pi_{\mathrm{vv}'}\,(\pi_{\mathrm{v}+} + \pi_{+\mathrm{v}'})^2 - 4\pi_e^2)}{(1-\pi_{\mathrm{e}})^4}.$$

Ahora, en el caso específico de la prueba de hipótesis mencionada arriba, $\hat{\kappa}$ se distribuye asintóticamente $N(0, \sigma^2)$, donde las ecuaciones anteriores se reducen a:

$\mu = \kappa,$

$$\sigma^2 = \frac{1}{N}\left(\frac{1}{(1-\pi_e)^2}\right)\left(\pi_e(1-\pi_e) + \sum_v \sum_{\mathrm{v}'} \pi_{\mathrm{vv}'}\,(\pi_{\mathrm{v}+} + \pi_{+\mathrm{v}'})^2 - 2\sum_{\mathrm{v}} \pi_{\mathrm{v}+}\pi_{+\mathrm{v}}(\pi_{\mathrm{v}+} + \pi_{+\mathrm{v}})\right).$$

Por simplicidad se estandariza la prueba. Sea $Z = \frac{\kappa-\mu}{\sigma}$; luego con la hipótesis nula se sigue que Z se distribuye asintóticamente $N(0,1)$. Dada una significancia α para la prueba el valor crítico, z se determina a partir de (por ejemplo para un nivel de confianza del 95% - prueba muy liberal, z=1.645):

$$P(Z \geq z) = \alpha$$

En el análisis realizado sobre todas las pruebas del Estado se utilizó un nivel de confianza del 99.9%.

Un supuesto clave en la derivación de esta prueba es que se supone que las probabilidades con las que c y s escogen una respuesta no dependen de la pregunta en consideración. Éste se usa en dos instancias. Primero, para aplicar el teorema Central del Límite (algo que no constituye una verdadera limitación ya que existen otra versiones del teorema Central del

Límite que son aplicables a esta situación) y segundo, para calcular las probabilidades π_{v+} y π_{+v} que son solo validas con la hipótesis de que la probabilidad de las respuestas no depende de la pregunta. Esta última hipótesis es muy fuerte y por esta razón se propone hacer una recodificación de las variables. Este es el contenido de la siguiente sección.

3.3. Índice Kappa: recodificación

La idea de la recodificación se basa en la observación de que una permutación de las opciones de las preguntas no afectaría la suma de la diagonal de la matriz mencionada anteriormente. Sin embargo, es en la diagonal que se registran las respuestas coincidentes. Luego, en principio, existe cierta libertad para modificar las celdas por fuera de la diagonal (las marginales). Utilizando esta idea, el objetivo es recodificar las respuestas de tal forma que la probabilidad conjunta de respuesta sea igual para todas las preguntas. El procedimiento propuesto es el siguiente.

Primero se calcula la distribución empírica de las respuestas a cada pregunta y se ordenan las respuestas a cada pregunta según su frecuencia (esto es lo que se denomina valor_a más abajo en la explicación del código). Luego se toman cada uno de los individuos de la base de datos y se cambian sus respuestas (A, B, C, etc.) por (1, 2, 3, etc.) respetando el ordenamiento de la distribución empírica mencionada anteriormente. El único cambio hecho es el de forzar que la respuesta 1 corresponda a la correcta. Las demás respuestas siguen el ordenamiento empírico entre las que no son correctas.

Una vez recodificada toda la base de datos de respuesta de los individuos para todo c y s, se calcula la matriz de respuestas mencionada anteriormente pero utilizando como cadena de repuestas la cadena recodificada.

Obsérvese que la recodificación no iguala la distribución de las respuestas entre preguntas pero reduce sus diferencias manteniendo la consistencia con los datos observados. Ahora, la diferencia de la distribución entre preguntas puede depender de la habilidad de los individuos. Por esta razón en la tercera versión de índice Kappa se condiciona la distribución empírica de las respuestas a la habilidad medida como el número de respuestas correctas.

3.4. Índice Kappa: recodificación condicional

Esta es la versión del índice Kappa que se ha implementado como un algoritmo de detección de copia. La diferencia principal con el método anterior es que ahora se calcula la distribución empírica de las respuestas por pregunta condicional al número de respuestas correctas. De esta forma para cada número posible de respuestas correctas y para cada pregunta se calcula una distribución empírica de las respuestas. Una vez hecho esto se toma la totalidad de la base de datos y se determina, según su cadena de respuestas, la habilidad de cada individuo (número de respuestas correctas) y la distribución empírica correspondiente a ese número de respuestas correctas. Ahora, dados c y s se utiliza el procedimiento anterior para determinar la distribución empírica con la misma habilidad de c y se recodifican c y s son esa misma distribución empírica de c para cada una de las preguntas. El efecto de esta recodificación es uniformizar aún más la distribución de las respuestas entre individuos condicional a la habilidad.

Ahora, con la recodificación condicional de c y s se sigue el mismo procedimiento de la versión simplificada del índice Kappa (sección 3.1).

4. Resultados y observaciones

Todos los resultados que se muestran a continuación utilizan un nivel de confianza del 99.9%. Sin embargo, es inmediato extraer el subconjunto de sospechosos con un nivel de confianza superior, usando los respectivos valores del estadístico. Asimismo, se consideran 36 formas como exámenes separados. Es decir, si un área toma varias formas, cada una se considera como un examen independiente. Así, no se analizan casos de copia de dos personas en una misma área, si tenían formas diferentes.

Con el 99.9% de confianza, el porcentaje de personas sospechosas de copia según el algoritmo es 48.29%. El porcentaje de parejas sospechosas de copia es de 2.94%. Este valor corresponde a las parejas *ordenadas* que son detectadas como sospechosas. Considerablemente mayor, es el porcentaje de salones en donde se sospecha hubo copia. Este es 96.94%.

La siguiente gráfica presenta el número de copias que realiza una persona, considerando distintas formas y fuentes de copia.

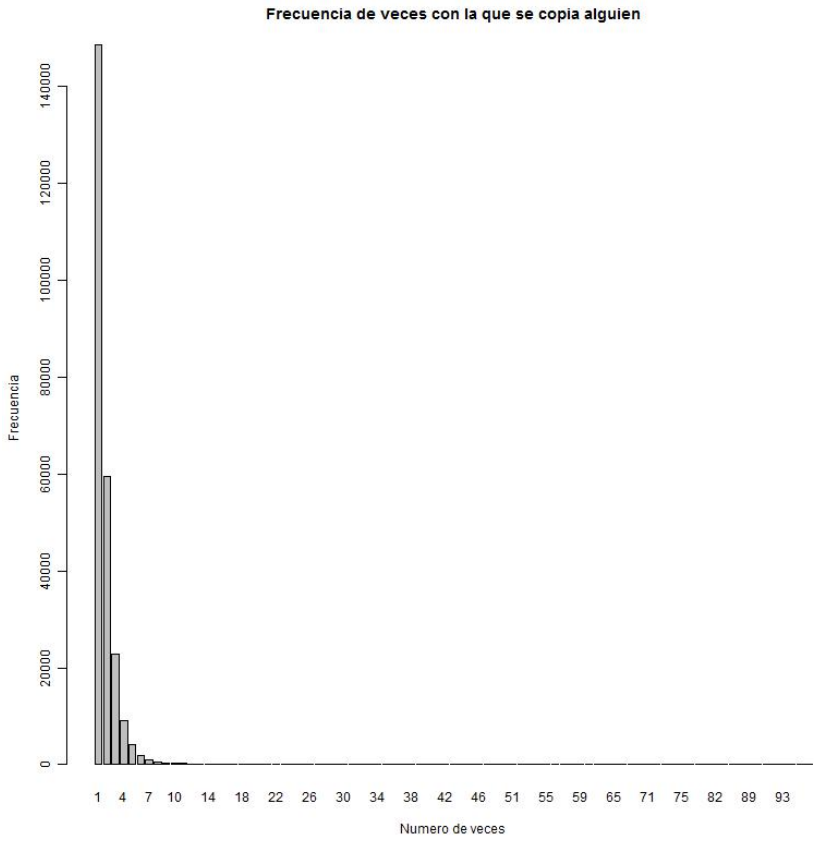

Figura 1. Número de veces que copia una persona (variando por salón, forma y fuente de copia).

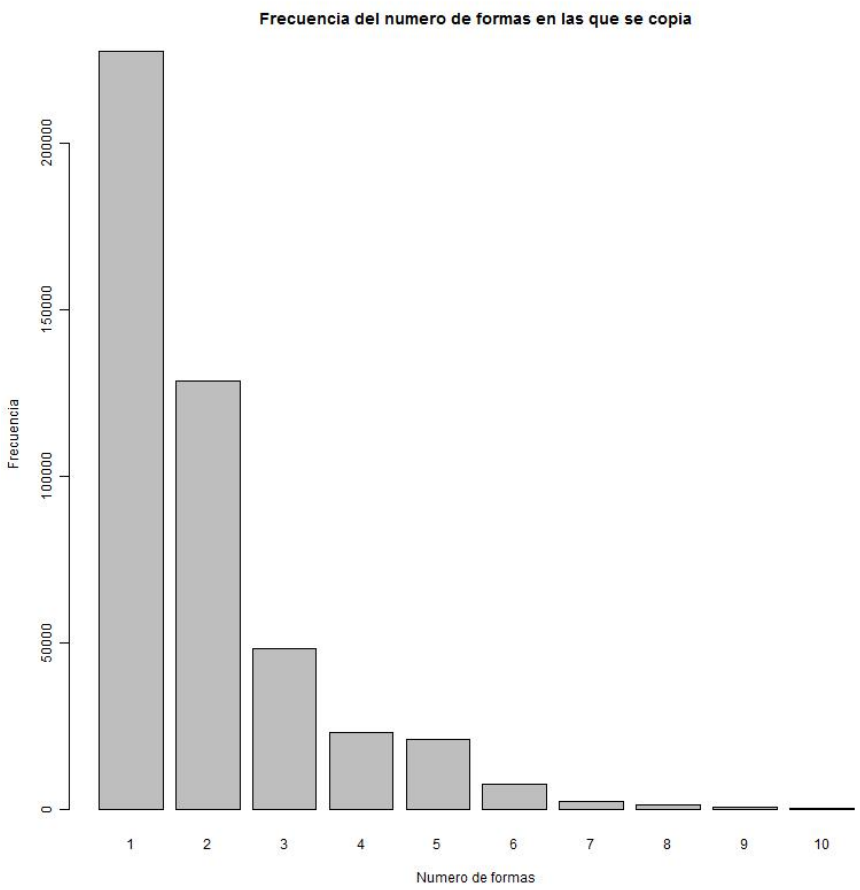

Figura 2. Número de formas en las que se detecta un sospechoso de copia.

En la gráfica anterior se encuentra la frecuencia del número de formas en las que se realiza copia. Es decir, se cuenta el número de formas en la que un individuo que sospechoso de copia aunque sea una vez, es detectado por el algoritmo.

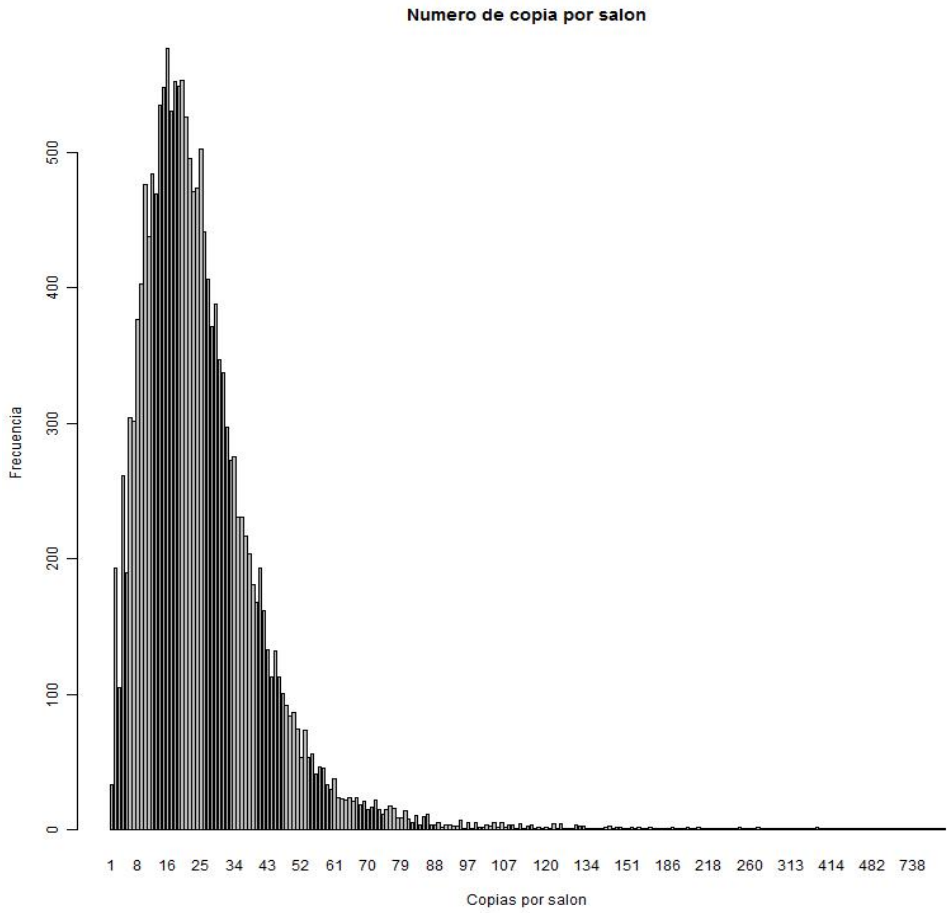

Figura 3. Número de parejas (ordenadas) detectadas como sospechosas en un salón.

Esta gráfica representa la frecuencia del número parejas sospechosas de copia por salón. Se cuenta cada pareja ordenada de sospechosos de copia detectado por el algoritmo. Es decir, si dos individuos A, y B, presentan la misma forma en el mismo salón, y se sospecha que A copió de B y que B copió de A, entonces se cuentan dos parejas como sospechosas de copia.

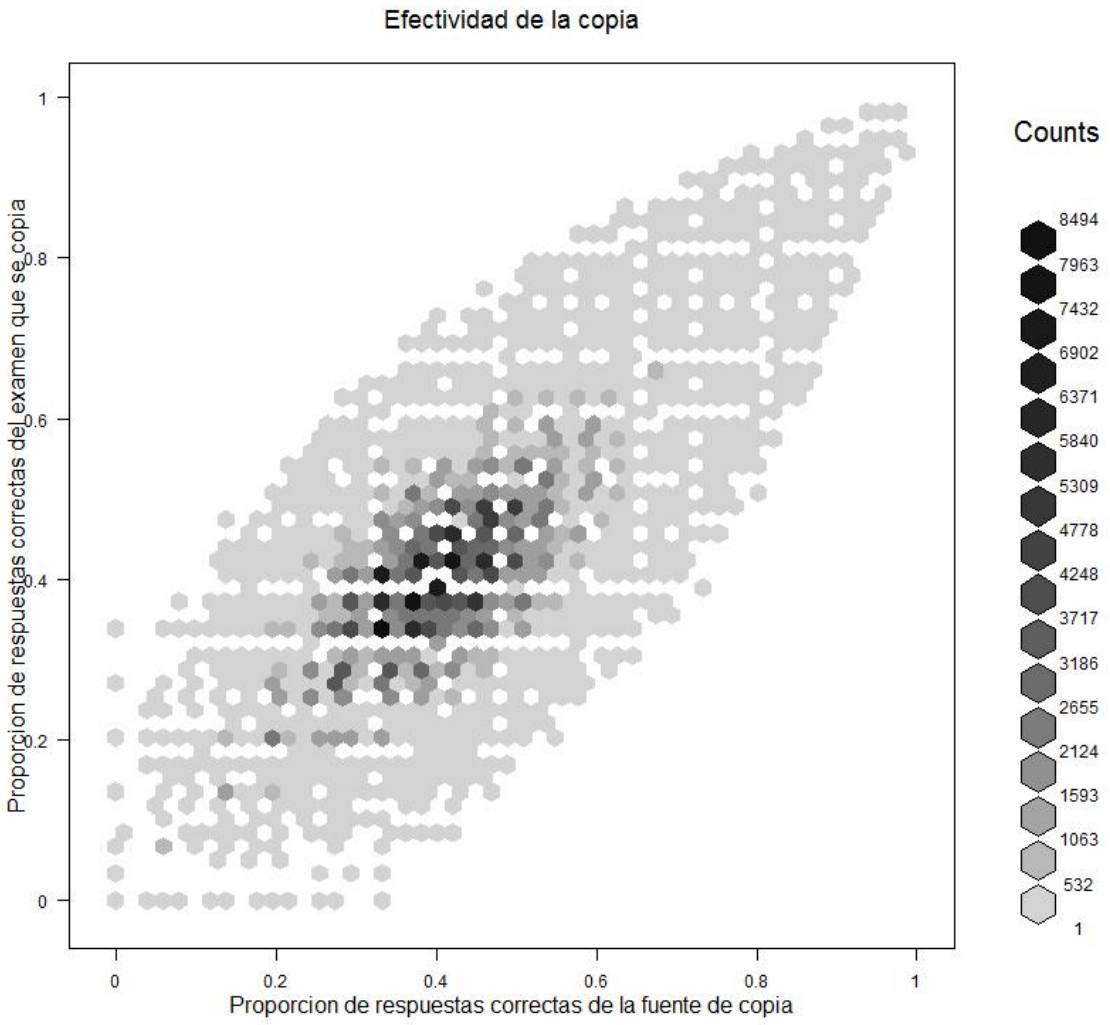

Figura 4. Relación del número de respuestas correctas entre individuo que copia y fuente, tomando las parejas sospechosas de copia.

En la gráfica anterior se exhibe la relación y la frecuencia que existe entre el número de respuestas correctas que tiene la fuente de la copia y el que copia. Como era de esperar, la proporción de respuestas correctas entre la fuente y la persona que copia suele ser similar. Adicionalmente la mayoría de los individuos sospechosos de copia no logran responder más del 50% de las preguntas correctamente.

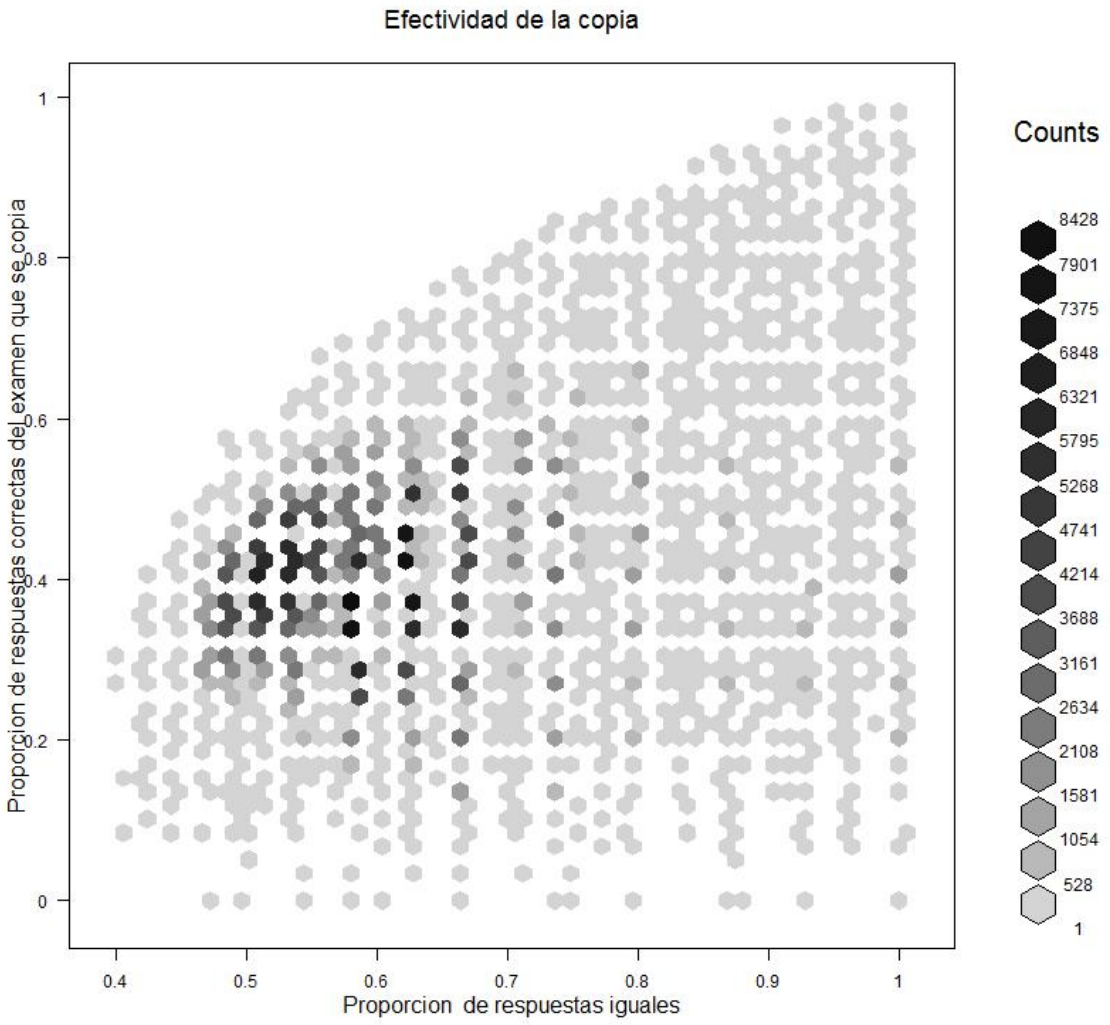

Figura 5. Relación entre número de respuestas correctas del sospechoso de copia y número de respuestas iguales con la fuente.

En esta gráfica se relaciona el número de respuestas correctas que tiene el que copia, con el número de respuestas iguales que tiene con la fuente. Como se ve, aunque el número de respuestas incrementa con el número de preguntas iguales, la copia no es del todo efectiva, pues el número de respuestas correctas es menor que el número de preguntas iguales.

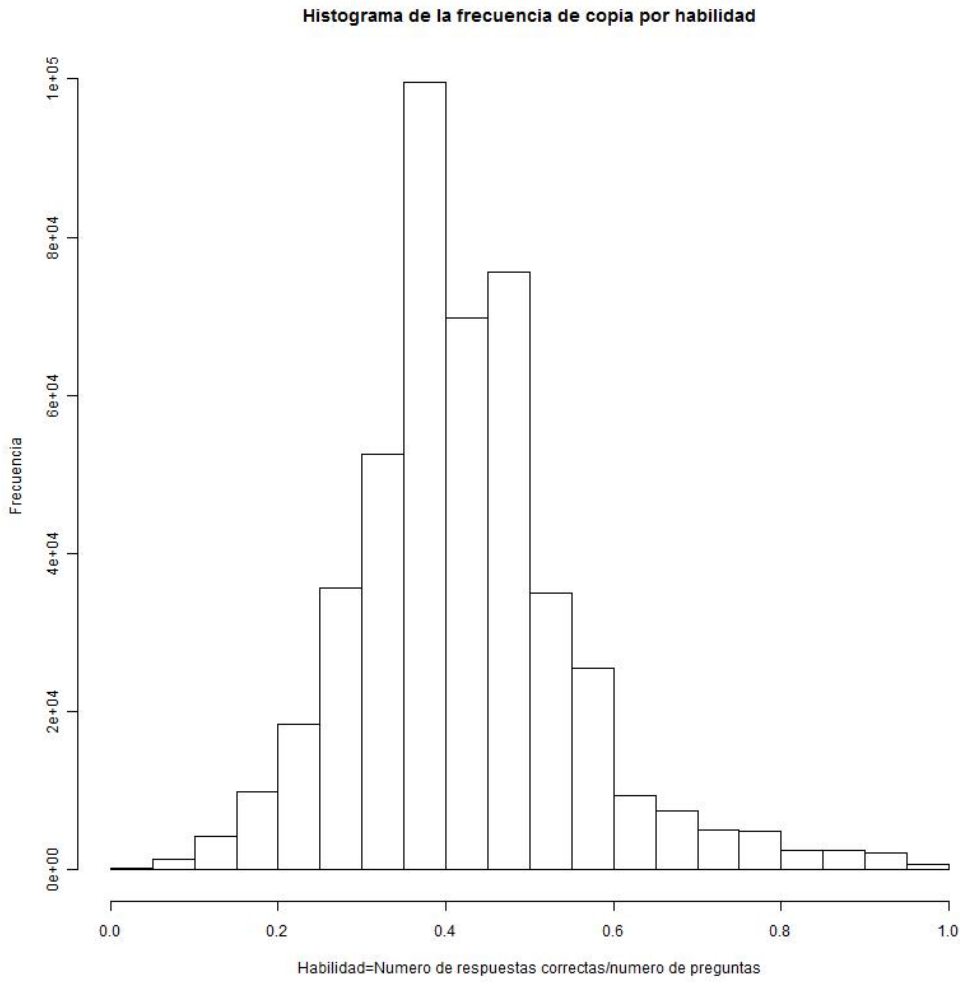

Figura 6. Distribución de la habilidad entre los sospechosos de copia.

La gráfica anterior representa la distribución de la habilidad entre aquéllos que se copian. Por habilidad se entiende el número de respuestas correctas sobre el número de preguntas total. Como se puede ver la mayoría de la gente sospechosa de copia presenta una habilidad menor al 50%.

En la siguiente gráfica (y tabla) se encuentra la proporción de copia por forma. Esta proporción se calcula como el número de estudiantes que presentan la forma y son sospechosos de copia sobre el número de estudiantes que presentan la forma. Dos formas sobresalen, la 51 y 149, correspondientes al examen de inglés, que solo es dividido en dos formas.

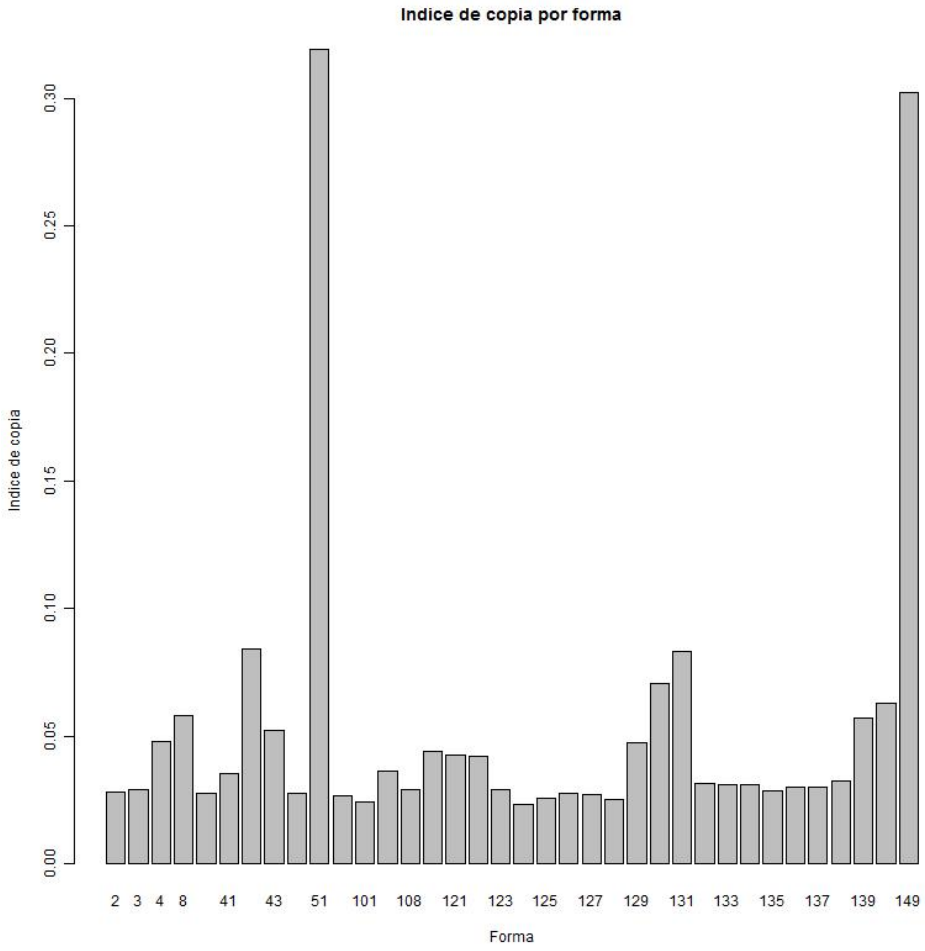

Figura 7. Proporción de estudiantes sospechosos de copia en cada forma.

Forma	Tasa de copia	Forma	Tasa de copia	Forma	Tasa de copia
2	0.028456223	106	0.036626887	130	0.070666793
3	0.029355835	108	0.029160714	131	0.083363414
4	0.048014035	120	0.044204353	132	0.031486583
8	0.058310887	121	0.042585205	133	0.030958223
10	0.027542242	122	0.042461895	134	0.030931415
41	0.035593925	123	0.029392433	135	0.028688525
42	0.084406941	124	0.0234282	136	0.030233348
43	0.052603249	125	0.025744734	137	0.03000903
44	0.027599007	126	0.027789574	138	0.032761648
51	0.319310562	127	0.027268235	139	0.05744152
100	0.026786064	128	0.025484987	140	0.063147962
101	0.024446639	129	0.047734147	149	0.30205696

Tabla 1. Proporción de estudiantes sospechosos de copia en cada forma.

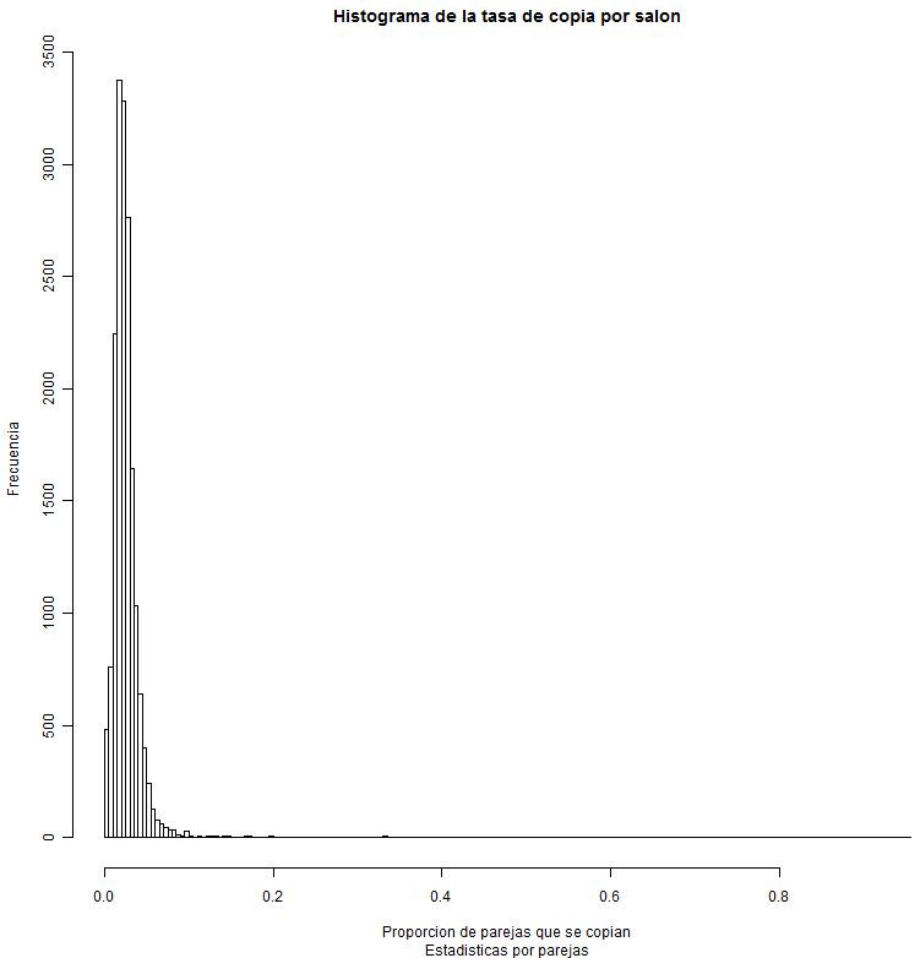

Figura 8. Proporción de parejas que se copian en cada forma.

El porcentaje de copia también se analiza por parejas, tomando en cuenta que el orden en la pareja importa (quién es la fuente y quién es el sospechoso). La figura anterior muestra los resultados. Tomando esta información, en la siguiente figura se muestra la distribución acumulada: para un nivel de porcentaje de parejas que copian, se muestra la proporción de los salones presentan a lo sumo ese porcentaje.

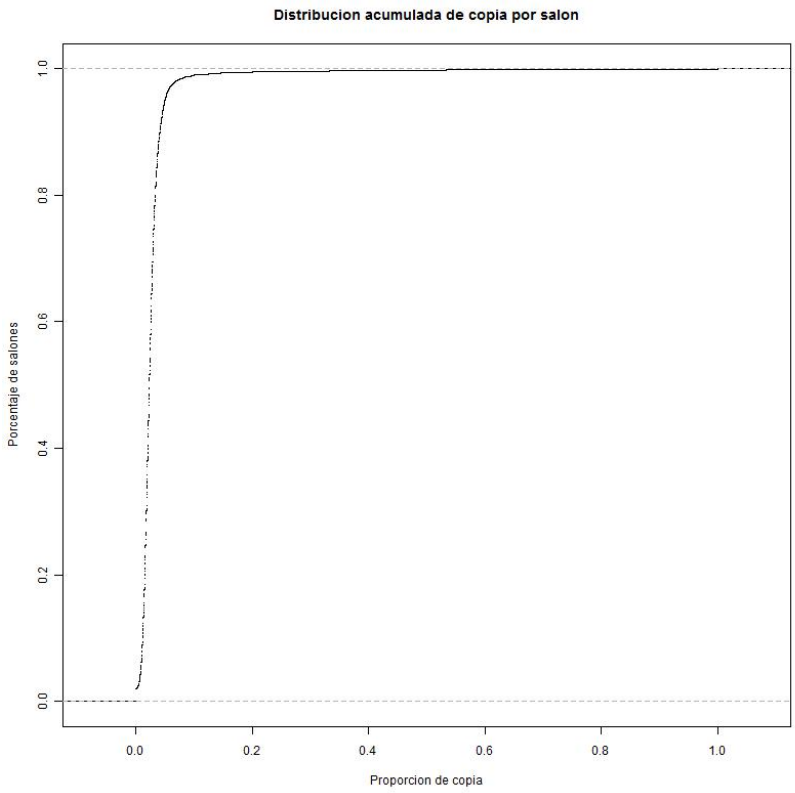

Figura 9. Porcentaje de salones que tienen a lo sumo la proporción correspondiente de parejas sospechosas de copia.

5. Resultados del Índice para Pruebas Independientes

Si se toman parejas compuestas por individuos que presentan pruebas en salones distintos, es de esperarse que el índice delate sospecha de copia en uno de cada mil casos, si las suposiciones de construcción del modelo son satisfechas en la práctica. Para verificar esto, se construyen 50,000 parejas de forma aleatoria e independiente (sobre distintos salones) para cada forma. Los resultados se incluyen en la siguiente tabla.

Como se observa, el índice detecta copia en alrededor de 3 por cada mil casos estudiados, y de hecho hay bastante dispersión en las formas consideradas. Las formas 51 y 149 (inglés) presentan resultados por encima de 15 casos en cada mil. Para dos formas (2 y 106) se repitió el experimento, tomando un nivel de confianza de uno en un millón, considerando 15 millones de parejas independientes, obteniendo 14.9 y 3.7 casos por cada millón, respectivamente.

Forma	# copias por mil	Forma	# copias por mil	Forma	# copias por mil
2	3.2083	106	1.0417	130	4.7750
3	3.2833	108	2.1417	131	5.2667
4	5.9000	120	5.2333	132	3.5833
8	3.2583	121	5.3333	133	3.6250
10	3.0750	122	5.4833	134	3.3833
41	4.1500	123	3.0333	135	3.1250
42	5.7917	124	2.6917	136	3.3167
43	2.3417	125	2.8833	137	3.2417
44	1.0167	126	3.2250	138	3.2667
51	16.5333	127	2.9000	139	3.5417
100	0.8333	128	2.8833	140	3.1417
101	1.2167	129	6.1667	149	19.5750

Tabla 2. Casos de copia por mil parejas independientes, para las 36 formas analizadas.

Los resultados sugieren una fuerte diferencia entre las probabilidades de respuesta en las preguntas, que se suponían idénticas para la definición del índice. En efecto, los patrones de respuesta varían bastante con las preguntas, como se puede observar en las siguientes figuras, que muestran – para tres formas distintas – las probabilidades de respuesta usadas para cuatro preguntas. Para cada prueba se escogió un nivel de habilidad arbitrariamente.

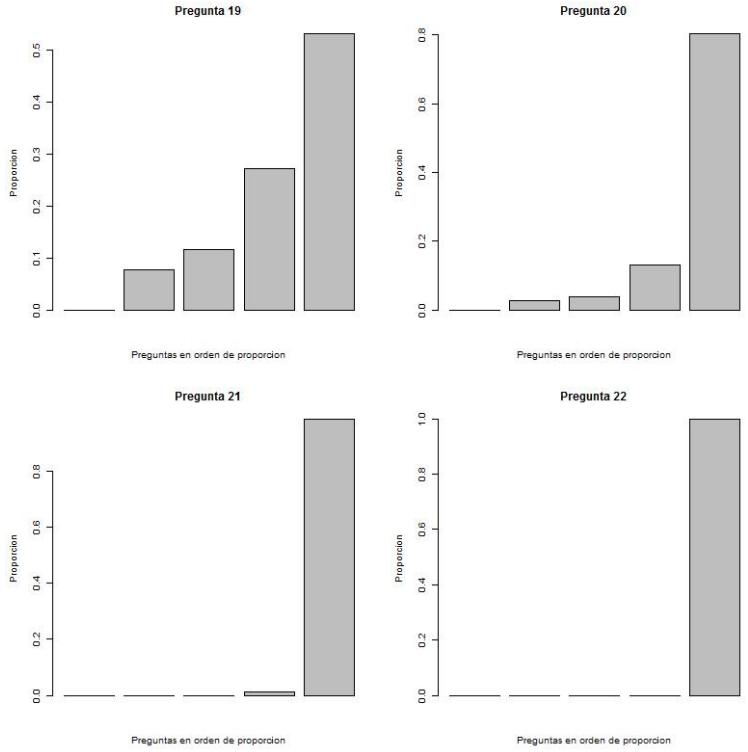

Figura 10. Probabilidades de respuesta, Forma 2, Habilidad 22, preguntas 19-22.

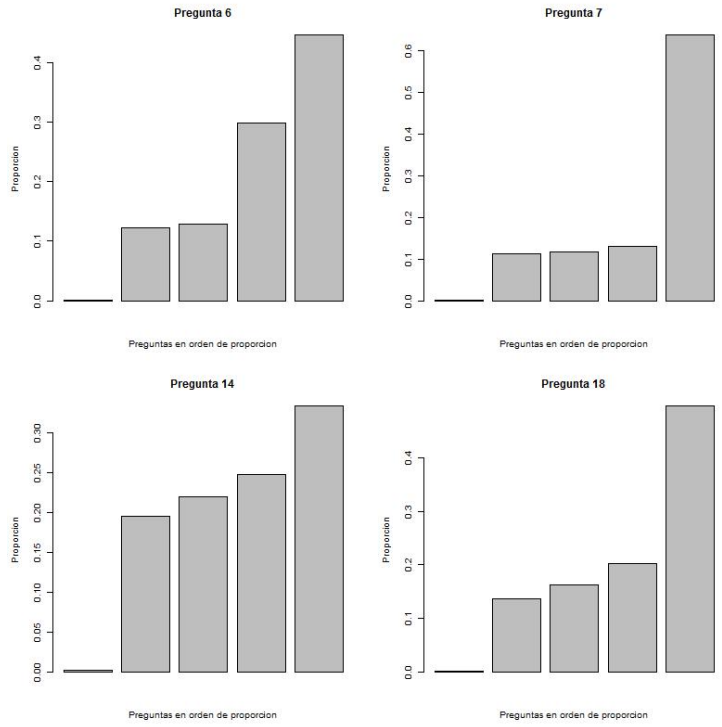

Figura 11. Probabilidades de respuesta, Forma 3, Habilidad 6, preguntas 6, 7, 14, 18.

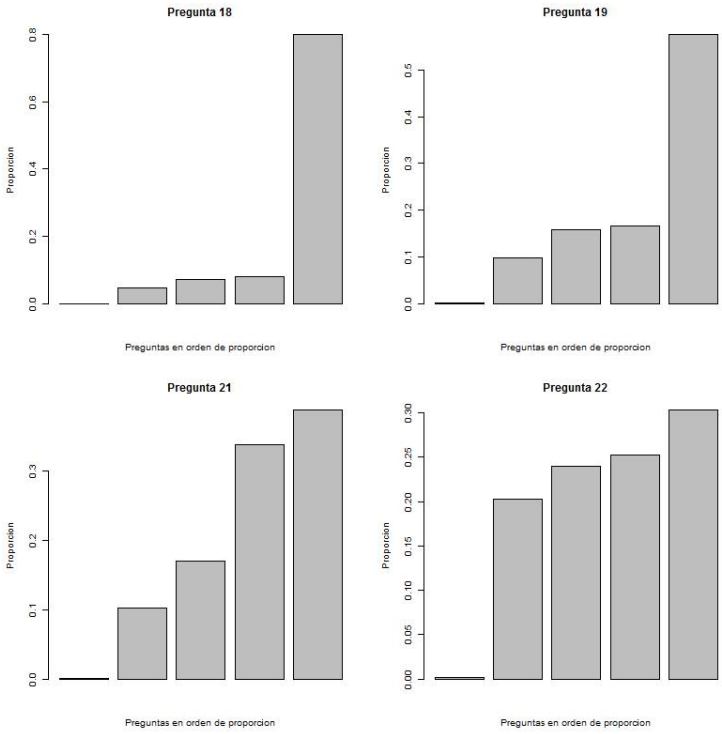

Figura 12. Probabilidades de respuesta, Forma 4, Habilidad 12, preguntas 18, 19, 21, 22.

Estas observaciones sugieren explorar direcciones para subsanar el impacto que las suposiciones de las distribuciones de las preguntas tienen sobre la detección de casos

potenciales de copia. Una posibilidad que no requiere alterar el esquema de trabajo es corregir la media del estadístico del índice con la media encontrada para pruebas independientes. La siguiente tabla exhibe los resultados que se obtendrían con este esquema. Como es de esperarse por construcción, el número de casos sospechosos por cada mil analizados se acerca a uno.

Forma	# copias por mil	Forma	# copias por mil	Forma	# copias por mil
2	1.0750	106	0.7167	130	1.0583
3	1.1250	108	0.9833	131	0.9500
4	1.2083	120	1.2583	132	0.9583
8	1.0000	121	1.1833	133	1.1000
10	0.9250	122	1.2417	134	1.0833
41	1.0250	123	0.9083	135	0.9833
42	1.2583	124	1.0083	136	1.0417
43	0.9917	125	1.0333	137	1.0417
44	0.6750	126	1.3000	138	1.0583
51	2.1833	127	1.0333	139	0.9333
100	0.6500	128	1.1750	140	0.7750
101	0.8250	129	1.0667	149	2.4750

Tabla 3. Casos de copia por mil parejas independientes, para las 36 formas analizadas, corrigiendo por la media del muestreo independiente.

Sin embargo, esta solución no viene acompañada de un soporte estadístico que justifique teóricamente este arreglo. Por lo tanto, la sugerencia es generalizar el índice construido mediante cambios estructurales que permitan considerar diferencias en las distribuciones de las respuestas. Esta dirección se elabora en las conclusiones.

6. Conclusiones

El presente proyecto concluye con la implementación del índice Kappa, según los objetivos planteados inicialmente. Es posible tener resultados para cada forma del examen de Estado para Calendario A en unas dos a tres horas (tiempos basados en un procesador de 2.33 GHz y 4GB de RAM). Los resultados se presentan usando un nivel de confianza de 99.9%. La siguiente figura muestra un resumen distribucional del estadístico del índice, agregando para las 36 formas. Las barras sombreadas corresponden a parejas que presentaron el examen en el mismo salón; las barras blancas corresponden a parejas independientes. Se

puede observar un sesgo a la derecha para las pruebas con posibilidad de copia, lo cual apunta a la presencia de un efecto presentado por la posibilidad de copia.

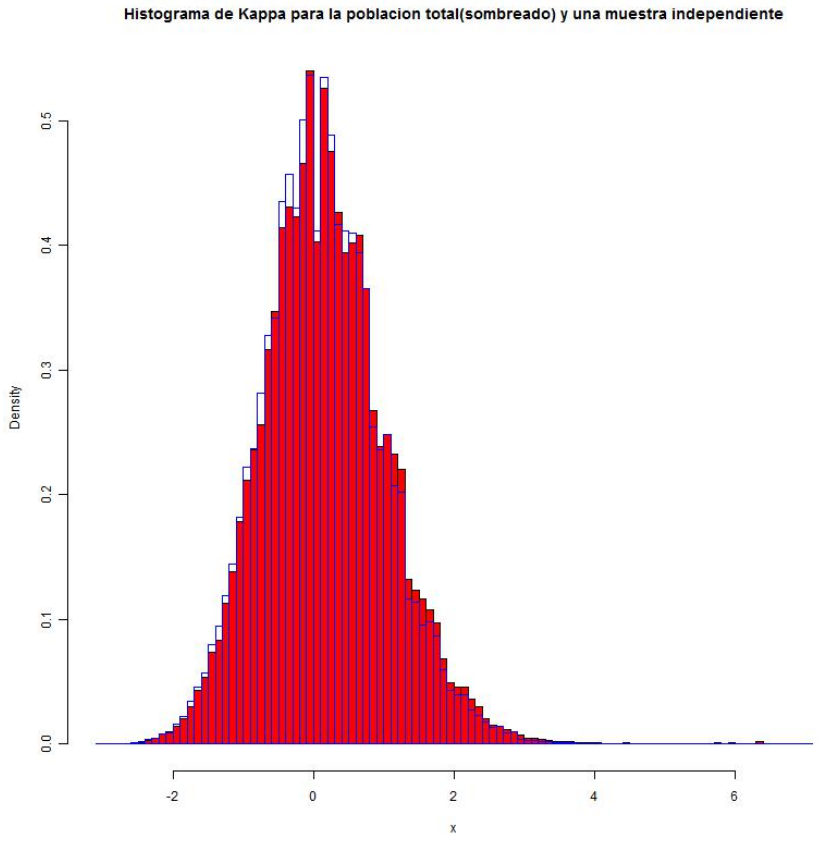

Figura 13. Estadístico de índice kappa, para el agregado de formas. Columnas sombreadas corresponden a parejas en el mismo salón. Las blancas corresponden a parejas independientes.

Ahora, la distribución conjunta de las respuestas de los individuos no es independiente de la pregunta. Esta era una suposición necesaria para la determinación de la distribución a la que converge el índice Kappa. De hecho, la media del índice, para formas independientes, no es cero, como se esperaría si se cumpliera la hipótesis. Esto permite sugerir futuras direcciones de trabajo que permitan generalizar el índice desarrollado. En particular, a continuación se describe una metodología basada en un modelo de respuesta nominal (NRM).[7]

Sea $i = 1, \dots, n$ el número de preguntas y $a = 1, \dots, k$ el número de alternativas para responder cada pregunta. Sea U_{ji} la respuesta del individuo j a la pregunta i. Supongamos

[7] Para mayores detalles véase (van der Linden & Sotaridona, 2006).

que un individuo con habilidad θ que no ha copiado de ningún otro individuo responde a en la pregunta i con probabilidad

$$\pi_{i,a}(\theta) = \frac{\exp(\zeta_{i,a} + \lambda_{i,a}\theta)}{\sum_{a=1}^{k} \exp(\zeta_{i,a} + \lambda_{i,a}\theta)}.$$

Sea γ_{js} el número de respuestas que j le copio a s. El objetivo es probar las siguientes hipótesis:

$$H_0: \gamma_{js} = 0$$

$$H_1: \gamma_{js} > 0$$

Sea I_{jsi} la función indicadora del conjunto $U_{ji} = U_{si}$. Es decir, es 1 cuando el individuo j y s coinciden es su repuesta a la pregunta i. Luego el número de respuestas en común entre j y s es

$$M_{js} = \sum_{i=1}^{n} I_{jsi}.$$

Si lo individuos no se copian (responden de forma independiente), la probabilidad de que j y s escojan la alternativa a en la pregunta i es

$$\pi_{jsi,a} = \pi_{i,a}(\theta_j)\pi_{i,a}(\theta_s).$$

Luego la probabilidad de una coincidencia entre j y s en la pregunta i es

$$P(I_{jsi} = 1) = \pi_{jsi} = \sum_{a=1}^{k} \pi_{jsi,a}.$$

Una forma equivalente de plantear las hipótesis de interés es utilizar la siguiente prueba:

$$H_0: \pi_{jsi} = \sum_{a=1}^{k} \pi_{i,a}(\theta_j)\pi_{i,a}(\theta_s),\ i = 1, \dots, n$$

$$H_1 = \left\{ \begin{array}{c} \pi_{jsi} = \sum_{a=1}^{k} \pi_{i,a}(\theta_j)\pi_{i,a}(\theta_s)\text{ , para } n - \gamma_{js} \text{ preguntas} \\ 1, \text{para } \gamma_{js} > 0 \text{ preguntas} \end{array} \right\}$$

Bajo la hipótesis nula el número de respuestas coincidentes,

$$M_{js} = \sum_{i=1}^{n} I_{jsi}$$

es la suma de n variables aleatorias de Bernoulli independientes cada una con una

probabilidad distinta de coincidencia. La suma de estas variables aleatorias sigue una distribución binomial generalizada que se puede calcular mediante un algoritmo recursivo.

Para implementar esta metodología a las pruebas del ICFES se podría aproximar la habilidad de los individuos por el número de respuesta correctas, estrategia que se ha utilizado a lo largo del documento, y estimar el modelo de respuesta nominal.

Finalmente, se observa que el índice Kappa fue implementado calculando el índice para todas las parejas c y s en cada salón y para cada forma. Estrictamente, para cada pareja c y s se hizo una prueba de detección de copia por cada una de los exámenes (formas) en las que los dos individuos participaron. Es decir, por pareja se calcularon varias pruebas de hipótesis, donde la hipótesis nula era que los individuos no se copiaron. Alternativamente, se podría trabajar con la prueba simultánea de que dos individuos no se copiaron en ningún examen (forma). Es un hecho bien conocido en estadística que aceptar la hipótesis nula de todas las pruebas (con el mismo nivel de confianza) no es equivalente a aceptar la hipótesis nula de la prueba simultánea. La implementación de una técnica que lleve en consideración esta observación sería muy valiosa para poder construir un índice de detección de copia para parejas (sin identificar las formas), llevando en consideración la información revelada por esta pareja en todas las formas presentadas.

Así, se sugieren dos líneas de trabajo futuro que permitan incorporar estas generalizaciones:

6.2.1. Implementación de una metodología basada en un modelo de respuesta nominal (NRM).

6.2.2. Pruebas simultáneas para cada pareja (que abarquen todas las formas simulatáneamente).

Referencias

Angoff, W. (1974). The Development of Statistical Indices for Detecting Cheaters. *Journal of the American Statistical Association*.

Baird, L. J. (1980). Current trends in college cheating. *Psychology in the schools*.

Belleza, F., & Belleza, S. (1989). Detection of Cheating on Mutiple -Choice Test by Using Error-Similarity Analysis. *Teaching of Psychology*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* , 37-46.

Frary, B., Tideman, N., & Watts, T. (2009). Indices of cheating on multiple choice tests.

Holland, P. (1996). Assesing unusual agreement between the incorrect answers of two examinees using the K index: statistical theory and empirial support. *ETS technical report*.

Sotadoridona, L., & Meijer, R. (2003). Two new indices to detect answer copying. *Journal of Educational Measurment*.

Sotaridona, L., & Mejier, R. (2001). Statistical Properties of the K-Index for Detecting Answer Copying. *Research Report Educational Science and Technology*.

Sotaridona, L., van der Linden, W., & Meijer, R. (2006). Detecting Answer Copying uisng Kappa Statistic. *Applied Psychology Measurement* , 412-431.

Van der Linden, W., & Hambleton, R. (1997). *Handbook of Modern Response Theory.* Srpinger.

Van der Linden, W., & Sotaridona, L. (2006). Detecting Answer Copying when the Regular Response Process Follows a Known Response Model. *Journal of Educational and Behavioral Statistics*.