Algoritmos para detectar copia en exámenes de opción múltiple y su aplicación en la prueba ENLACE

Nombre del Autor .....

Lic. Brenda Jesús Rodríguez Alcántar

Director de tesis .....

Dr. Johan Jozef Lode Van Horebeek

Aceptado por .....

M. en I. Maximino Tapia Rodríguez
Revisor de tesis

Aceptado por .....

Dr. Rogelio Hasimoto Beltrán
Revisor de tesis

Algoritmos para detectar copia en exámenes de opción múltiple y su aplicación en la prueba ENLACE

por

Brenda Jesús Rodríguez Alcántar

Resumen

En pruebas académicas, el formato que comúnmente se utiliza es el de multi-opción. El proyecto de tesis se centra en el problema de detección de copia en este tipo de exámenes, con el apoyo de datos provenientes de la aplicación de la prueba **ENLACE** del 2007. Uno de los procedimientos para la evaluación de exámenes irregulares son los métodos estadísticos que modelan las probabilidades de las respuestas de los sujetos bajo el supuesto que no copian en busca de patrones de respuesta similares entre los sujetos examinados.

A mis padres y hermanas.

A Edel con todo mi corazón.

Agradecimientos

Muchas han sido las personas que de manera directa o indirecta me han ayudado en la realización de esta tesis. Quiero dejar constancia de todas ellas y agradecerles con sinceridad su participación.

Mi más sincero agradecimiento al Dr. Johan Jozef Lode Van Horebeek, profesor-investigador del Centro de Investigación en Matemáticas, por la oportunidad de trabajar bajo su coordinación, una vez más. A los sinodales de tesis, los cuales con sus comentarios y sugerencias han mejorado considerablemente la calidad de este trabajo.

¡A mi familia por su apoyo incondicional!

Este trabajo ha sido parcialmente subvencionado por becas del **C**onsejo de **C**iencia y **T**ecnología del **E**stado de **G**uanajuato (**CONCyTEG**) y el **C**onsejo **N**acional de **C**iencia y **T**ecnología (**CONACyT**).

Índice general

Capítulo 1

Introducción

En pruebas académicas, el formato que comúnmente se utiliza es el de multi-opción, ya que proporciona facilidad y confiabilidad a la hora de calificar la prueba para un gran número de sujetos examinados. Como ejemplo de este tipo de exámenes se pueden mencionar: ***G**raduate **M**anagement **A**dmissions **T**est* (**GMAT**), ***T**est of **E**nglish as a **F**oreign **L**anguage* (**ToEFL**) y ***G**raduate **R**ecord **E**xamination* (**GRE**), aplicados por la ***E**ducational **T**esting **S**ervice* (**ETS**) y en México la **E**valuación **N**acional de **L**ogro **A**cadémico en **C**entros **E**scolares (**ENLACE**), aplicada por la **S**ecretaria de **E**ducación **P**ublica (**SEP**).

La prueba **ENLACE** se aplica en los diferentes tipos de escuelas primarias a nivel nacional: CONAFE, general, indígena y particular. Así como también en las distintos tipos de escuelas secundarias: general, particular, telesecundaria y técnica.

Se elabora una prueba dividida por secciones con temas de español y matemáticas para cada grado y asignatura. Las preguntas son de opción múltiple y son presentadas en un cuaderno y una hoja donde los alumnos registran sus respuestas. Tiene 101 preguntas como mínimo (para tercer grado de primaria) y hasta 138 como máximo (para tercer grado de secundaria), variando en cada asignatura-grado. Las diferentes pruebas evalúan los contenidos establecidos en los planes y programas de estudio oficiales vigentes en la **SEP**. Ver apéndice A para más información sobre esta prueba.

Existen varias cuestiones sobre la calidad de dicha prueba, por ejemplo: ¿Hasta qué grado las respuestas entregadas son resultados individuales y son obtenidas en las condiciones de aplicación preestablecidas de la prueba? [9].

Los factores más importantes que afectan de manera negativa la aplicación idónea de la prueba, son:

- A nivel de alumno: copiar o dejar copiar, cambiar su identidad, nivel o grupo, no hacer un esfuerzo para resolver la prueba correctamente, consultar material ó personas externas, cambiar respuestas de una sesión anterior (esto es posible ya que se usa una sola hoja de respuestas para todas las sesiones).

- A nivel de grupo: permitir copiar, dictar respuestas o indicar respuestas erróneas, proporcionar explicación no permitida, rellenar o cambiar hojas de respuestas, cambiar las condiciones preestablecidas de aplicación (por ejemplo dar más tiempo), controlar la población al cual se aplica la prueba.

- A nivel de escuela: organizar sesiones de preparación, controlar los grupos a los cuales se les aplicará la prueba.

En particular esté proyecto de tesis se centra en el problema de detección de copia en este tipo de exámenes, con el apoyo de datos provenientes de la aplicación de la prueba **ENLACE** en el 2007.

Métodos para evaluación de exámenes irregulares

Los llamados "métodos de observación" utilizan un observador humano para establecer si algún sujeto ha copiado respuestas. Este tipo de métodos se basan en observar ciertos tipos de comportamientos físicos de los sujetos, para determinar si han copiado.

Los métodos estadísticos, por otro lado, modelan las probabilidades de las respuestas de los sujetos bajo el supuesto que no copian en busca de patrones de respuesta similares entre los sujetos examinados.

1.1. Descripción

Para un grupo de N alumnos, donde cada uno esta realizando un examen de M preguntas, se tiene un conjunto de cadenas (o trenes) de respuesta, donde cada cadena es dada por el i-ésimo alumno y cada elemento de dicha cadena pertenece a la respuesta de la j-ésima pregunta, ver cuadro 1.1. Se define u_{ij} como la respuesta del alumno i a la pregunta j.

		Preguntas					
		1	2	3	...	$M-1$	M
	1	A	A	D		C	A
	2	A	A	B	...	C	A
Alumnos	3	A	A	A		C	A
	⋮		⋮		⋱		⋮
	$N-1$	A	A	A		C	A
	N	A	A	B	...	C	A

Cuadro 1.1: Ejemplo de un examen de grupo.

El problema principal es: ¿como detectar copia entre alumnos en un examen de esté tipo a nivel *micro* y *macro*?.

- Nivel *micro*: realizar detección de alumnos que copian respuestas de sus exámenes a otros alumnos dentro de un mismo grupo de estudiantes.

- Nivel *macro*: proporcionar información acerca del comportamiento de un determinado grupo sobre la cantidad de copia detectada.

Hay que tomar en cuenta los siguientes puntos:

- No existe un modelo verdadero. Hasta el momento no hay en la literatura un modelo matemático que sea capaz de describir el comportamiento de un grupo de alumnos al contestar un examen. ¡Ni existirá!

- Existen varios métodos de copia. Por la naturaleza del proceso de copia, existen varios "trucos" o formas de copiar en un examen. Por lo tanto no es posible emular todos estos trucos estudiantiles.

- Existe siempre una probabilidad de equivocación.

Como parte de la realización del proyecto de tesis se realizan un sin fin de experimentos. Muchos de los datos con los cuales se trabajo han sido simulados, así también se ha simulado el proceso de copia, tratando de abarcar varios casos reales. A continuación se enlistan los métodos seguidos para la generación de datos y los métodos seguidos para el proceso de copia.

1.1.1. Obtención/Generación de datos

Se hace uso de tres tipos de datos: datos simulados (los cuales son 100 % artificiales), datos ficticios (se crean a partir de datos reales) y datos reales (tomados directamente de los datos reales).

Datos simulados

Se crean exámenes donde cada pregunta cuenta con cuatro posibles opciones de respuesta. El proceso por el cual el i-ésimo sujeto del conjunto contesta su examen es de manera aleatoria, suponemos que cada pregunta cuenta con las opciones de respuesta {A, B, C, D}. El hecho de fijar en cuatro opciones de respuesta para cada pregunta, es tratar de emular la prueba **ENLACE** donde cada ítem tiene esté número de opciones. Cabe mencionar que se establece como respuesta correcta la opción A. Se idearon dos procedimientos para generar exámenes con M preguntas y N alumnos, los cuales se describen en seguida:

1. Siguiendo una distribución uniforme; tomando en cuenta que cada pregunta tiene cuatro opciones posibles de respuesta, se tiene que la probabilidad de seleccionar alguna de estás opciones es $\frac{1}{4}$. El algoritmo 1 muestra el proceso utilizado para

generar trenes de respuestas siguiendo una distribución uniforme.

Algorithm 1: Genera examen siguiendo una distribución uniforme

```
for i = 1 to N do
    for j = 1 to M do
        q ~ U(0, 1);
        if q ∈ [0.00, 0.25) then
            u_ij = A;
        if q ∈ [0.25, 0.50) then
            u_ij = B;
        if q ∈ [0.50, 0.75) then
            u_ij = C;
        if q ∈ [0.75, 1.00) then
            u_ij = D;
```

2. Siguiendo un modelo **IRT**; estos modelos se explican en apéndice A. A nivel pregunta utiliza tres parámetros para caracterizar cada ítem: nivel de dificultad, probabilidad de discriminación y probabilidad de adivinar la respuesta correcta. A nivel individuo utiliza un parámetro para representa la habilidad o capacidad del alumno de contestar bien el examen. El algoritmo 2 muestra el proceso utilizado para generar trenes de respuestas siguiendo un modelo **IRT**.

Algorithm 2: Genera examen siguiendo un modelo **IRT**

```
a, b, θ ~ U;
c = 1/4;
for i = 1 to N do
    for j = 1 to M do
        P(X_ij = 1|θ_i, a_j, b_j, c_j) = c_j + (1 − c_j)/(1 + exp(−1.7a_j(θ_i − b_j)));
        P(X_ij = 0|θ_i, a_j, b_j, c_j) = (1 − P(X_ij = 1|θ))/3;
        q ~ U(0, 1);
        if q ∈ [P(X_ij = 1|θ), (P(X_ij = 1|θ) + P(X_ij = 0|θ))) then
            u_ij = B;
        else if q ∈ [(P(X_ij = 1|θ) + P(X_ij = 0|θ), (P(X_ij = 1|θ) + 2P(X_ij = 0|θ))) then
            u_ij = C;
        else if q ∈ [(P(X_ij = 1|θ) + 2P(X_ij = 0|θ)), 1) then
            u_ij = D;
        else
            u_ij = A;
```

Datos ficticios

Son trenes de respuestas de un grupo que fue creado artificialmente a partir de respuestas de alumnos reales. Para obtener un alumno, se toma un grupo al azar sin repeticiones de la base de datos (= resultados de la prueba **ENLACE** aplicada en 2007), a su vez dentro del grupo seleccionado se toma al azar y sin repeticiones un alumno. Una vez que se tienen N alumnos se crea el grupo. La intención para crear esté tipo de grupos es tener la certeza que no existe copia entre sus integrantes.

Datos reales

Hacen referencia a grupos completos tomados, en su mayoría, del estado de Puebla. Dichos datos fueron obtenidos de la prueba **ENLACE** aplicada en 2007.

1.1.2. Proceso de copia

Para valorar la eficiencia de las técnicas de detección de copia, ha sido necesario crear situaciones "controladas" en las cuales existe un grupo de sujetos, donde sus integrantes copian respuestas entre ellos. Para esté tipo de situaciones controladas, llamadas simulaciones, se han ideado varios prototipos, los cuales se describen a continuación:

1. Del grupo total de N alumnos y M preguntas, se selecciona al azar un porcentaje de sujetos, que van a copiar un porcentaje de preguntas seleccionadas aleatoriamente. Los sujetos a los cuales se les va a copiar también son seleccionados al azar, sin embargo los sujetos que van a copiar no pueden ser sujetos a los que se les va a copiar.

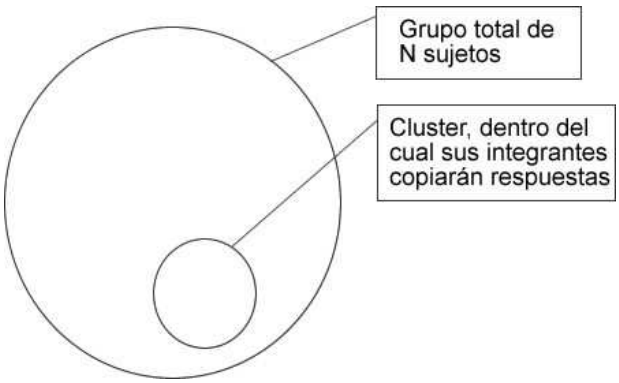
2. Igual al anterior, salvo que en este caso se restringe la selección de ítems a copiar a aquellos que están mal contestados.

3. Siguiendo la idea del punto anterior, con la diferencia que los alumnos fuente inician con una probabilidad de ser copiados igual para todos; al seleccionar al i-ésimo sujeto fuente, para tal caso, dicha probabilidad aumenta. Por lo tanto, aquellos sujetos con mayor probabilidad de ser copiados serán seleccionados más de una vez.

4. Del grupo total de N alumnos, se toma una porción aleatoria de estos, los cuales crean un *cluster*, ver cuadro 1.2. Los sujetos dentro de este *cluster* realizarán copia de alguna de las siguientes formas:

   a) Del total de alumnos/sujetos dentro del *cluster*, se toma al sujeto con un mayor número de aciertos, dicho alumno funge como representante del *cluster*. El resto de los sujetos, dentro del *cluster*, copian al representante sólo las respuestas que cada uno de ellos tiene mal contestadas.

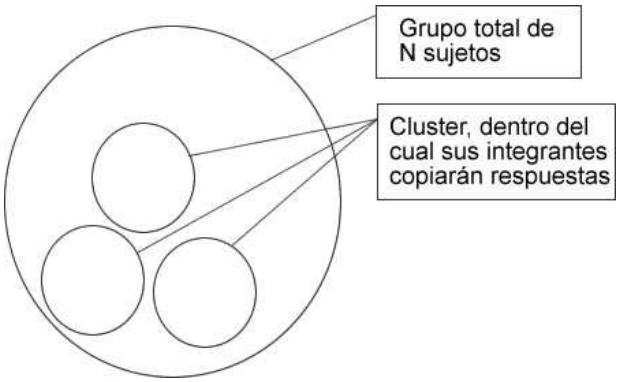
   b) Igual al anterior, salvo que en este caso se restringe la selección de ítems a un porcentaje del total de respuestas.



Cuadro 1.2: Formar un *cluster*.

c) Igual al anterior, salvo que en este caso se restringe la selección de ítems a copiar a aquellos que están mal contestadas.

5. Siguiendo la misma idea que el punto anterior, se crean varios *cluster*, donde el número de *cluster* esta dado por el valor de *K* y el *cluster* tiene el mismo número de alumnos contenidos, ver cuadro 1.3. Los sujetos dentro de cada *cluster* realizarán copia de alguna de las siguientes formas:

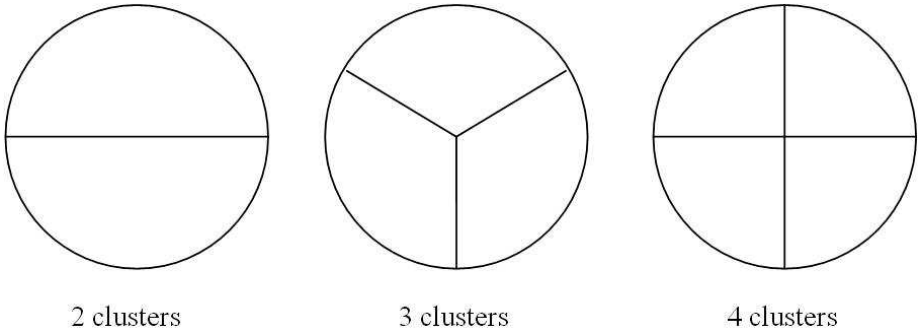


Cuadro 1.3: Formar *K cluster*.

a) Del total de alumnos/sujetos dentro de cada *cluster*, se toma al sujeto con un mayor número de aciertos, dicho alumno funge como representante del *cluster*. El resto de los sujetos, dentro del *cluster*, copian al representante sólo un porcentaje de las respuestas del total de respuestas.

b) Se crean todas las posibles parejas dentro del *cluster*, dichas parejas copiaran entre ellas un porcentaje del total de preguntas. La selección de preguntas a copiar es de forma aleatoria para cada pareja.

6. Siguiendo con la idea anterior, se crean *K cluster*, ver cuadro 1.4. La diferencia con el punto anterior es que todos los alumnos están contenidos en algún *cluster*. El número de alumnos por *cluster* esta dado por $\frac{N}{K}$. Las formas de copiar preguntas son las mismas que en el inciso anterior.



Cuadro 1.4: Formar *K cluster*, todos los alumnos están contenidos en algún *cluster*.

7. Del grupo total de N alumnos se crean todas las posibles triadas {(1, 2, 3), (1, 2, 4), . . . , (1, N -1, N), (2, 3, 4), . . . , (2, N-1, N), . . . , (N-2, N-1, N)}. Cada triada copia, a aquel involucrado con un mayor número de respuestas correctas, un porcentaje aleatorio del número total de respuestas, restringiéndose sólo a aquellas respuestas mal contestadas.

8. Del número total de triadas que se pueden formar con N alumnos {(1, 2, 3), (1, 2, 4), . . . , (1, N -1, N), (2, 3, 4), . . . , (2, N-1, N), . . . , (N-2, N-1, N)} se toma una al azar. Dicha triada es la única que se involucra en el proceso de copia. Sea (i, j, k) la triada seleccionada, los sujetos j y k copian un porcentaje del total de respuestas dadas por el sujeto i, restringiéndose aquellas respuestas que los sujetos j y k tienen mal contestadas. Lo cual indica que no necesariamente copiaron ambos las mismas respuestas.

9. Del grupo total de triadas que se pueden formar con N alumnos {(1, 2, 3), (1, 2, 4), . . . , (1, N -1, N), (2, 3, 4), . . . , (2, N-1, N), . . . , (N-2, N-1, N)} se toma como subgrupo todas las triadas únicas, es decir que los alumno que conforman la i-ésima triada no se encuentran en ningún otro trío. Dentro de dicho subgrupo de tríos se toma sólo un porcentaje de estás, finalmente este conjunto de tríos son los que se involucran en el proceso de copia. La copia dentro de cada trío se crea a partir de la selección del elemento de la triada con un mayor número de respuestas correctas, los otros dos elementos copian al representante las mismas respuestas, seleccionadas al azar.

1.2. Análisis exploratorio

Como ya se mencionó se cuenta con una base de datos, proveniente de los resultados de la prueba **ENLACE** del 2007, aplicada en los estados Aguascalientes, Nayarit y Puebla, a grupos de tercero a sexto grado de primaria y tercero de secundaria. Cada examen cuenta con preguntas sobre temas de español y matemáticas. En el cuadro 1.5 se muestra un desglose de las características de los datos.

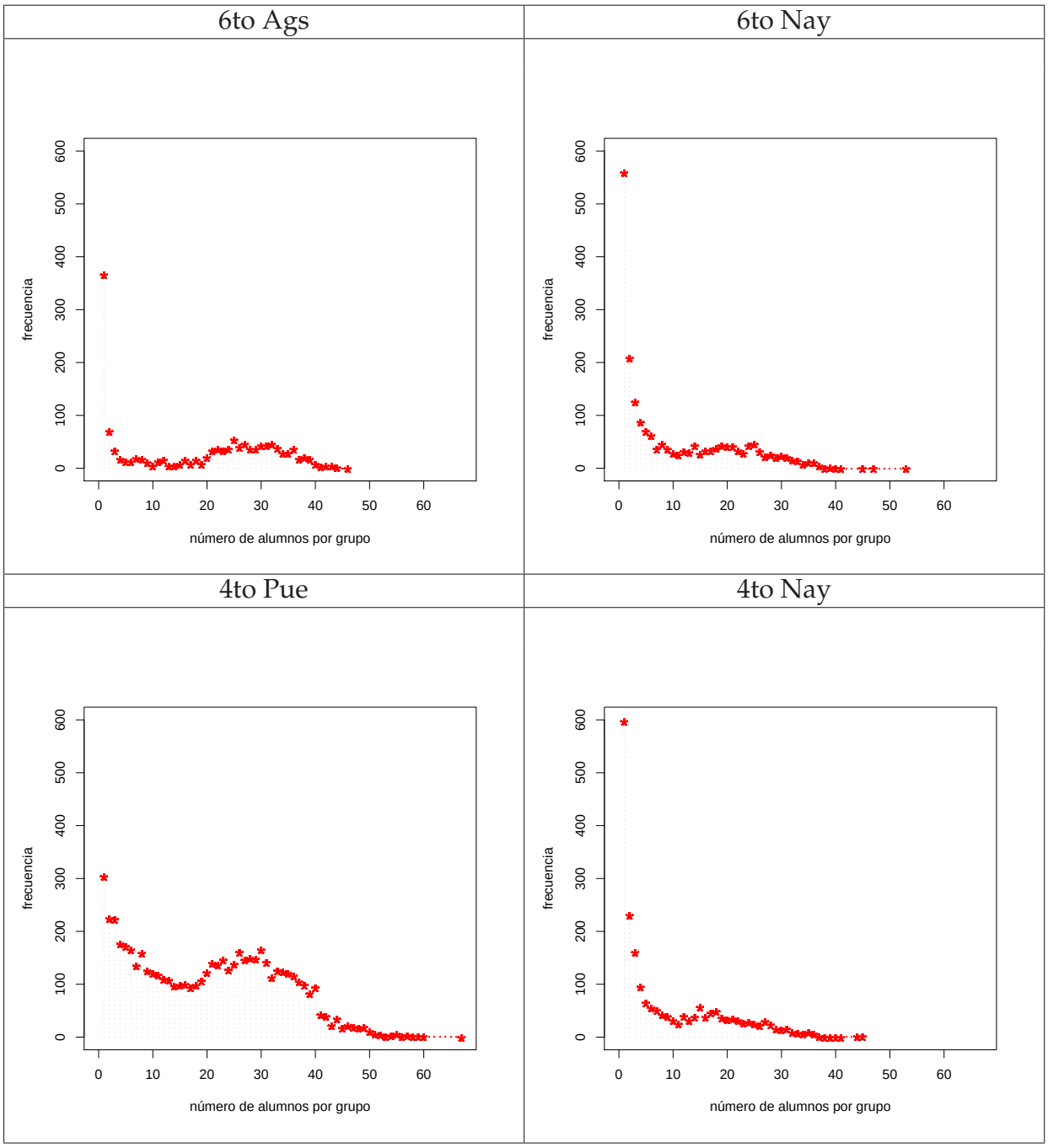
	Ags	Nay					Pue
	6to p	3ro p	4to p	5to p	6to p	3ro s	4to p
Num. escuelas	690	1,085	1,039	1,068	1,028	465	4,284
Num. grupos	1,405	1,984	2,102	1,964	2,041	872	5,755
Num. alumnos	24,376	19,233	19,029	19, 812	20,726	15,417	115,247
Num. ítems en examen	125	101	127	118	125	138	127
Num. ítems anulados[1]	2	3	2	5	2	11	2

Cuadro 1.5: Base de datos obtenida de la prueba **ENLACE** 2007.

Dado que la prueba **ENLACE** se aplica en las diferentes escuelas primarias a nivel nacional (CONAFE, general, indígena y particular), así como también en las distintas escuelas secundarias (general, particular, telesecundaria y técnica), se cuenta con una gran diversidad de grupos y tamaños. Siguiendo esta línea el cuadro 1.6 muestra la variedad que existe en el tamaño de los grupos evaluados. Se observa que constan más de 300 grupos con un solo estudiante, lo que genera la interrogante de si los datos están bien capturados. Asimismo se puede apreciar la gran diversidad de tamaños de grupos existentes en los estados, esto es, existen grupos muy grandes con hasta 50 ó 60 alumnos.

Como se muestra en el cuadro 1.5 se tienen más datos (trenes de respuesta) del estado de Puebla; es por esta razón que son estos datos los utilizados para realizar el análisis exploratorio. Usando los resultados del *II Conteo de Población y Vivienda 2005*, realizado por el **I**nstituto **N**acional de **E**stadística y **G**eografía (**INEGI**) y datos proporcionados por el **C**onsejo **N**acional de **P**oblación (**CONAPO**).

[1]Preguntas que no fueron tomadas en cuenta para la evaluación de los alumnos.



Cuadro 1.6: Frecuencia de grupos, según el número de alumnos.

La información con la que se cuenta se puede desglosar como sigue:

$$
\begin{aligned}
&\textit{por alumno} = \begin{cases} - \textit{indicador de copia} \ (\textit{si o no}) \\ - \textit{calificacion en espanol} \\ - \textit{calificacion en matematicas} \end{cases} \\
&\Downarrow \\
&\textit{por grupo} = \begin{cases} - \textit{tamano} \ (\textit{numero de alumnos en el grupo}) \\ - \textit{porcentaje de alumnos que copian} \\ - \textit{calificacion promedio en espanol} \\ - \textit{calificacion promedio en matematicas} \\ - \textit{nivel de tamano} \ \{\textit{muy chico}, \ \textit{chico}, \ \textit{mediano}, \ \textit{grande}\} \\ - \textit{nivel de copia} \ \{\textit{nada}, \ \textit{muy poco}, \ \textit{poco}, \ \textit{medio}, \ \textit{mucho}\} \end{cases} \\
&\Downarrow \\
&\textit{por escuela} = \begin{cases} - \textit{tipo} \ \{\textit{conafe}, \ \textit{general}, \ \textit{particular}, \ \textit{indigena}\} \end{cases} \\
&\Downarrow \\
&\textit{por municipio} = \begin{cases} - \textit{grado de marginacion} \ \{\textit{muy alto}, \ \textit{alto}, \ \textit{medio}, \ \textit{bajo}, \ \textit{muy bajo}\} \end{cases}
\end{aligned}
$$

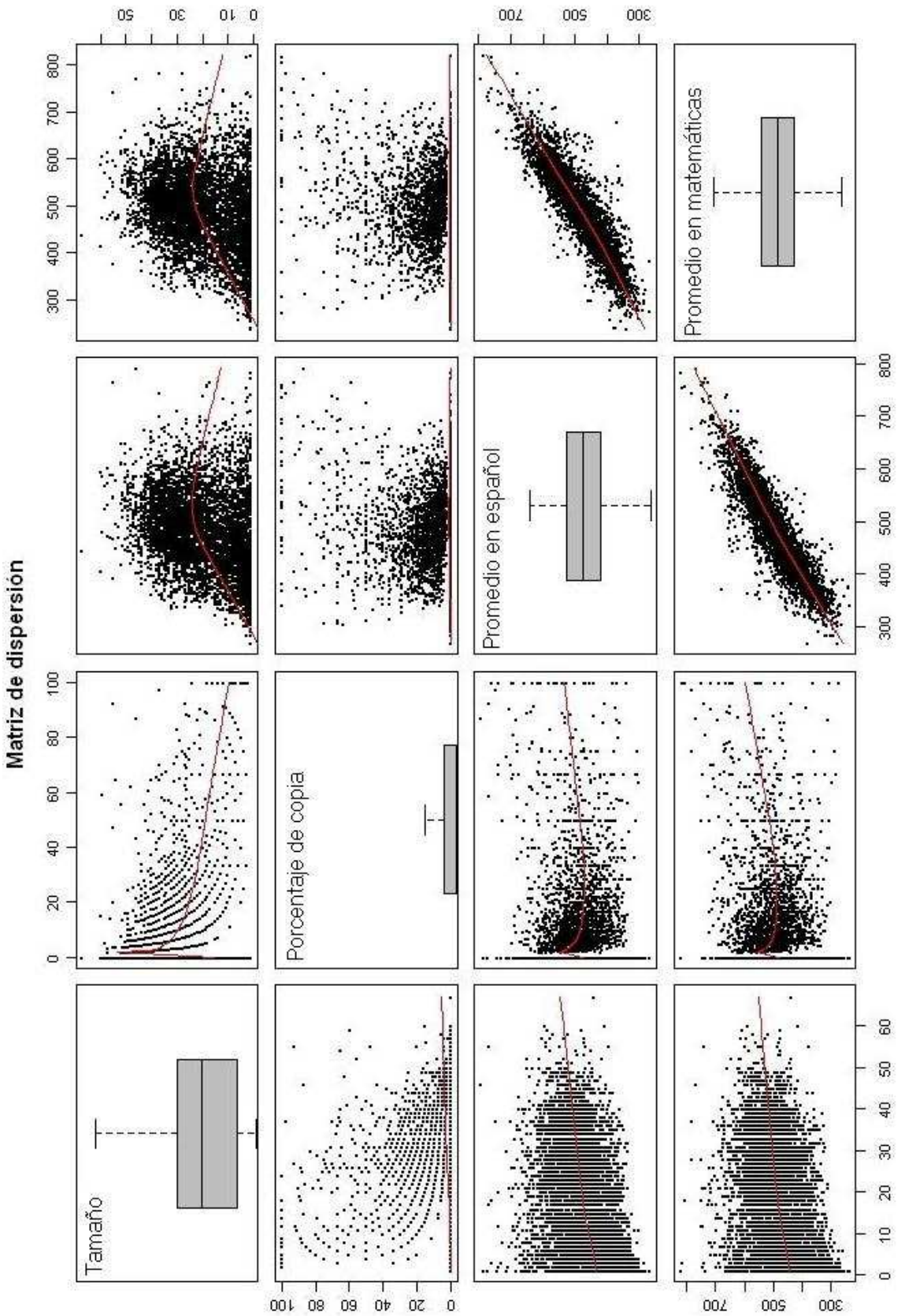
Cabe señalar que el análisis exploratorio se hace a nivel de grupo. Como un primer acercamiento se obtiene el cuadro 1.7, donde se muestra la relación que existe entre las variables no categóricas, estas son: tamaño (número de alumnos en el grupo), porcentaje de alumnos que copia, calificación promedio en español y calificación promedio en matemáticas.

Se puede resaltar el hecho que las calificaciones de español y matemáticas son bastante similares entre ellas. Incluso el comportamiento de las mismas con respecto al resto de las variables de interés es muy similar. Por ejemplo, la variabilidad de ambas calificaciones disminuye conforme el tamaño del grupo crece, aunque también hay que tomar en cuenta el hecho de que existen muy pocos grupos con muchos alumnos. Incluso con respecto al porcentaje de copia no se muestra mucha diferencia entre las calificaciones. Por otro lado, el hecho de que las calificaciones estén centradas en 500 no es de sorprender ya que así fueron normalizadas por los **SEP**.

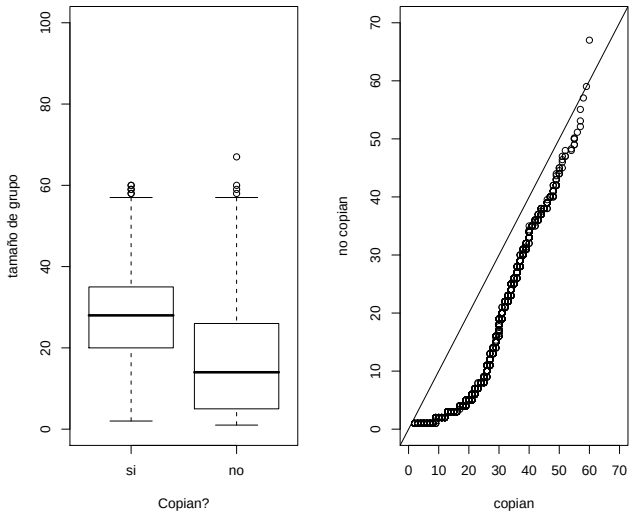
Sobre la grafica de interacción entre las variables tamaño y porcentaje de copia, es notorio el particular comportamiento de los datos, no obstante se observa la existencia de copia en todos los tamaños de grupo, sin embargo hay una mayor densidad de datos entre los grupos con 20 a 45 alumnos. Así como también una mayor densidad de datos en el rango entre 0 % y 25 % de copia.

Una de las variables con mayor importancia es aquella que mide el porcentaje de copia. Para entender un poco como se comporta el porcentaje de copia, se muestran las graficas de los cuadros 1.8, 1.9 y 1.10.

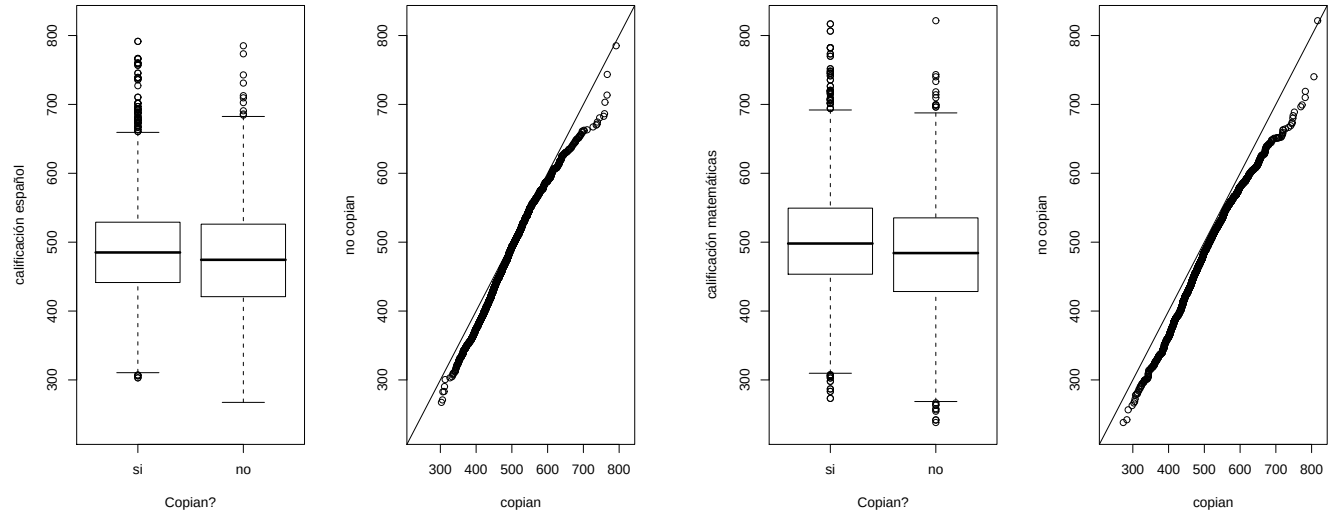
En las graficas mostradas en el cuadro 1.8 se separan los resultados de los alumnos detectados como sospechosos de copia y los que no fueron detectados. Se puede resaltar que la copia se reporta en grupos un tanto grandes.



Cuadro 1.7: Matriz de dispersión.



a) % de copia vs. tamaño de grupo

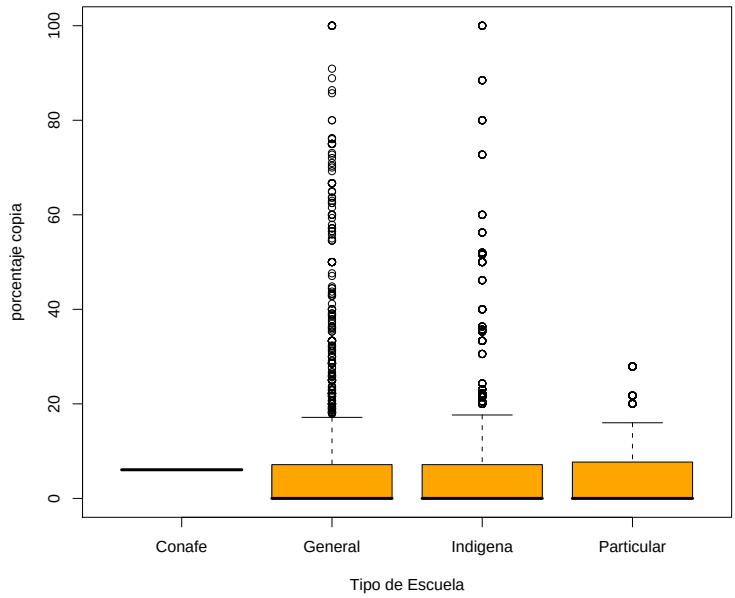


b) % de copia vs. calif. en español **c)** % de copia vs. calif. en matemáticas

Cuadro 1.8: Resultados para la variable de porcentaje de copia.

El hecho que las graficas de *qqplots* de las figuras b) y c) sean muy cercanas a la diagonal, indica que las calificaciones (ya sea de español ó matemáticas) tanto en alumnos que copiaron como los que no lo hicieron, son muy semejantes entre si. Por otro lado, la grafica de *qqplot* de la figura a) es muy desigual a la diagonal, lo que indica que el tamaño de los grupos entre los que copiaron y los que no lo hicieron, son muy desiguales entre si.

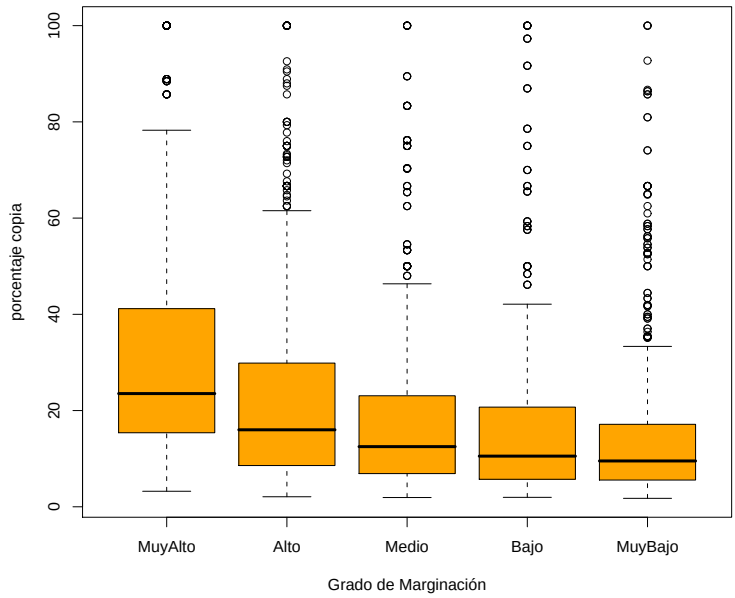
En el cuadro 1.9 se puede ver que en escuelas tipo general es donde existe un mayor número de porcentajes de copia, en comparación a escuelas tipo conafe, indígena y particular. Por otro lado, el cuadro 1.10 muestra la relación que existe entre el grado de marginación y el porcentaje de copia detectado. Aparentemente mientras menor sea el grado de marginación, del municipio donde provenga el grupo, es también menor el nivel de copia detectado. Hay que tener cuidado con está interpretación porque existen otras variables, sobre los diferentes tipos de escuelas y grados de marginación, que hay que tomar en cuenta.



longitud: conafe= 1, general= 1657, indígena= 209, particular= 49

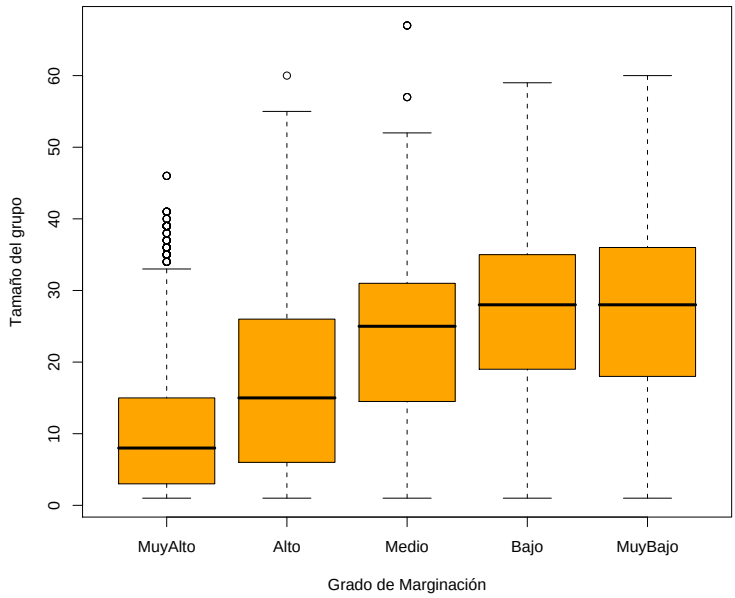
Cuadro 1.9: Porcentaje copia por tipo de escuela.

Por ejemplo, una variable importante es el tamaño del grupo. En cuadro 1.11 muestra la relación entre el tamaño del grupo y el grado de marginación, donde se puede observar que entre menor sea el grado de marginación, los grupos evaluados son cada vez más grandes. Dentro de los datos estudiados se tiene un mayor número de escuelas tipo general, seguidas de las escuelas particulares. Por lo tanto no es de sorprender que el cuadro 1.12 reporte una mayor variabilidad en los tamaños de grupos para esté tipo de escuelas.



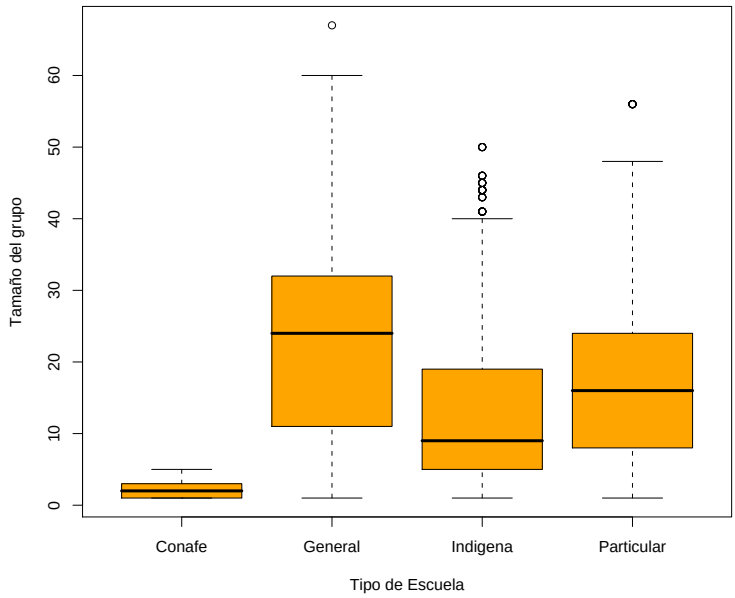
longitud: muy alto= 133, alto= 795, medio= 259, bajo= 263, muy bajo= 466

Cuadro 1.10: Porcentaje copia por grado de marginación.



longitud: muy alto= 537, alto= 2484, medio= 707, bajo= 606, muy bajo= 1266

Cuadro 1.11: Tamaño de grupo por grado de marginación.



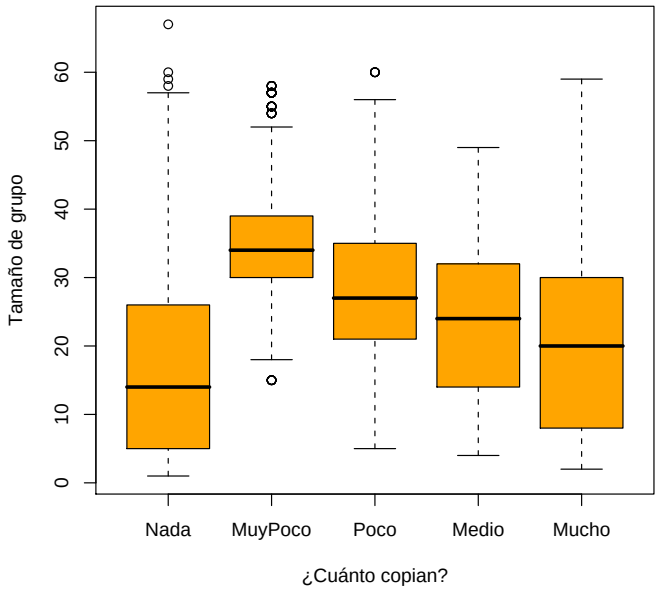
longitud: conafe= 191, general= 4221, indígena= 774, particular= 477

Cuadro 1.12: Tamaño de grupo por tipo de escuela.

Existen grupos donde no se detectó copia y sería un error no contemplar estos en el estudio, por lo tanto se discretizo el nivel de copia de los grupos, usando información proporcionada por los cuartiles (Q_1 = 25 % de los datos, Q_2 = 50 % de los datos, Q_3 = 75 % de los datos) las categorías según el nivel de copia son:

- Nada= 0
- Muy poco= $(0, Q_1]$
- Poco= $(Q_1, Q_2]$
- Medio= $(Q_2, Q_3]$
- Mucho= $(Q_3$, máximo valor]

Del cuadro 1.13 se puede observar que el promedio del tamaño de los grupos donde no se copió es menor que el promedio del tamaño de los grupos donde se copio. De hecho llama la atención como el promedio del tamaño de los grupos va disminuyendo conforme copian de muy poco a mucho, lo cual sugiere que es más "fácil" copiar en grupos pequeños que en grupos grandes.



longitud: nada= 3747, muy poco= 494, poco= 795, medio= 148, mucho= 479

Cuadro 1.13: Tamaño de grupo por nivel de copia.

1.3. Visualización de datos

En la sección anterior se trabajó con un par de características a la vez. Por otro lado, en esta sección se trata de obtener una visualización de todos los trenes de respuesta. Al trabajar con datos multidimensionales, la primera necesidad que surge es la visualización de los mismos con la simple intención de entender el comportamiento de los datos. ***M**ulti-**D**imensional **S**caling* (**MDS**) es una técnica a menudo utilizada en la visualización de información para explorar las similitudes o diferencias en los datos.

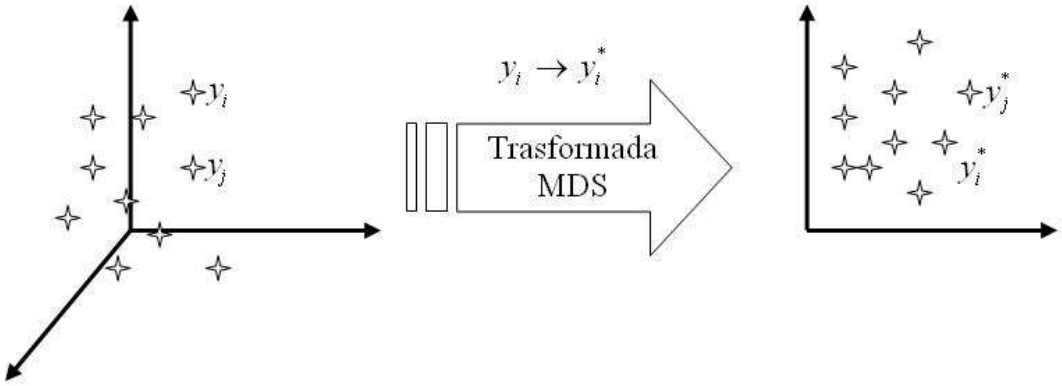
La idea general de **MDS** es reducir la dimensión de los datos, ver cuadro 1.14, con una transformación tal que

$$d(y_i^\star, y_j^\star) \approx d(y_i, y_j)$$

donde se busca

$$\text{mín} \sum_{i,j} \left(d(y_i, y_j) - d(y_i^\star, y_j^\star) \right)^2$$

Se tiene un conjunto con N objetos y una forma de determinar la disimilaridad entre cada par de objetos, con lo cual se crea una matriz de dimensiones R_{NxN}, donde los valores de R_{NxN} estan dados por $r_{ij} = d(y_i, y_j)$. **MDS** trata de buscar una configuración de N puntos en el espacio euclidiano q-dimensional, de tal forma que el k-ésimo punto, en el nuevo espacio, represente al k-ésimo objeto.



Cuadro 1.14: Escalamiento multidimensional.

Ahora bien, si y_i e y_j son cadenas de texto ¿cómo definir $d(y_i, y_j)$? Para implementar **MDS** con datos provenientes de la aplicación de un examen, se ideó como medida de distancia entre los trenes de respuesta (donde cada tren de respuesta pertenece a un alumno) $d(y_i, y_j)$ = número de ítems distintos entre cada pareja de alumnos (y_i, y_j).

Cabe señalar que se trabajo con otras medidas de distancia. Por ejemplo, siguiendo una distribución binomial, donde se toma en cuenta el número de respuestas que ambos trenes de respuesta (y_i, y_j) tienen mal y el número de respuestas que ambos trenes de respuesta (y_i, y_j) tienen mal e igual. Otra medida estaba basada en la suma del logaritmo de la probabilidad de contestar bien el i-ésimo ítem, y la suma del logaritmo de la probabilidad de no contestar bien. Sin embargo estás medidas de distancia no reportaron mejorías significativas.

1.3.1. Experimentos

Con la finalidad de mostrar el alcance del método **MDS** en cadenas de texto, se realizaron un par de experimentos usando datos simulados y datos reales. A continuación se presentan los resultados de datos simulados.

Datos simulados

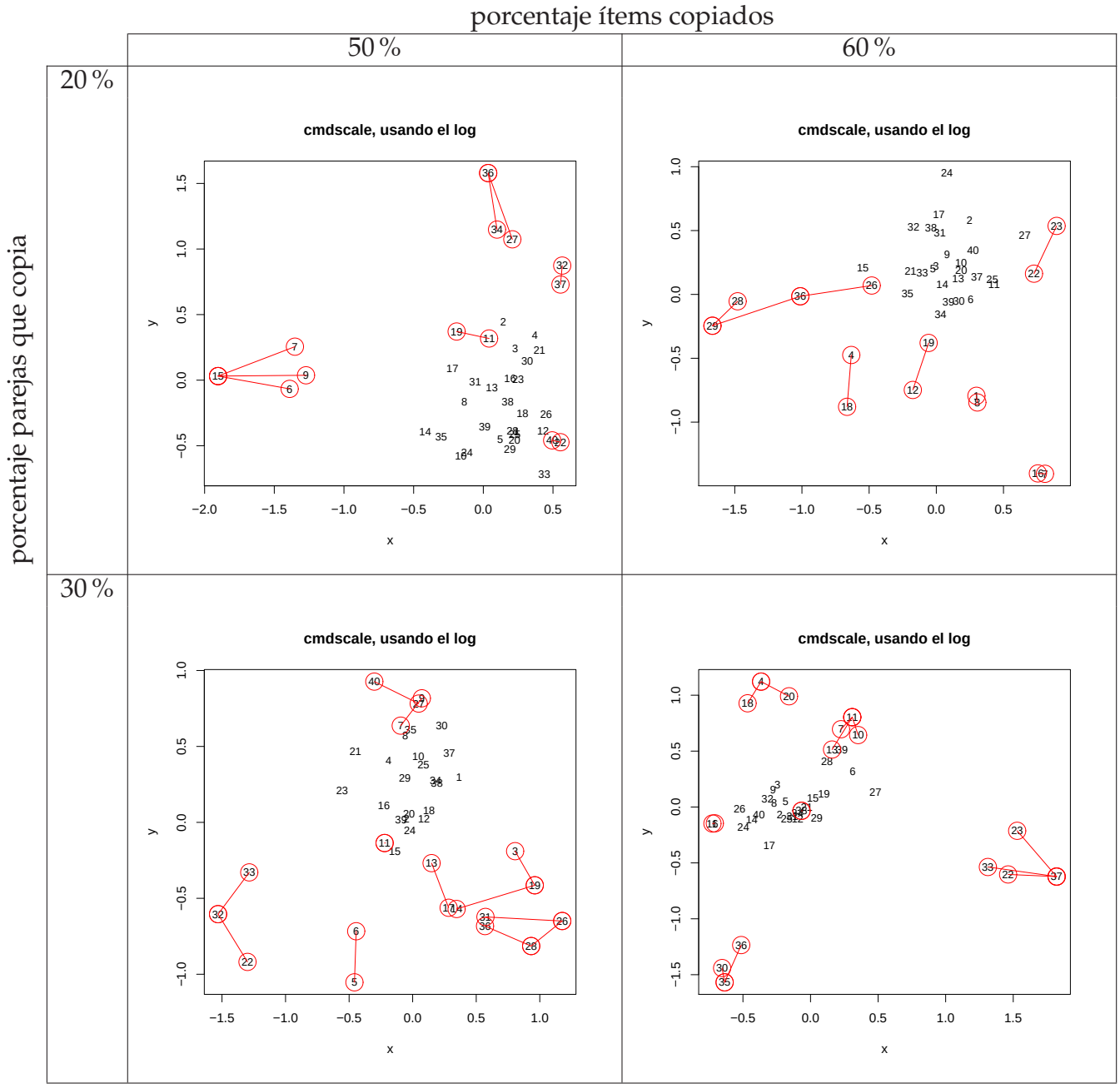
Con el fin de realizar un par de ejemplos "controlados", se generan exámenes siguiendo una distribución uniforme según lo mostrado en la sección 1.1.1, mientras que el proceso de copia se realiza según la opción 3 de la sección 1.1.2.

Se genera un grupo con las siguientes características: 40 alumnos, 80 ítems, para 20 % y 30 % de parejas que realizan copia, 50 % y 60 % de ítems copiados, por las parejas. En el cuadro 1.15 se muestran los resultados obtenidos. Para cada experimento se muestra la grafica de la transformación logarítmica sobre la matriz de distancias, antes de aplicar el método **MDS**. Con un círculo se resalta a los alumnos que participaron en alguna de las parejas que realizaron copia, mientras que con una línea se une a los alumnos pertenecientes a cierta pareja. De las graficas se puede observar que, cada

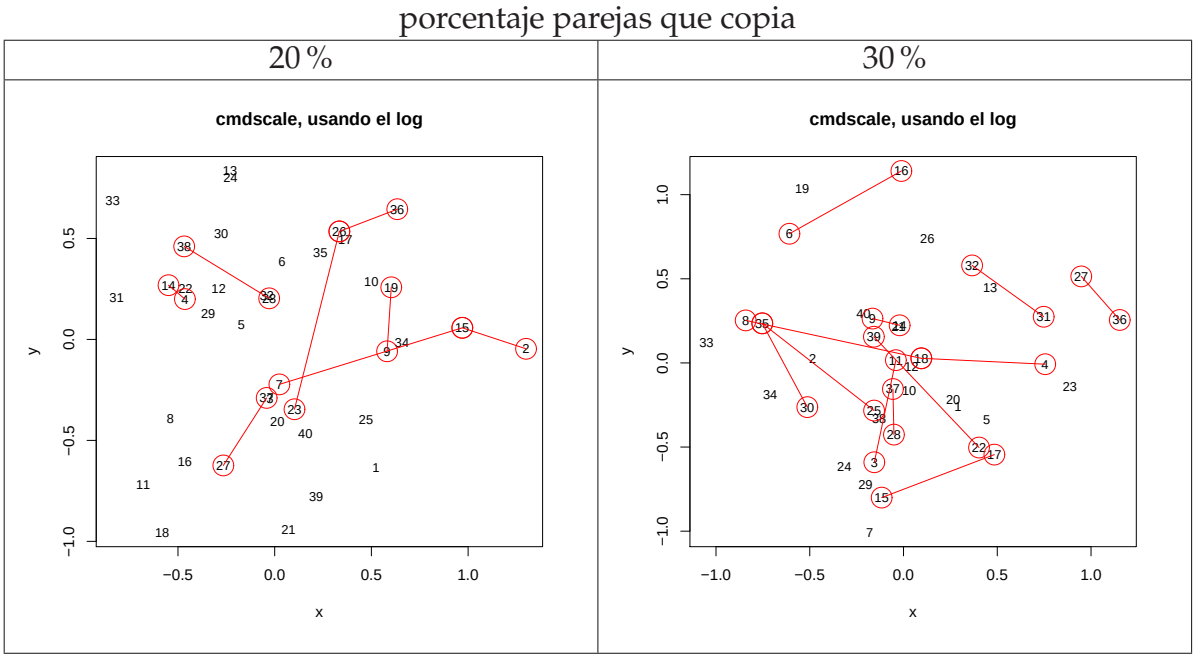
una de las parejas que realizaron copia se encuentra en las periferias del resto de los alumnos que no están involucrados en la copia.

El hecho de que los alumnos que copian se encuentren en la periferia, refleja lo distinto que pueden llegar a ser de aquellos que no copiaron. Lo que indica que es posible llegar a separar o discriminar entre ambos casos, alumnos que copian y alumnos que no copian.

Sin embargo, dado que **MDS** es una técnica utilizada con la única finalidad de visualizar datos y obtener un panorama más especifico del comportamiento de los mismos, no es de extrañar que para un porcentaje de ítems copiados relativamente bajo la separación de parejas involucradas no sea tan clara, ver cuadro 1.16.



Cuadro 1.15: Resultados después de aplicar **MDS** a los trenes de respuesta.



Cuadro 1.16: Resultados después de aplicar **MDS** a los trenes de respuesta, con 20 % de ítems copiados.

Datos reales

Para probar el método **MDS** con datos reales se utilizó trenes de respuesta del estado de Puebla. En particular para grupos con 35 y 45 alumnos, ver cuadro 1.17. En datos reales no se tiene información irrevocable a priori de cuales alumnos han copiado, sin embargo como parte de los resultados de la prueba **ENLACE** se indica cuales alumnos son sospechosos de copia, en siguientes capítulos se hablará más sobre esté tema. Usando está información se resalta con un circulo los alumnos sospechosos de copia detectados por la **SEP**. Con esto se puede observar que al igual que en resultados con datos simulados, aquellos alumnos catalogados como sospechosos de copia se encuentran en la periferia del resto de los alumnos.

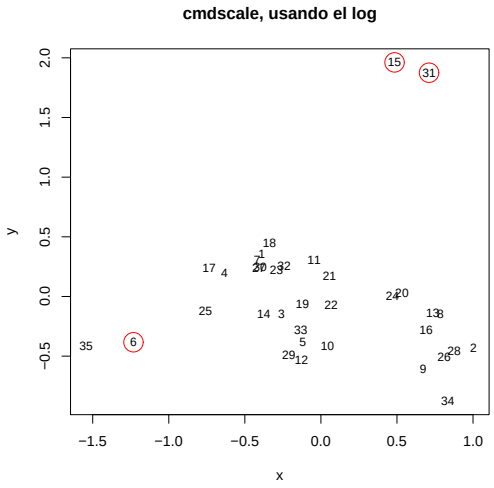
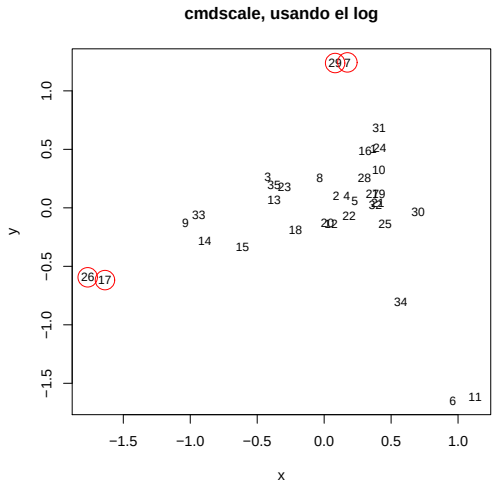
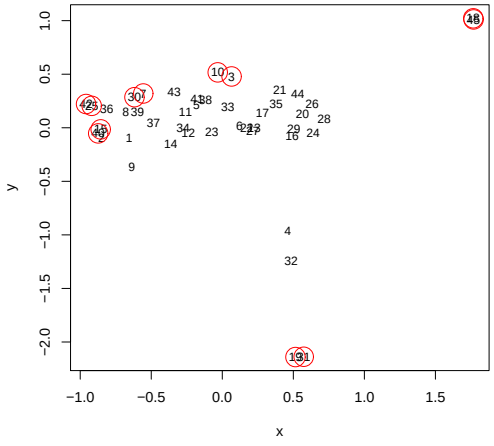
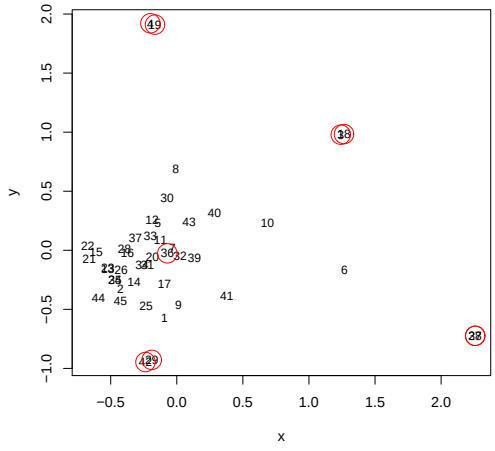
Cabe señalar que dentro de los resultados de la **SEP** no se indica cuales alumnos son los que copiaron entre si, no obstante dada la obvia cercanía mostrada entre varios alumnos sospechosos de copia, se puede concluir que son precisamente estos los que copiaron respuestas entre si.

Esta visualización también ayuda a detectar sospechosos de copia, tomando en cuenta como sospechosos de copia aquellos alumnos que se encuentran en las periferias del conjunto de alumnos, por ejemplo:

- En la grafica (b) para N= 35 se muestran los alumnos 6 y 11 (parte inferior derecha de la imagen) relativamente juntos, al indagar en sus respectivos trenes de respuesta, se encontró que existe un 29.6 % de respuestas mal contestadas e igual entre ambos.

- En la grafica (a) para N= 45 se muestran los alumnos 4 y 32 (parte central de la imagen) muy apartados del resto, al indagar en sus respectivos trenes de respuesta, se encontró que existe un 32.8 % de respuestas mal contestadas e igual entre ambos.

En los experimentos se puede observar que **MDS** reporta buenos resultados, es decir llega a mostrar en las periferias a los alumnos involucrados en copia, pero en datos reales no es suficiente con esta herramienta para discriminar entre alumnos sospechosos de copia y alumnos no sospechosos. Sin embargo **MDS** cumple con el propósito de mostrar el comportamiento de los datos. Antes de trabajar de lleno con ellos.

N=	a)	b)
35		
45		

Cuadro 1.17: Resultados después de aplicar **MDS** a los trenes de respuesta reales.

1.4. Estructura de tesis

El presente documento de tesis cuenta con la siguiente estructura:

El capítulo 2 enlista y explica los modelos actuales y típicamente utilizados para la detección de copia. Existe una serie de herramientas estadísticas para ayudar en la detección de posibles copias, las cuales proporcionan una cuantificación de la probabilidad que una irregularidad se pueda atribuir puramente a la casualidad.

El capítulo 3 muestra una de las contribuciones de esta tesis, la cual es, detección de copia por medio de tríos. El problema de detección de copia es un caso de estudio de varios años atrás, donde originalmente se basa en evaluar el comportamiento de los sujetos examinados por medio de parejas.

El capítulo 4 muestra otra de las contribuciones de esta tesis, la cual es, explorar el comportamiento del grupo examinado en el proceso de copia en un cierto examen.

En el capítulo 5 se exponen las conclusiones y contribuciones de este proyecto de tesis.

En el apéndice A se expone el proceso de evaluación a nivel nacional por parte de la **SEP** por medio del proyecto **ENLACE**.

Capítulo 2

Índices actuales para detección de copia

2.1. Introducción

Típicamente el problema de detección de copia ha sido abordado mediante el uso de pruebas de hipótesis, basados en diferentes estadísticos de prueba. Varios de estos estadísticos se definen con base en las respuestas de los sujetos sospechosos de haber copiado y el sujeto a quien presuntamente se le ha copiado. Algunos ejemplos son los estadísticos de copia *k-index* [4] y *scrutiny* [11].

En general, los estadísticos propuestos en la literatura operan bajo la hipótesis nula que no existe copia, esto es, que los alumnos contestan sus exámenes de forma independiente. Bajo hipótesis nula, los estadísticos siguen una distribución conocida, por ejemplo, distribución *binomial* o distribución *normal*.

La probabilidad con la que se está dispuesto a correr el riesgo de cometer un error de tipo I, rechazar la hipótesis nula cuando está es cierta (falso positivo), se llama nivel de significancia. Esta probabilidad se denota por α. Mientras que un error tipo II, consiste en no rechazar la hipótesis nula cuando está es falsa (falso negativo).

Por la naturaleza del proceso de copia este tipo de estudios no está bien definido, es imposible comprender como se desarrolla el proceso para identificar a los sujetos que copian. El proceso está sujeto a diferentes condiciones tales como el número de ítems por examen, tamaño del grupo a examinar, el número de sujetos que copian y el número de ítems copiados, así como la técnica que usaron para copiar.

Los métodos estadísticos, a describir, se utilizan para evaluar el grado de inusual acuerdo entre las respuestas incorrectas en una prueba de selección múltiple de dos personas examinadas. Sea *s: source* el sujeto/alumno al cual se le plagian ciertas respuestas de su examen y *c: copier* el sujeto/alumno que realiza dicho plagio de respuestas. Muchos de los estadísticos, hacen uso de alguno, o ambos, de los Teoremas 1 y 2.

Teorema 1 (Aproximación de la binomial a la normal) *Sea X una variable aleatoria con*

distribución binomial de parámetros n y p

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ para } k \in \{0, \ldots, n\}$$

si n es grande, entonces la distribución de X es aproximadamente normal con esperanza $\mu = np$ *y varianza* $\sigma^2 = np(1-p)$.

Teorema 2 (Teorema del límite central) *Si la variable aleatoria X sigue una distribución con media* μ *y desviación* σ *conocida, entonces para una muestra* $\{X_i\}$

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \sim \mathcal{N}(0,1)$$

se aproxima a una distribución normal estándar, si $n \to \infty$.

A continuación se da una descripción de algunos índices de copia. A excepción del primero, están basados en pruebas de hipótesis.

2.2. Método de diferencias

El método de diferencia es en sí el más sencillo. El número de respuestas concidentes esta dado por

$$h_{cs} = \sum_{j=1}^{M} I_j$$

donde,

$$I_j = \begin{cases} 1 & si \quad u_{cj} = u_{sj} \\ 0 & en \quad otro\ caso. \end{cases}$$

Si h_{cs} es mayor que un valor determinado, entonces se etiqueta a c y s como sospechosos de copia. Típicamente se consideran como copia cuando $\frac{h_{cs}}{M} \geq 0.9$. Cabe señalar que este índice sólo considera respuestas concidentes y no distingue entre respuestas correctas e incorrectas. Un par de personas que contestaron bien todos los ítems serán marcados como sospechosos de copia bajo este método.

2.3. Índice *kappa*

El índice *kappa* [7] se basa en el supuesto que las respuestas son variables aleatorias. Las respuestas dadas por la pareja (c, s) se almacenan en una tabla de dimensiones kxk, donde k es el número de opciones de respuesta a los ítems del examen. Cada celda (x, y) contiene el número de preguntas donde el alumno c contestó con la opción x y el alumno s contestó con la opción y. Las respuestas en común se encuentran en la diagonal principal de la tabla, mientras que aquellas donde no están de acuerdo se

hallan en las celdas restantes.

Si se almacena la cuenta de respuestas en una matriz R_{kxk}, entonces los valores de R_{kxk} están dados por

$$r_{ij} = \sum_{l=1}^{M} I_l(i,j)$$

$$I_l(i,j) = \begin{cases} 1 & si \quad u_{cl} = i \; e \; u_{sl} = j \\ 0 & en \quad otro \; caso \end{cases}$$

Por ejemplo, para un examen con $M = 55$, donde cada ítem tiene $k = 4$ opciones de respuesta, se resumen los trenes de respuesta de una pareja (c, s) en el cuadro 2.1. En la celda $(x = 1, y = 1)$ se indica el número de veces que ambos alumnos coincidieron en la respuesta con opción 1, por otro lado la celda $(x = 3, y = 4)$ indica el número de veces donde el alumno c contestó la opción 3 y el alumno s optó por la opción 4.

		s				
	x\y	1	2	3	4	total
	1	7	2	1	0	10
	2	9	3	5	4	21
c	3	8	2	0	4	14
	4	6	2	2	0	10
	total	30	9	8	8	55

Cuadro 2.1: Tabla de respuestas de la pareja (c, s).

Se define $\pi_{vv} = P(X = v, Y = v)$, $\pi_{v+} = P(X = v)$ y $\pi_{+v} = P(Y = v)$, con X como respuesta del alumno c e Y como respuesta del alumno s. Estas probabilidades permiten definir el parámetro κ para medir el acuerdo entre los examinados:

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e},$$

donde,

$$\pi_o = \sum_v \pi_{vv}$$

es la probabilidad de observar un acuerdo entre las respuestas de los examinados y

$$\pi_e = \sum_v \pi_{v+}\pi_{+v}$$

es la probabilidad de un acuerdo si los examinados trabajan de forma independientemente (i.e. debido al azar).

La variable κ mide la diferencia entre lo que se observa y lo que se espera bajo independencia entre X e Y. Por otro lado, κ toma valores entre $[-1, 1]$; un valor de

1 implica un perfecto acuerdo entre las respuestas de los examinados mientras que valores inferiores a 1 refieren un menor acuerdo entre estos. Si el valor de κ es negativo es una señal que los dos examinados coinciden menos de lo que se esperaba, bajo independencia.

2.3.1. Hipótesis

Se observa que el valor de κ es 0 si las respuestas de la pareja (c, s) son v. a. independientes (lo que implica que no copiaron). Para verificar si existe copia, se toma entonces como hipótesis nula:

$$H_0 : \kappa = 0$$

y por hipótesis alternativa:

$$H_1 : \kappa > 0$$

2.3.2. Estadístico de prueba

Sea $\hat{\kappa}$ el estadístico que se obtiene al remplazar π_o y π_e, en la definición de κ, por los estimadores $\hat{\pi}_o$ y $\hat{\pi}_e$, respectivamente, donde

$$\hat{\pi}_o = \sum_v \hat{p}_{vv}$$

$$\hat{\pi}_e = \sum_v \hat{p}_{v+}\hat{p}_{+v}$$

y $\hat{p}_{vv}$, $\hat{p}_{v+}$, $\hat{p}_{+v}$ son las proporciones de las observaciones en la celda (v, v) del renglón y columna v, respectivamente.

Si el número de observaciones es suficientemente grande entonces, el estadístico $\hat{\kappa}$ se distribuye *normal* [7], con media

$$\mu_{\hat{\kappa}} = \kappa$$

y varianza

$$\sigma^2_{\hat{\kappa}} = \frac{1}{M}\left\{\frac{\hat{\pi}_o}{1-\hat{\pi}_e} + a + b\right\}$$

donde,

$$a = \frac{2}{(1-\hat{\pi}_e)^2}\left\{2\hat{\pi}_o\hat{\pi}_e - \sum_i \hat{\pi}_{ii}(\hat{\pi}_{i+} + \hat{\pi}_{+i})\right\}$$

$$b = \frac{1-\hat{\pi}_o}{(1-\hat{\pi}_e)^3}\left\{\sum_i \sum_j \hat{\pi}_{ij}(\hat{\pi}_{i+} + \hat{\pi}_{+j})^2 - 4\hat{\pi}_e^2\right\}$$

El valor de $\hat{\kappa}$ estandarizada es definido como

$$z_{\hat{\kappa}} = \frac{\hat{\kappa} - \mu_{\hat{k}}}{\sigma_{\hat{\kappa}}}$$

Bajo la hipótesis nula, se tiene que

$$\mu_{\hat{\kappa}} = 0,$$

entonces, la distribución del estadístico de prueba, bajo la hipótesis nula, está dado por

$$z_{\hat{\kappa}} = \frac{\hat{\kappa}}{\sigma_{\hat{\kappa}}} \sim \mathcal{N}(0,1)$$

Finalmente, la evaluación del índice de *kappa* se resumen en el algoritmo 3.

Algorithm 3: Aplicación del índice

Input: Dado s y c
- Calcular $z_{\hat{\kappa}} = \frac{\hat{\kappa}}{\sigma_{\hat{\kappa}}}$;
- Si $P(|Z| > z_{\hat{\kappa}}) < \alpha$ entonces, se marca como sospechoso de copia, con $Z \sim \mathcal{N}(0,1)$ y α nivel de significancia dado;

2.4. Índice *Scrutiny*

El índice *scrutiny* [11] compara los trenes de respuesta de una pareja (c, s); entre las preguntas mal contestadas de cada uno se cuenta el número de respuestas iguales de ambos. Si este número es demasiado alto entonces, se marca como sospechoso de copia.

Para cada pareja de sujetos examinados (c, s), el número de respuestas mal contestadas e igual está dado por

$$W_{cs} = \sum_{j=1}^{M} Z_j$$

donde,

$$Z_j = \begin{cases} 1 & si \quad \textit{ambos } (c,s) \textit{ contestaron mal e igual el item } j \\ 0 & \quad \textit{otro caso} \end{cases}$$

2.4.1. Hipótesis

Sea p_j la probabilidad estimada de contestar con la misma opción errónea. Si no se copia, $\{Z_j\}$ será una v. a. independiente con $P(Z_j = 1) = p_j$, donde $p_j = \frac{\sum_{k\ opc.\ mal} p_{j,k}^2}{\left(1 - p_{j,opc.\ bien}\right)^2}$ y $p_{j,k}$ representa la probabilidad de elegir la opción k para el j-ésimo ítem. La hipótesis nula para evaluar el acuerdo entre los sujetos inspeccionados está formulada por

$$H_0 : P(Z_j = 1) = p_j$$

2.4.2. Estadístico de prueba

Si h denota el número de preguntas que ambos, c y s, tienen mal, entonces bajo H_0 se tiene

$$\begin{aligned} E[W_{cs}] &= E\left(\sum_{j=1}^{M} Z_j\right) \\ &= \sum_{j=1}^{M} E[Z_j] \\ &= h\bar{p} \end{aligned}$$

donde,

$$\bar{p} = \frac{\sum_{j=1}^{M} p_j * I_j}{h},$$

$$I_j = \begin{cases} 1 & si \quad ambos\ (c,s)\ contestaron\ mal\ el\ item\ j \\ 0 & \quad\ otro\ caso \end{cases}$$

y la varianza está dada por

$$\begin{aligned} Var[W_{cs}] &= \sum_{j=1}^{M} Var[Z_j] \\ &\approx h\bar{p}(1-\bar{p}) \end{aligned}$$

Esto último porque se aproxima cada Z_j con una misma *bernoulli* con θ como el promedio de $P(Z_j = 1)$. Con lo anterior se establece que el índice de *scrutiny* esta dado por

$$z_s = \frac{\left(W_{cs} - \frac{1}{2}\right) - h\bar{p}}{\sqrt{h\bar{p}[1-\bar{p}]}} \sim \mathcal{N}(0,1)$$

para h suficientemente grande.

Finalmente, la evaluación del índice de *scrutiny* se resumen en el algoritmo 4.

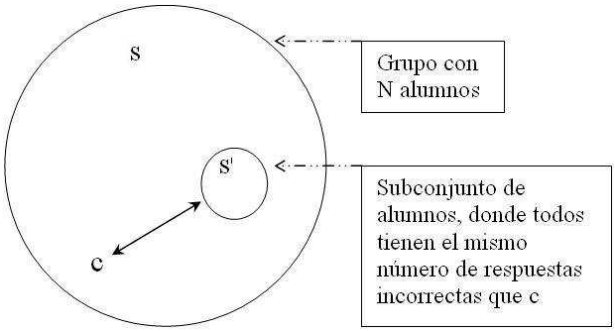
Algorithm 4: Aplicación del índice

Input: Dado s y c
- Calcular $z_s = \frac{\left(W_{cs} - \frac{1}{2}\right) - h\bar{p}}{\sqrt{h\bar{p}[1-\bar{p}]}}$;
- Si $P(|Z| > z_s) < \alpha$ entonces, se marca como sospechoso de copia, con $Z \sim \mathcal{N}(0,1)$ y α nivel de significancia dado;

2.5. Índice K (*k-index*)

Un problema con el índice *scrutiny* es que estima el valor de p_j de manera global. En general la probabilidad de coincidencias en preguntas mal contestadas dependerá por

ejemplo de la habilidad de los alumnos involucrados. Tal vez sea más realista comparar el número de coincidencias entre alumnos con una habilidad similar. Esto último se puede traducir en comparar al alumno *c* contra todos aquellos alumnos con el mismo número de respuestas mal contestadas que él (lo cual no significa que dichos alumnos tengan exactamente las mismas preguntas contestadas incorrectamente por *c*). Para el calculo del índice *K* [11] se calcularán todas las probabilidades condicionales en este subconjunto, ver cuadro 2.2.



Cuadro 2.2: Estructura de *k-index*.

Sea w_c, w_s el número de respuestas incorrectas para los alumnos examinados *c* y *s*, respectivamente y w_{cs} el número de respuestas erroneas coincidentes. El índice *K* se calcula mediante la determinación de la probabilidad que *s* y *c* respondan de manera idéntica con al menos w_{cs} respuestas incorrectas.

En la literatura consultada no existe una definición única de *K*. Por ejemplo los autores Wollack [11] y Holland [4] lo definen de manera distinta. No obstante, todos pretenden calcular

$$P(W_{c's} > w_{cs} | \textit{respuestas de } c, W_c = W_{c'} = w_c). \tag{2.1}$$

2.5.1. Estadístico de prueba

Para calcular 2.1 existen diferentes caminos:

1. Usando la distribución empírica; tomando el porcentaje de alumnos con el mismo número de preguntas mal contestadas y con al menos w_{cs} respuestas coincidentes.
2. Usando una aproximación *binomial* con

$$\bar{p} = \frac{\sum_{j:w_j=w_c} \frac{w_{js}}{N_{w_c}}}{min(w_j, w_s)} \tag{2.2}$$

donde, w_j representa el número de ítems contestadas incorrectamente por el sujeto j, w_{js} es el número de ítems contestadas incorrectamente por ambos alumnos (j, s)

y N_{w_c} es el número de alumnos examinados con el mismo número de respuestas incorrectas que c. Esté valor, $\bar{p}$, hace referencia a una estimación del parámetro θ de la distribución *binomial* igualando su promedio teórico con el promedio muestral.

3. Usando una aproximación *binomial* con una $\bar{p}$ distinta. En [2] y [6] se muestran distintos caminos para su cálculo.

La formula para calcular el valor de K[1]esta dada por

$$z_k = \sum_{a \geq w_{cs}} \binom{min(w_s, w_c)}{a} (\bar{p})^a \, (1 - \bar{p})^{min(w_s, w_c) - a}$$

donde $\bar{p}$ esta dado por 2.2.

Finalmente, la evaluación del índice K se resume en el algoritmo 5.

Algorithm 5: Aplicación del índice

Input: Dado s y c
- Calcular w_{cs};
- Si $P(W_{cs} > w_{cs}) < \alpha$ entonces, se marca como sospechoso de copia, con $W_{cs} \sim \mathcal{B}(\text{mín}(w_s, w_c), \bar{p})$ y α nivel de significancia dado;

2.6. Índice g_2

El índice g_2 [10] es muy similar a la formulación del índice *scrutiny* pero no se limita a respuestas mal contestadas. Se usa:

$$v_{cs} = \sum_{j=1}^{M} I_j$$

donde

$$I_j = \begin{cases} 1 & si \quad u_{cj} = u_{sj} \\ 0 & \quad otro\ caso \end{cases}$$

2.6.1. Estadístico de prueba

Se establece que el índice g_2 esta dado por

$$z_g = \frac{v_{cs} - E[v_{cs}]}{\sqrt{Var[v_{cs}]}} \sim \mathcal{N}(0, 1)$$

[1]La formulación de *k-index* está mal realizada en [11].

Finalmente, la evaluación del índice g_2 se resume en el algoritmo 6.

Algorithm 6: Aplicación del índice

Input: Dado s y c
- Calcular $z_g = \frac{v_{cs} - E[v_{cs}]}{\sqrt{Var[v_{cs}]}}$;
- Si $P(|Z| > z_g) < \alpha$ entonces, se marca como sospechoso de copia, con $Z \sim \mathcal{N}(0,1)$ y α nivel de significancia dado;

2.7. Índice ω

Al igual que el índice g_2, el índice ω [10] no se limita sólo a las respuestas mal contestadas. De hecho es bastante similar a g_2 en su formulación, salvo por el cálculo de la probabilidad de elegir la opción k para el ítem j, $p_{j,k}$. En este caso para calcular dicha probabilidad se hace uso de la teoría de **IRT**, ver sección A.1.2, en particular de su extensión a respuestas multi-opción como *nominal response model* (**NRM**).

Bajo **NRM**, la probabilidad que el i-ésimo sujeto, con un nivel de habilidad θ_i, seleccione la opción k para el j-ésimo ítem esta dado por

$$P(X_{ij} = 1|\theta_i, \zeta_{jk}, \lambda_{jk}) = \frac{\exp\{\zeta_{jk} + \lambda_{jk}\theta_i\}}{\sum_{v=1}^{V} \exp\{\zeta_{jv} + \lambda_{jv}\theta_i\}}$$

donde ζ_{jk} y λ_{jk} son parámetros del ítem; intercepción y discriminación del j-ésimo ítem en la k-ésima opción de respuesta, respectivamente. El valor de θ y los parámetros de los ítems pueden ser estimados usando el método *maximum likelihood*.

2.8. Otros índice

Existen un par de índices adicionales a los anteriormente descritos, no tan populares. Estos son:

ACT Pair1 & Pair2, [1]: El método Pair1 utiliza dos índices para cada par de candidatos: (1) el número de errores coincidentes y (2) la longitud de la cadena más larga de las respuestas coincidentes. El método Pair2 usa: (1) el número de errores en la cadena más larga de las respuestas coincidentes y (2) el número de respuestas idénticas de todos los candidatos.

IMS, OMS and L_z, [5]: Similar a ω al usar teoría **IRT**.

2.9. Comparación de índices

Al final, se tiene una diversidad de índices de copia, desde el índice de copia que sólo cuenta el número de similitudes entre los trenes de respuesta, hasta aquel que hace uso de modelos estadísticos más fuertes. El cuadro 2.3 rescata los pros y contras de cada uno de los índices descritos en este capítulo.

Si se agruparan los índices según sus características, se puede observar que los índices *scrutiny* y *k-index* son bastante parecidos entre si, ambos se basan en las respuestas igual y mal contestadas de la pareja (c, s), sin embargo su diferencia radica en la forma que tienen de calcular el valor p_j, probabilidad de contestar con la misma opción. Por otro lado, se tienen los estadísticos ω y g_2, ambos se basan en las respuestas igual de la pareja (c, s), al igual que los índices *scrutiny* y *k-index*, ambos calculan el valor de p_j de distinta manera.

Uno de los problemas más visibles al trabajar con esté tipo de índices es el hecho que no se toma en cuenta el tamaño del grupo evaluado. Por poner un ejemplo, *k-index* regresa un valor de copia aun en grupos con dos alumnos en el grupo, algo que es muy criticado para esté índice en particular y que se ha observado en la prueba **ENLACE**.

2.9.1. Error tipo I y error tipo II

Existen, al menos, dos maneras para definir el error tipo II: a nivel de parejas de alumnos y a nivel de alumnos.

1.- Crear varias parejas de sujetos de las cuales sólo una porción de estas realiza copia. Dentro de las parejas que copiaron contar el número de parejas detectadas, bajo cierto índice de copia. El cuadro 2.5 presenta resultados bajo esté camino.

2.- Crear un grupo donde un número aleatorio de sujetos copian respuestas. Generar todas las parejas posibles de dicho grupo y evaluarlas, bajo cierto índice de copia. En este caso, es más estricto ya que un solo individuo aparece en varias parejas y es suficiente detectarlo en alguna de estas combinaciones para indicar que es sospechoso de copia. El cuadro 2.6 presenta resultados bajo esté camino.

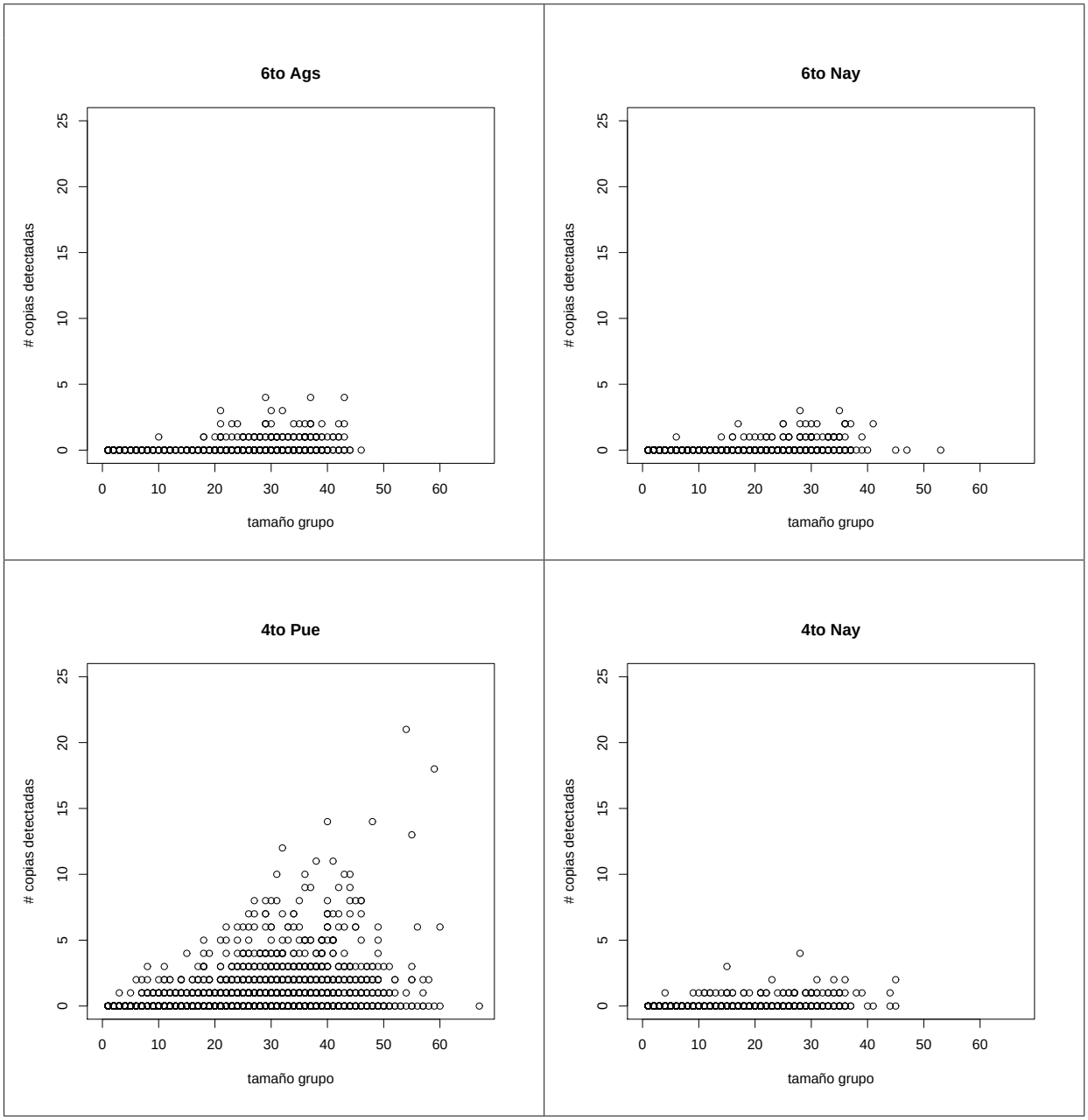
Una importante diferencia entre los dos caminos es que en el segundo el tamaño del grupo influye mucho. Cabe señalar que en la prueba **ENLACE** no se toma en cuenta este aspecto.

Las graficas 2.5 y 2.6 están basadas en 100 grupos simulados con $N = 150$ y $M = 80$ bajo el índice *Scrutiny*. En las graficas se aprecia el fenómeno de disminucioń de variabilidad conforme el número de parejas aumenta.

Como ya se menciono al inicio de esté capítulo, el error tipo I refiere a falsos positivos. En el caso particular de detección de copia, se puede traducir esto como el número de alumnos detectados como sospechosos de copia cuando en realidad no copiaron.

Estadístico	Pros	Contras
diferencias	Es sumamente sencillo de calcular, solamente hay que contar el número de ítems iguales de la pareja (c, s).	Alumnos que contestaron un examen (casi) perfecto, son considerados como sospechosos de copia.
kappa	No se basa en modelos de respuestas dicotomicas, sino en el supuesto de que las respuestas son probabilísticas.	Dado que no hace distinción entre respuestas bien y mal contestadas, alumnos con exámenes (casi) perfectos, son considerados como sospechosos de copia. Por otro lado, necesita la creación de una tabla para cada pareja (c, s) lo convierte en uno de los índices más lentos computacionalmente.
scrutiny	No asume ningún modelo **IRT**. Por lo tanto es fácil de aplicar.	Es necesario estimar el valor del parámetro p de la *binomial*. Dado que el índice está basado en el número de respuestas incorrectas esté debe ser lo suficientemente grande para obtener una estimación fiable de la estimación de la *binomial* a la normal.
k-index	No asume ningún modelo **IRT**. Por lo tanto es fácil de aplicar.	Al basarse en el número de sujetos con el mismo número de respuestas erróneas que c, es necesario contar con un grupo/conjunto de alumnos/sujetos suficientemente grande para una buena estimación de $\overline{p}$. Un ejemplo de está falla se muestra en el cuadro 2.4 donde aun en grupos muy chicos (menos de 15 alumnos) este método reporta sospechosos de copia.
g_2	Compara el número de respuestas iguales de un par de sujetos contra el valor esperado del número de coincidencias.	Requiere la estimación de la probabilidad que c seleccione cada alternativa del ítem.
ω	Esta basado en los modelos **IRT**, en particular en el modelo **NRM**.	El valor de θ y los parámetros de los ítems deben ser estimados. Es difícil estimar bien θ si se copió.

Cuadro 2.3: Resumen de los estadísticos de prueba.



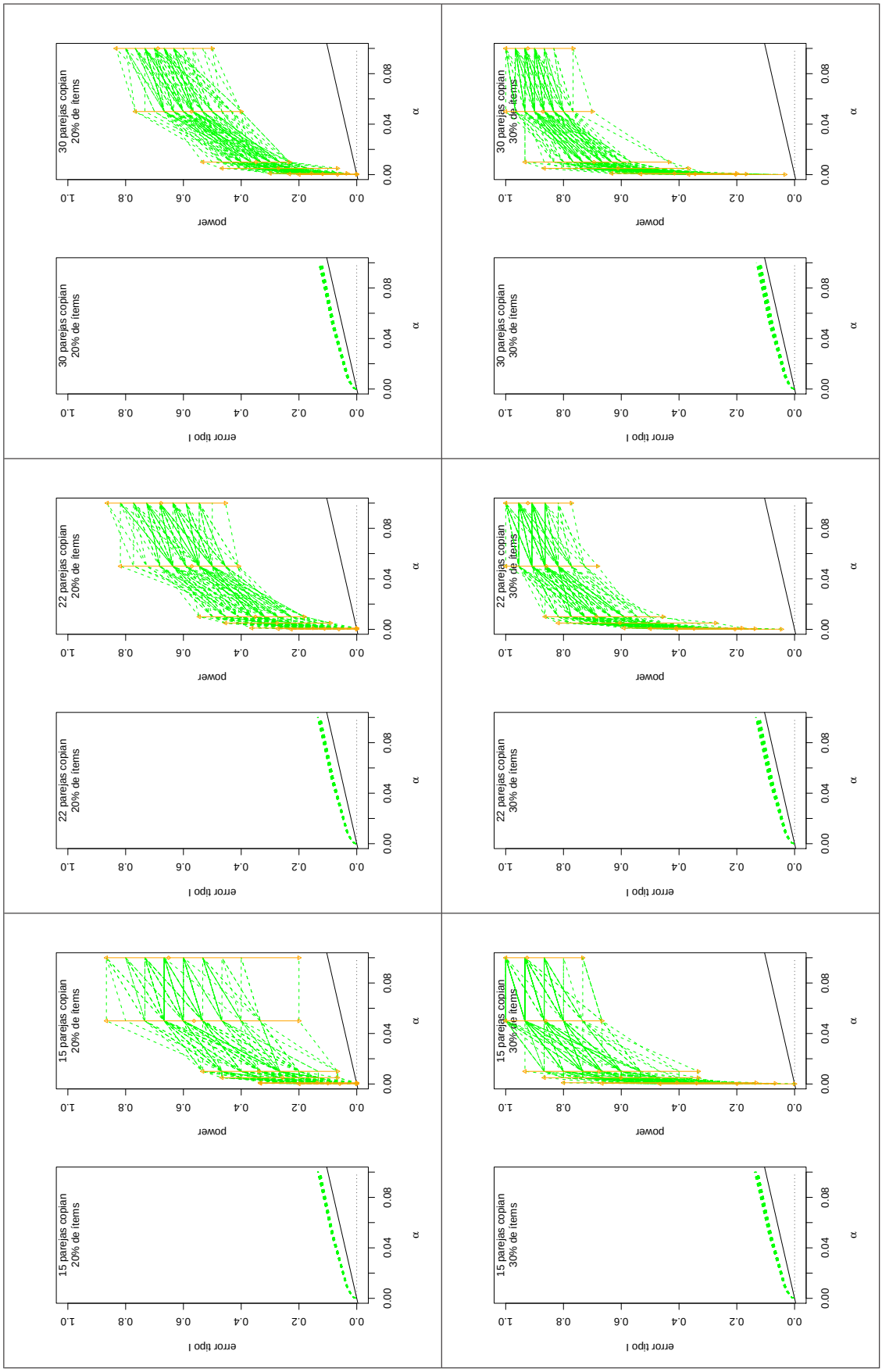
Cuadro 2.4: Tamaño de grupo por el número de copias detectadas, bajo *k-index*.

El error tipo II que refiere a falsos negativos, se traduce como el número de alumnos detectados como no sospechosos de copia cuando en realidad sí copiaron. Mientras que el poder (power) se define como 1-error tipo II, lo cual se traduce como el número de alumnos detectados como sospechosos de copia que se sabe a priori que si copiaron.

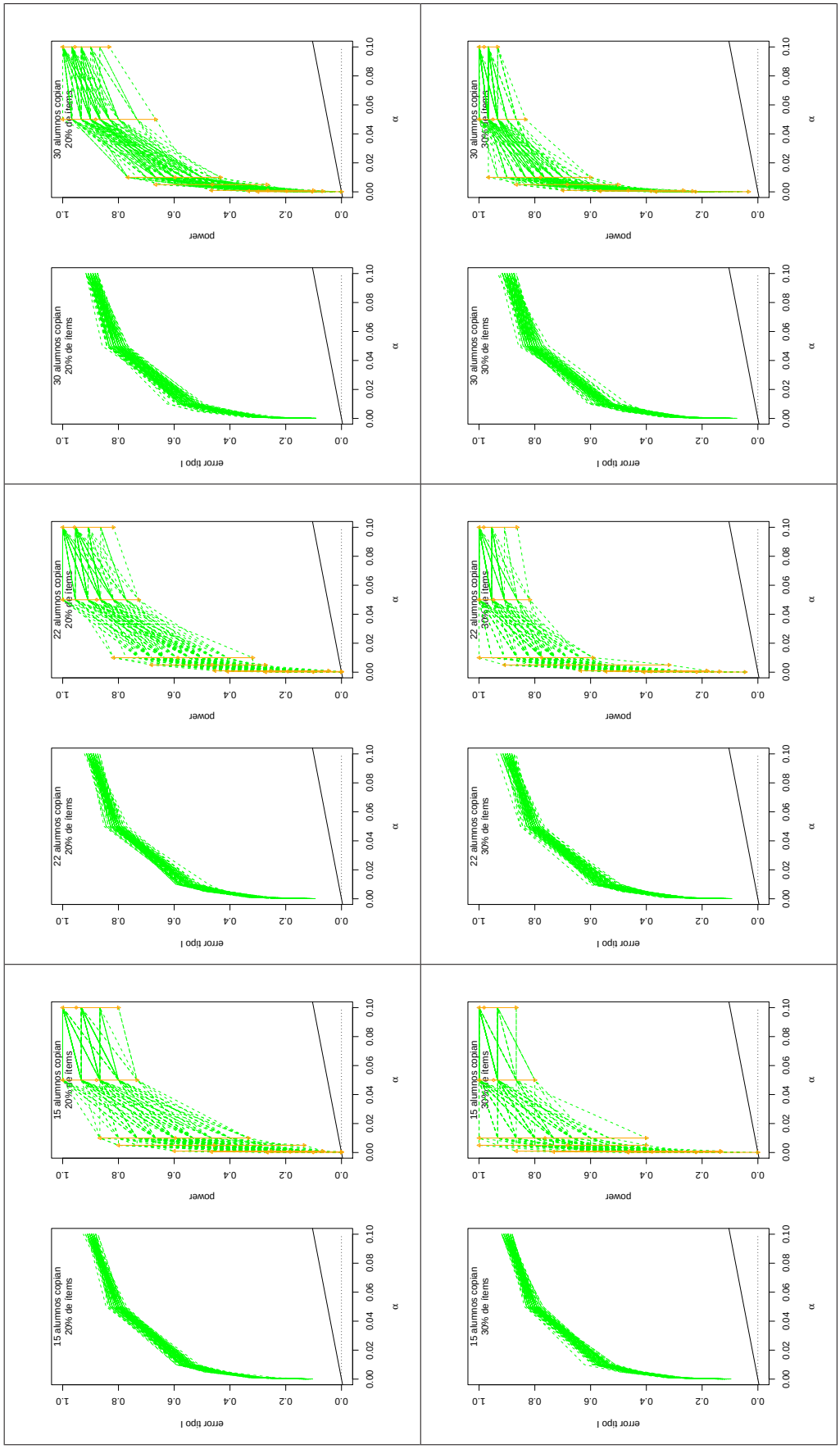
Por otro lado, α indica la probabilidad de cometer un error tipo I. Así que al incrementar α el error tipo I también aumenta, no obstante el error tipo II disminuye lo que se ve reflejado en el aumento del poder. Este comportamiento se puede observar en los cuadros 2.5 y 2.6.

Uno de los experimentos más sencillos para probar el poder, es generar varios grupos donde sólo un alumno copia por grupo y contar el número de veces en que dicho alumno ha sido detectado como sospechoso de copia, bajo distintos niveles de α. Mientras, para el error tipo I se toma al azar un alumno, donde se sabe a priori que dicho alumno no copio y se cuenta el número de veces que ha sido detectado bajo distintos niveles de α.

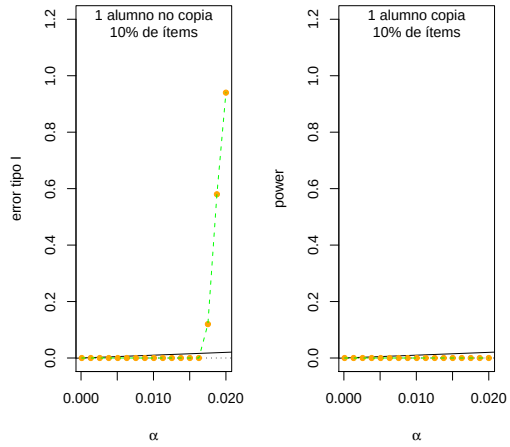
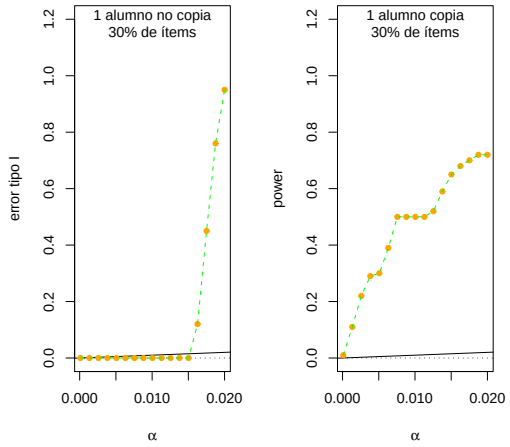
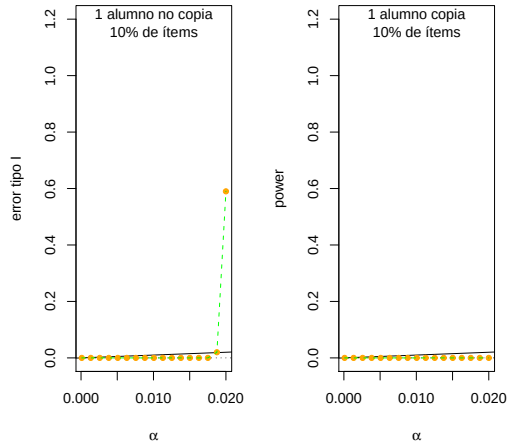
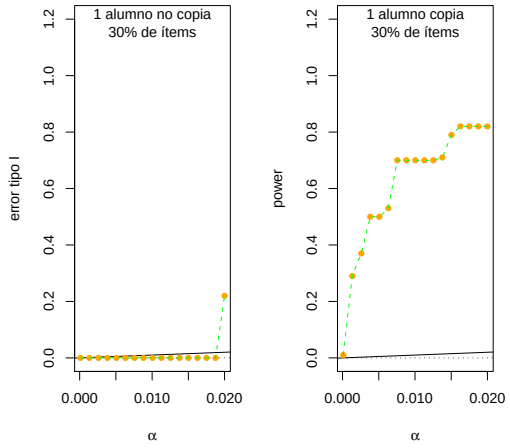
El cuadro 2.7 muestra el resultado de 100 iteraciones con M= 80, distintos N y porcentajes de ítems copiados. Hay que resaltar el hecho de que el comportamiento del error tipo I y del poder varia según el tamaño del grupo. El que varié según el número de ítems copiados no es de sorprender, ya que se espera que entre mayor sea el porcentaje de ítems copiados se facilita el detectar alumnos sospechosos de copia.



Cuadro 2.5: Graficas para *scrutiny*, siguiendo primer camino.



Cuadro 2.6: Graficas para *scrutiny*, siguiendo segundo camino.

N		
15		
45		

Cuadro 2.7: Graficas para *scrutiny*.

2.9.2. Uso de los índices en ENLACE

Como ya se menciono, en la prueba **ENLACE** se utilizan métodos de detección de copia como parte de la evaluación de los resultados. En particular los estadísticos que se utilizó en el 2007 son: *k-index*, *scrutiny* y diferencias. Los umbrales utilizados para cada uno de estos índices [3] se enlistan a continuación.

- Para diferencias, se fija un número máximo de ítems iguales permitidos en cada grado, por lo tanto si se rebasa dicho umbral se cataloga como sospechoso de copia al alumno evaluado. Debido a que los exámenes son de extensión diferente para cada grado evaluado, el valor de dicho umbral es distinto. El cuadro 2.8 muestra el valor del umbral para cada grado escolar evaluado.

Grado	Número de ítems en el examen	Número máximo de ítems permitidos (umbral)	Porcentaje de ítems permitidos
3ro	101	10	9.9 %
4to	127	12	9.4 %
5to	118	11	9.3 %
6to	125	12	9.6 %
3ro (secundaria)	138	13	9.4 %

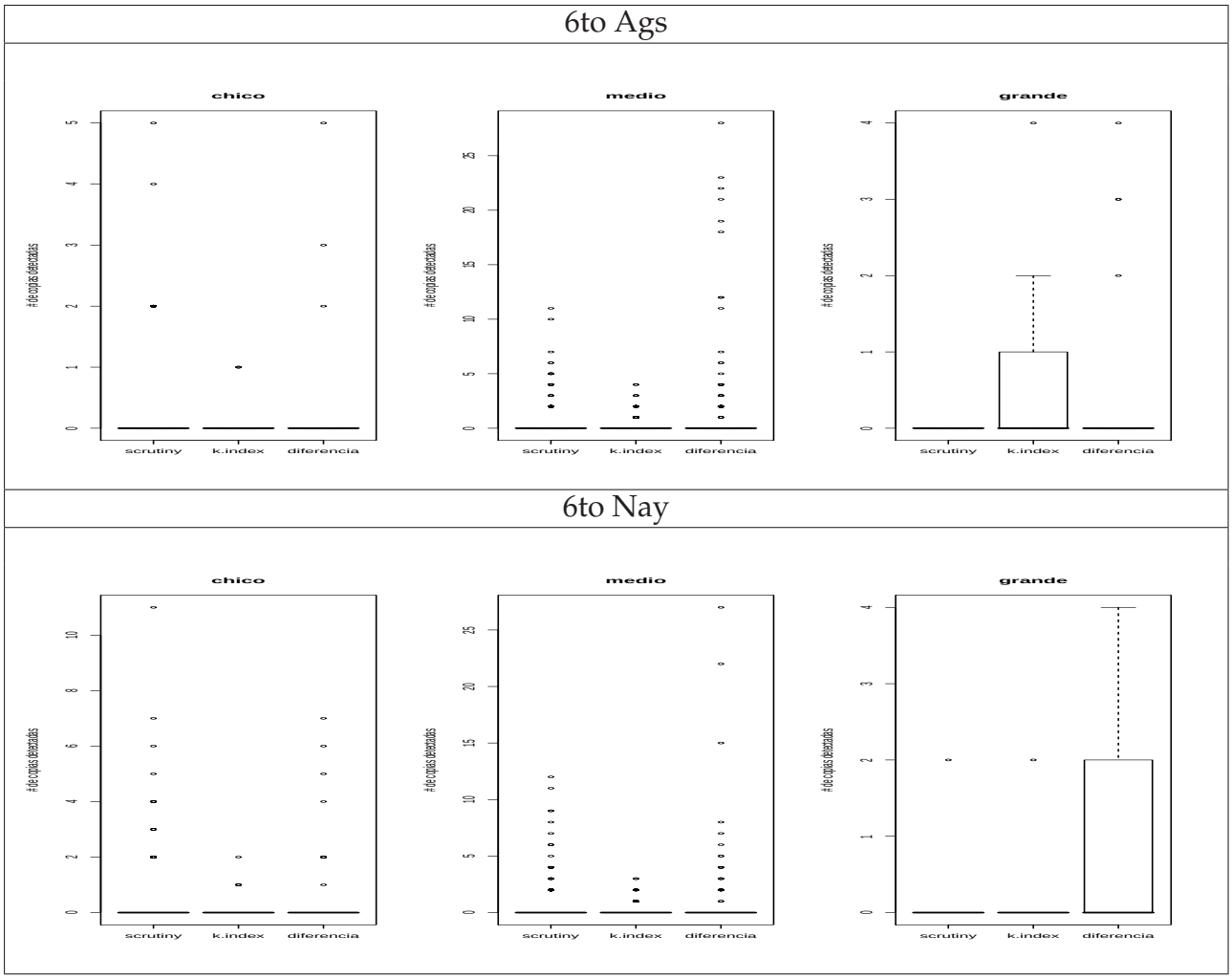
Cuadro 2.8: Valor del umbral para diferencias según grado escolar.

- Para *scrutiny*, se establece un $\alpha = 4.595x10^{-13}$ lo que equivale a tener un valor critico $z^{\star} \approx 7.1$, por lo tanto todo alumno con un valor *scrutiny* ≥ 7.1 es marcado como sospechoso de copia.
- Para *k-index*, en este caso todo aquel sujeto con un valor *k-index* $\leq 4.369x10^{-6}$ es marcado como sospechoso de copia.

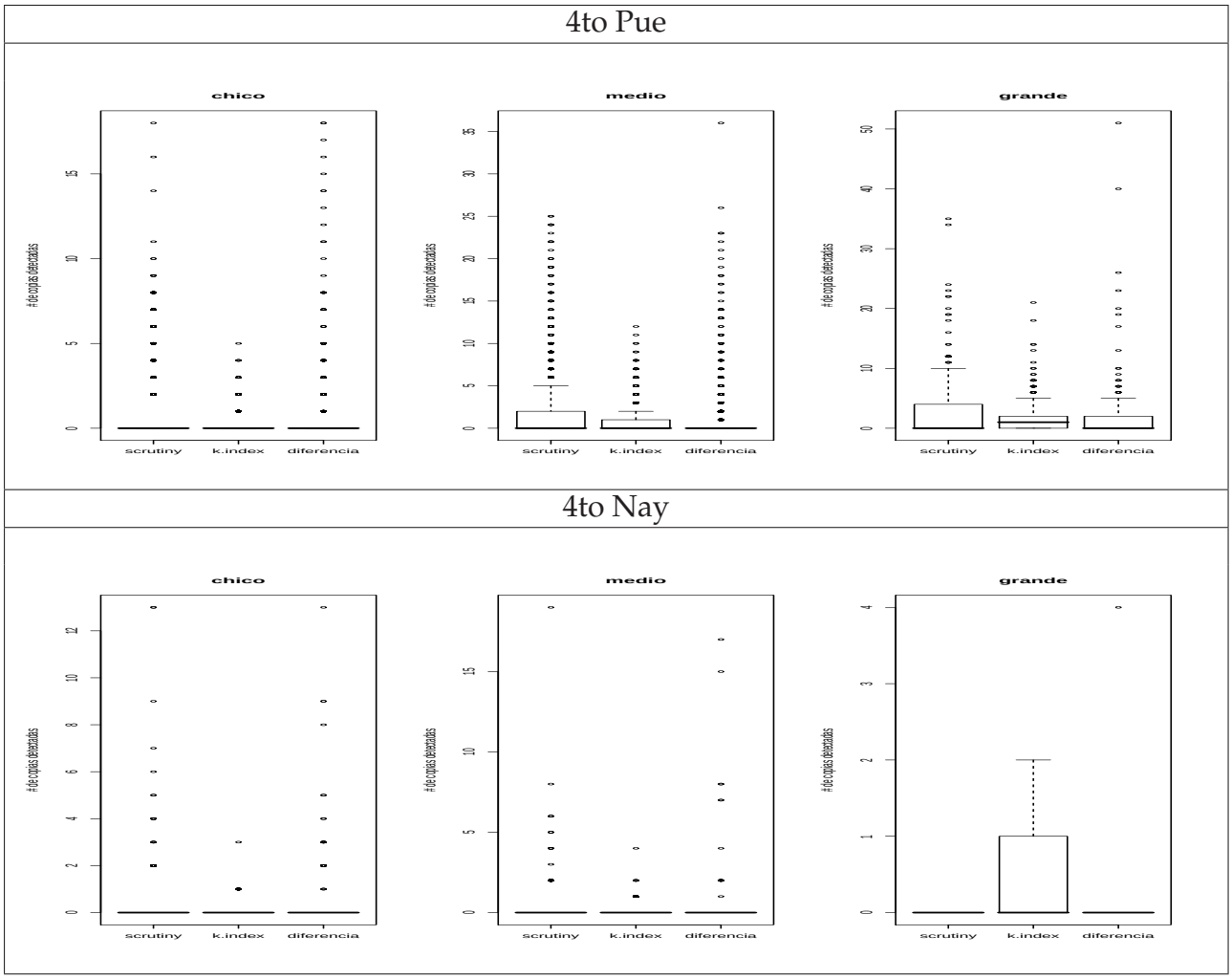
En datos reales uno de los problemas es la gran variabilidad de tamaños de grupos existentes. De hecho se tienen grupos con un solo alumno hasta grupos con más de 60. Por lo anterior, se reagruparon de la siguiente manera.

- Grupos chico son aquellos con menos de 20 alumnos.
- Grupos medianos son aquellos con [20, 40) alumnos.
- Grupos grandes son aquellos con más de 40 alumnos.

Los cuadros 2.9 y 2.10 muestran la densidad de alumnos detectados como sospechosos de copia por cada uno de los índices utilizados por la **SEP**, bajo los tres tamaños de grupo definidos como chico, mediano y grande.



Cuadro 2.9: Tamaño de grupo contra el índice de copia.



Cuadro 2.10: Tamaño de grupo contra el índice de copia.

Capítulo 3

Extensión del índice *scrutiny* a tríos

3.1. Descripción

En el capítulo 2 se abordó el tema de detección de copia basado en la comparación de respuestas de parejas de alumnos (c, s). Ignorando la posibilidad de copia entre más de dos alumnos.

En este capítulo se introduce un método desarrollado para detección de copia entre tres alumnos a la vez. Para este fin se extiende el índice de copia *scrutiny* a triadas de alumnos.

La base de esta extensión consiste en el cálculo de la probabilidad de que dos personas seleccionen la misma respuesta equivocada para el ítem j. Esté valor fue definido como

$$p_j = \frac{\sum\limits_{k\ opc.\ mal} p_{j,k}^2}{\left(1 - p_{j,opc.\ bien}\right)^2}$$

con $p_{j,k}$ representa la probabilidad de elegir la opción k para el ítem j.

Para el caso de triadas de alumnos se trabaja con

$$p_j = \frac{\sum\limits_{k\ opc.\ mal} p_{j,k}^3}{\left(1 - p_{j,opc.\ bien}\right)^3}$$

Como el cálculo de todas las triadas es computacionalmente caro se presenta también una variante más sencilla. El algoritmo 7 muestra como se calcula el valor *scrutiny* para un trío de sujetos.

Algorithm 7: Scrutiny para tríos

Para cada triada (s, c_1, c_2);

Input: Dado $p_j = \frac{\sum_{k\ opc.\ mal} p_{j,k}^3}{(1-p_{j,opc.\ bien})^3}$, Perr[j]= p_j, para cada ítem j

```
for i = 1 to M do
    I[i]= 0;
    W_cs[i]= 0;
    if u_si, u_c1i y u_c2i están mal contestadas then
        I[i]= 1;
        if u_si = u_c1i = u_c2i then
            W_cs[i]= 1;
for i = 1 to M do
    SumaI+= I[i];
    SumaW_cs+= W_cs[i];
    Pmedio+= Perr[i]* I[i];
Pmedio/ = SumaI;
```

StdDev= $\sqrt{SumaI * Pmedio * (1 - Pmedio)}$;

scrutiny= $\frac{\left(SumaW_{cs} - \frac{1}{2}\right) - SumaI * Pmedio}{StdDev}$;

3.2. Experimentos

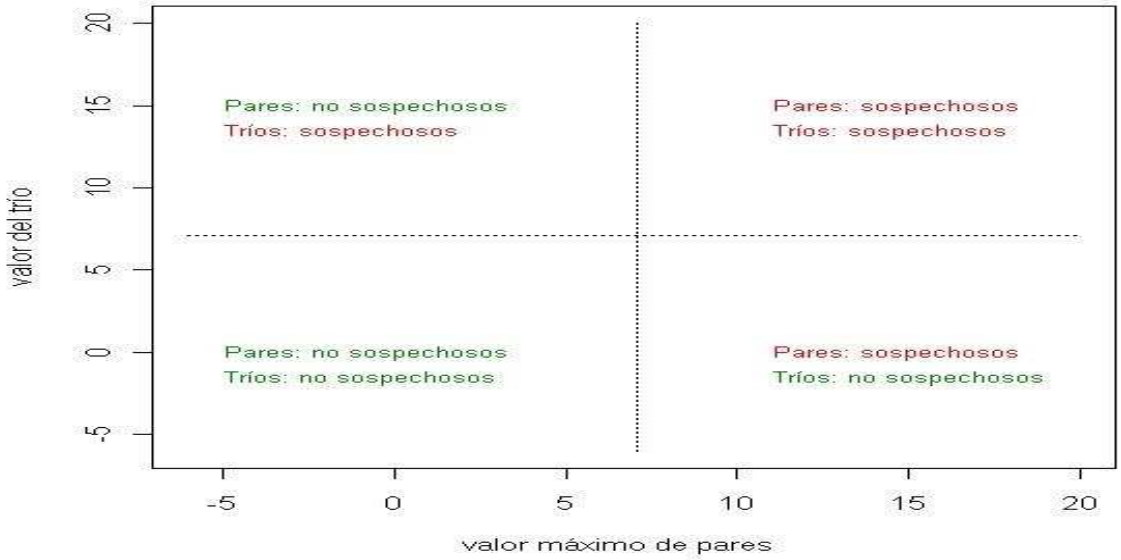
A continuación se presentan experimentos con base en la detección de triadas, utilizando datos simulados, datos ficticios y datos reales (para estos dos últimos los trenes de respuesta provienen del estado de Puebla de cuarto grado de primaria). La meta de esta sección es encontrar un *threshold* que facilite la discriminación entre los tríos que realizan copia y aquellos que no lo hacen.

Para mostrar los resultados con datos ficticios y datos reales, se utilizan graficas como la mostrada en el cuadro 3.1. La grafica se ha creado, con información sobre cada triada evaluada. Para cada trío se grafica un punto con coordenadas x el máximo valor *scrutiny* de los tres pares formados por un trío, y el valor *scrutiny* del trío. Dicha grafica esta dividida en cuatro secciones/cuadrantes (el valor de corte de las divisiones está basado en el α determinado por la **SEP** para el índice *scrutiny*, ver sub-sección 2.9.2), las cuales representan:

1. Sujetos que en pares no pasan el umbral y como triadas tampoco. (Pares: no sospechosos y Tríos: no sospechosos)
2. Sujetos que en pares no pasan el umbral y como triadas si. (Pares: no sospechosos y Tríos: sospechosos)
3. Sujetos que en pares superan el umbral y como triadas no. (Pares: sospechosos y Tríos: no sospechosos)

4. Sujetos que en pares superan el umbral y como triadas también. (Pares: sospechosos y Tríos: sospechosos)

En general, es de gran interés el estudio del cuadrante perteneciente a "Pares: no sospechosos y Tríos: sospechosos", ya que, es aquí donde los índices de detección de copia tradicionales no detectan a los alumnos sospechosos de copia.



Cuadro 3.1: Distribución de secciones de interés.

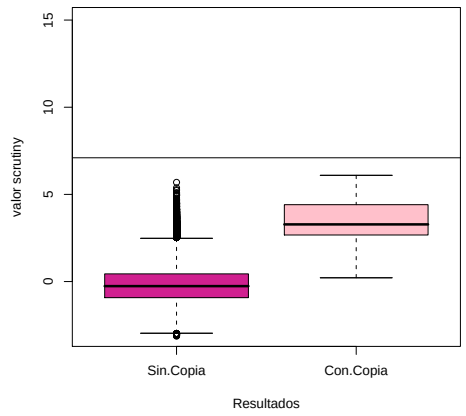
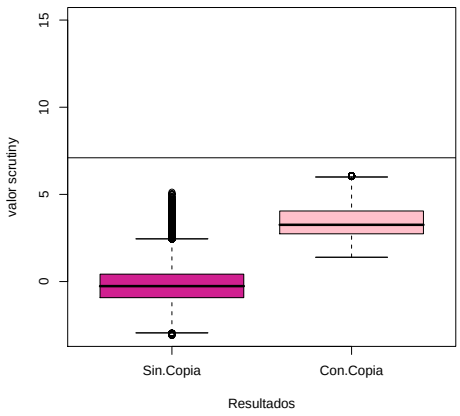
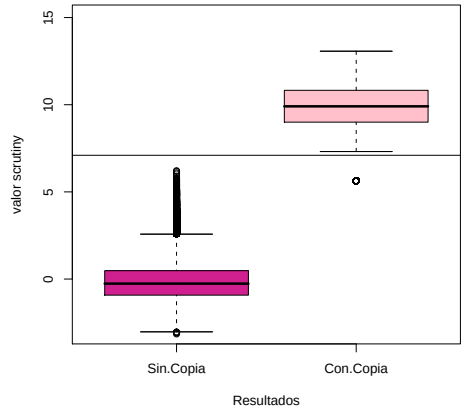
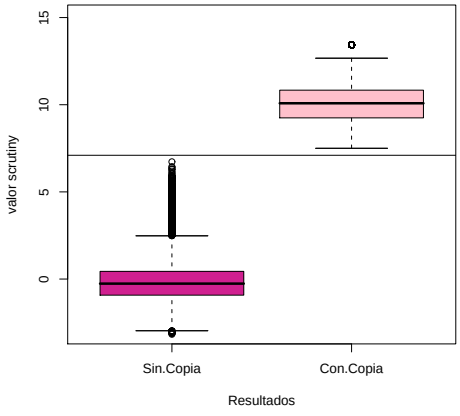
3.2.1. Datos simulados

El experimento consiste en generar un conjunto con 100 grupos, donde el tamaño de todos los grupos dentro de este conjunto, esta dado por N. De todas las triadas posibles por grupo {(1, 2, 3), (1, 2, 4), . . . , (1, 2, N), (2, 3, 4), (2, 3, 5), . . . , (2, 3, N), . . . , (N-2, N-1, N)}, sólo una de estas es la que realiza copia. El porcentaje de ítems que se copia esta dado por PC (= porcentaje de copia). En resumen, se tienen tres alumnos que participan en el proceso de copia por grupo.

Para la generación de cada grupo se ha utilizado una distribución uniforme, ver sección 1.1.1 para más detalles. Mientras que el proceso de copia se desarrolla con la octava opción de la sección 1.1.2.

El experimento tiene la finalidad de verificar si es posible discriminar/separar los tríos que han copiado de los que no. Lo que se espera obtener es, que las triadas que han copiado reporten un índice *scrutiny* mayor a las triadas constituidas con los sujetos que no han copiado.

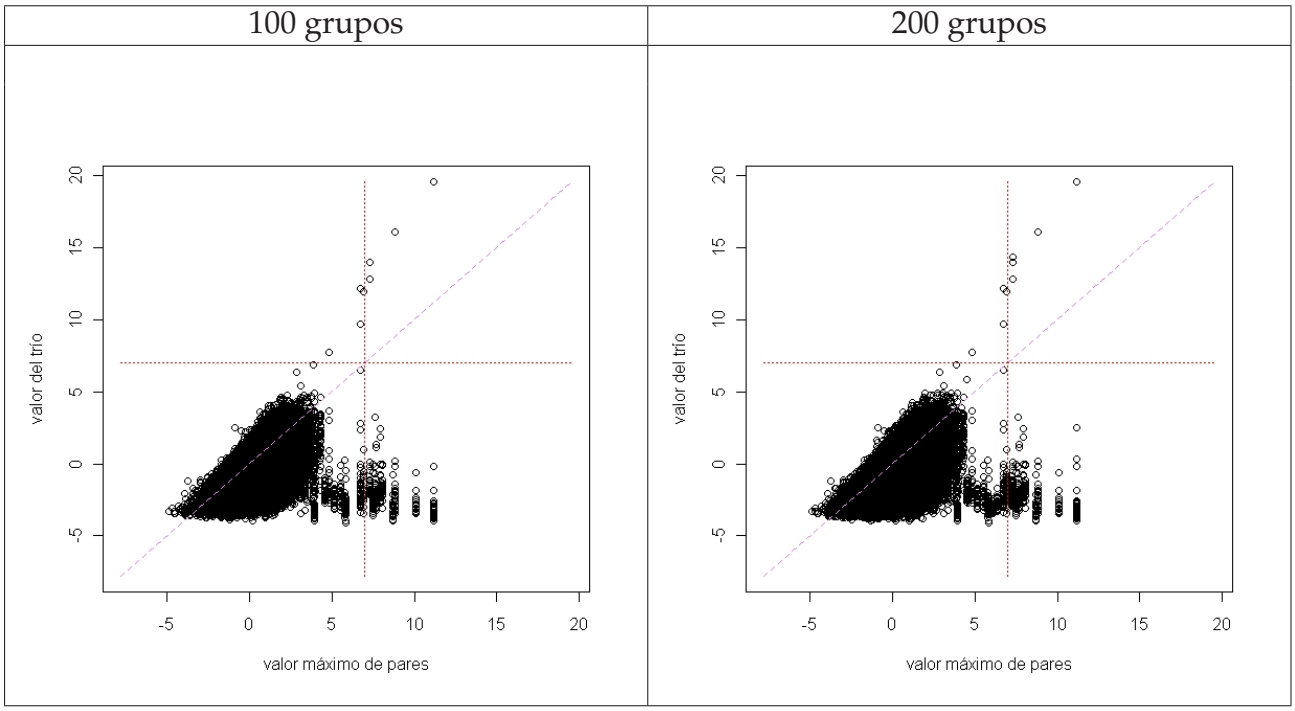
El cuadro 3.2 muestra un par de graficas donde en cada una de las graficas se muestran dos *boxplot*, uno con los valores *scrutiny* de aquellas triadas que no copiaron (Sin.Copia) y otro con las triadas que si copiaron (Con.Copia). Se puede rescatar que la sospecha de que es posible discriminar entre triadas con copia y triadas sin copia, no ha sido incorrecta. Salvo en grupos donde existe un PC= 10 %, donde los *boxplot* se llegan a traslapar un poco, sin embargo con PC= 30 % es clara la diferencia de los tríos que han copiado de los que no. Adicionalmente en cada grafica se ha agregado una línea que representa el umbral/*threshold* definido por la **SEP** para el índice *scrutiny*, sub-sección 2.9.2. Para valores de PC muy bajo ambos casos, tríos con copia y sin copia, no superan el umbral establecido. En cambio para PC un poco más grande, los tríos con copia están claramente por arriba del umbral.

PC	Número de alumnos	
	$N = 30$	$N = 40$
10 %		
30 %		

Cuadro 3.2: Datos simulados para 100 grupos, donde la longitud del examen es de 127 ítems.

3.2.2. Datos ficticios

La descripción de la generación de grupos ficticios se puede encontrar en la sub-sección 1.1.1. El cuadro 3.3 muestra los resultados obtenidos para 100 y 200 grupos ficticios, donde cada grupo cuenta con 25 alumnos. El supuesto que no copian entre ellos se puede valorar en dichas graficas, ya que el grueso de los resultados se encuentran en el cuadrante perteneciente a "Pares: no sospechosos y Tríos: no sospechosos", los puntos que se salen de esta colección de datos se pueden atribuir al azar o al mismo tipo de enseñanza en las escuelas. Esto es, que los alumnos contestaron lo que creen es correcto y no tanto que hayan copiado respuestas.



Cuadro 3.3: Grupos ficticios con $N = 25$ para cada grupo.

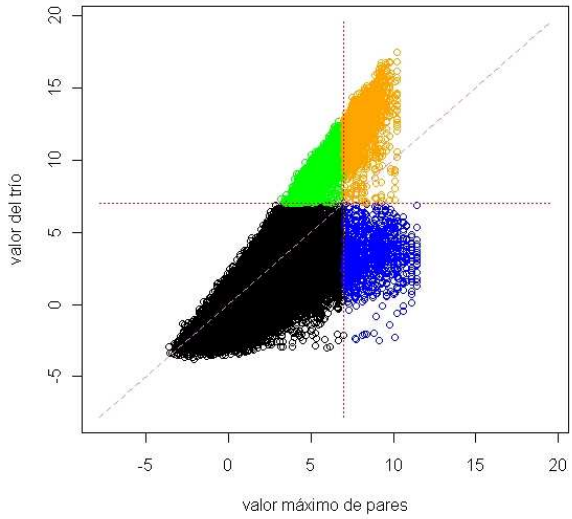
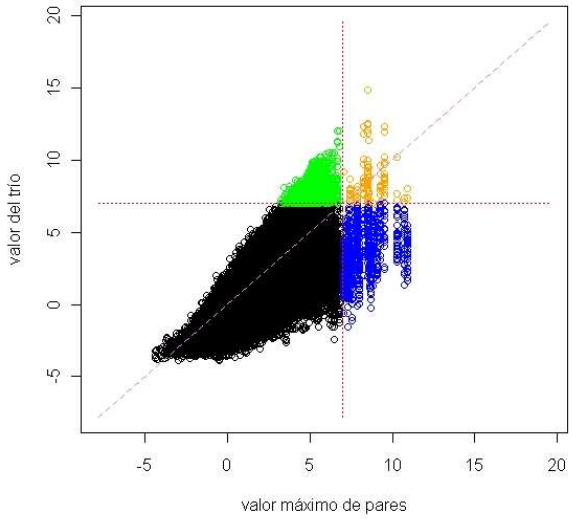
3.2.3. Datos reales

La descripción de la obtención de datos reales se puede encontrar en la sub-sección 1.1.1. Sólo se han utilizado grupos donde todos sus integrantes han contestado los M ítems de la prueba. Esta aclaración es porque existen grupos donde sus alumnos no contestaron, en ocasiones, hasta $\frac{2}{3}$ del examen. Con la intención de no entrar en conflicto con dichos alumnos es que se han descartado los grupos donde estos aparecen. Por otro lado en general no se toman en cuenta en el cálculo de alguno de los índices, esto es, si se cuenta con un grupo con 35 alumnos de los cuales 5 no contestaron una parte de los ítems, entonces dicho grupo se reduce a un total de 30 alumnos y es entre estos con los que se calcula el índice en cuestión.

Se crean dos conjuntos:

- **Conjunto A**.- Consta de 69 grupos con 25 alumnos cada uno,
- **Conjunto B**.- Consta de 50 grupos con 35 alumnos cada uno.

El cuadro 3.4 muestra los resultados obtenidos con cada conjunto.

Conjunto A	Conjunto B
	

Cuadro 3.4: Grupos reales.

Una vez más, el grueso de los resultados se encuentran en el cuadrante perteneciente a "Pares: no sospechosos y Tríos: no sospechosos". De hecho se desea que todos estuvieran en esté cuadrante, ya que indicaría que no hay evidencia para sospechar que ha existido copia.

En el conjunto A se detectaron 15 grupos dentro de la sección "Pares: no sospechosos y Tríos: sospechosos", el cuadro 3.5 muestra las tablas de confusión para cada uno de estos 15 grupos. Se observan dos grupos donde la versión de tríos detecto varios alumnos sospechosos de copia mientras que la versión de pares [1]no detecto ningún alumno. Para cada uno de estos dos grupos se realizaron todas las parejas posibles, mientras que, para cada pareja se contaron los ítems mal contestados e igual, del total de $\binom{25}{2}$ parejas posibles se tiene que en el primer grupo el mínimo de ítems mal e igual es de 12 y máximo 25. Mientras que en el segundo grupo el mínimo de ítems mal e igual es de 18 y máximo 39.

[1]Resultados tomados directamente de la prueba **ENLACE**. Número de alumnos detectados como sospechosos de copia bajo alguno de los tres índices de copia utilizados.

En el conjunto B se detectaron 12 grupos dentro de la sección "Pares: no sospechosos y Tríos: sospechosos", el cuadro 3.6 muestra las tablas de confusión para cada uno de estos 12 grupos.

		tríos			información	
		si	no	total	escuela	grupo
pares	si	18	0	18	21DPR0633V	14A
	no	6	1			
total		24				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	14	5	19	21DPR0913E	14A
	no	1	5			
total		15				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	4	4	21DPR0940B	14B
	no	3	18			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	9	3	12	21DPR1095U	14A
	no	4	9			
total		13				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	2	2	21DPR1644H	14A
	no	7	16			
total		7				

		tríos			información		mal e igual[1]
		si	no	total	escuela	grupo	
pares	si	0	0	0	21DPR1668R	14A	min= 9.6 %
	no	23	2				max= 20 %
total		23					

		tríos			información	
		si	no	total	escuela	grupo
pares	si	22	0	22	21DPR2122H	14B
	no	2	1			
total		24				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	0	0	21DPR2285S	24A
	no	3	22			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	0	0	21DPR2525A	14A
	no	3	22			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	5	2	7	21DPR2903L	14A
	no	5	18			
total		10				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	2	4	6	21EPR0002G	14A
	no	1	18			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	3	3	21EPR0307Z	24A
	no	3	19			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	12	0	12	21EPR0455H	14B
	no	5	8			
total		17				

		tríos			información		mal e igual
		si	no	total	escuela	grupo	
pares	si	0	0	0	21DPB0514G	14A	min= 14.4 %
	no	23	2				max= 31.2 %
total		23					

		tríos		total	información	
		si	no		escuela	grupo
pares	si	2	8	10	21DPR2401S	14B
	no	2	13			
total		4				

Cuadro 3.5: Grupos con 25 alumnos, grupos que están en el conjunto de "Pares: no sospechosos y Tríos: sospechosos"

		tríos		total	información	
		si	no		escuela	grupo
pares	si	0	0	0	21DPR0573X	14B
	no	3	32			
total		3				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	0	6	6	21DPR0959Z	14A
	no	3	26			
total		3				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	2	0	2	21DPR2380W	24A
	no	1	32			
total		3				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	0	4	4	21DPR2420G	14A
	no	4	27			
total		4				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	5	8	21DPR2818O	14A
	no	16	11			
total		19				

[1]Porcentaje de respuesta mal e igual, sobre todos los pares

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	2	5	21DPR3694C	24A
	no	6	24			
total		9				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	18	0	18	21EPR0325O	14C
	no	12	5			
total		30				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	0	3	21EPR0380H	14B
	no	2	30			
total		5				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	2	3	5	21EPR0399F	14C
	no	3	27			
total		5				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	5	0	5	21EPR0629H	24B
	no	22	8			
total		27				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	1	1	2	21DPB0472Y	14A
	no	7	26			
total		8				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	19	1	20	21DPR0291P	14B
	no	9	6			
total		28				

Cuadro 3.6: Grupos con 35 alumnos, grupos que están en el conjunto de "Pares: no sospechosos y Tríos: sospechosos"

Exploración

Por otro lado, de manera aleatoria se tomaron puntos pertenecientes a la colección "Pares: no sospechosos y Tríos: sospechosos" de grupos con 25 alumnos. Con la finalidad de verificar si los elementos, (s, c_1, c_2), pertenecientes a cada una de las triadas seleccionadas han sido marcados como sospechosos de copia por la **SEP**. Como se recordara dentro de los resultados reportados sobre la prueba **ENLACE** 2007, se indica si el alumno evaluado es sospechoso de copia, bajo tres distintos índices: *scrutiny*, *k-index* y diferencias.

Para el i-ésimo punto seleccionado al azar (perteneciente a la colección "Pares: no sospechosos y Tríos: sospechosos") se encontró que ninguno de sus elementos han sido detectados como sospechosos de copia por la **SEP**, sin embargo al calcular el valor *scrutiny* se obtuvo que el índice de copia de cada elemento, de la triada en cuestión, es por arriba de 6.0, con el umbral determinado por la **SEP** efectivamente dichos alumnos no son sospechosos de copia sin embargo, al comparar los trenes de repuesta se observa que existen varias coincidencias entre los tres sujetos, de hecho tienen $\frac{1}{4}$ del total de respuestas mal contestadas e igual.

3.3. Optimización

El proceso utilizado hasta ahora para la evaluación de copia a nivel de triadas, es por medio de una búsqueda en bruto, esto es, se evalúan todas las triadas posibles lo cual involucra un gran costo de tiempo de computo, sobre todo cuando se esta trabajando con varios grupos y adicionalmente dichos grupos tienen muchos alumnos cada uno.

El número de pares posibles dentro de un grupo con N alumnos esta dado por $\binom{N}{2}$, mientras que el número de triadas posibles esta dado por $\binom{N}{3}$, el cuadro 3.7 muestra el número de evaluaciones necesarias para distintos valores de N, donde se puede apreciar que entre más alumnos por grupo, es necesario un mayor número de evaluaciones.

[1]Porcentaje de respuesta mal e igual, sobre todos los pares

N	# de pares	# de tríos	total de evaluaciones
3	3	1	4
4	6	4	10
5	10	10	20
6	15	20	35
10	45	120	165
20	190	1,140	1,330
30	435	4,060	4,495
40	780	9,880	10,660
50	1,225	19,600	20,825
60	1,770	34,220	35,990

Cuadro 3.7: Número de evaluaciones necesarias.

Es por esta razón que se ha ideado una estrategia para no realizar las evaluaciones de todas las triadas posibles.

Como se recordará, de las graficas del cuadro 3.4, es de gran interés detectar aquellos puntos pertenecientes a la colección "Pares: no sospechosos y Tríos: sospechosos". Como primer paso se evalúan todos los pares posibles y se conservan sólo aquellos que se encuentran dentro del intervalo [3.5, 7.1], dicho intervalo ha sido tomado de las graficas antes mencionadas. Se trabaja sólo con los pares que caen dentro del citado intervalo.

El algoritmo 8 muestra el método desarrollado para no realizar las evaluaciones de todas las posibles triadas.

Algorithm 8: Optimización

forall *Grupo con N alumnos* **do**
- Calcular el índice *scrutiny* para todas las parejas posibles;
- Se crea la lista de todas las posibles triadas;
- Se evalúan sólo las triadas donde al menos dos de las parejas formadas por el trío, {par(i, j), par(i, k), par(j, k)}, están en el intervalo $3.5 \leq par(a, b) < 7.1$;

3.3.1. Datos reales

Después de aplicar el algoritmo 8 al conjunto A se detectaron 16 grupos que reportaron la presencia de algún trío, donde al menos dos de las parejas formadas por los elementos de la triada están en el intervalo [3.5, 7.1). El cuadro 3.8 muestra las tablas de confusión para cada uno de estos 16 grupos.

Así mismo, se aplico el algoritmo 8 al conjunto B y se detectaron 11 grupos que reportaron la presencia de algún trío. El cuadro 3.9 muestra las tablas de confusión para cada uno de estos 11 grupos.

Para el conjunto A ambos cuadros 3.5 y 3.8 tienen resultados muy parecidos. En el caso del conjunto B los cuadros 3.6 y 3.9, también muestran un comportamiento muy semejante. Sin embargo en este caso no fue necesaria la evaluación de todas las triadas. Por lo tanto es factible el optar por esta opción que es más económica computacionalmente y que da resultados igual de buenos que la evaluación de todas las triadas posibles.

		tríos			información	
		si	no	total	escuela	grupo
pares	si	18	0	18	21DPR0633V	14A
	no	6	1			
total		24				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	16	3	19	21DPR0913E	14A
	no	1	4			
total		17				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	4	4	21DPR0940B	14B
	no	3	18			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	12	0	12	21DPR1095U	14A
	no	6	0			
total		18				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	2	2	21DPR1644H	14A
	no	7	16			
total		7				

		tríos			información		mal e igual
		si	no	total	escuela	grupo	
pares	si	0	0	0	21DPR1668R	14A	min= 9.6 %
	no	19	6				max= 20 %
total		19					

		tríos			información	
		si	no	total	escuela	grupo
pares	si	22	0	22	21DPR2122H	14B
	no	2	1			
total		24				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	0	0	21DPR2285S	24A
	no	3	22			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	2	8	10	21DPR2401S	14B
	no	2	13			
total		4				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	2	3	5	21DPR2481U	14B
	no	1	19			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	0	0	0	21DPR2525A	14A
	no	3	22			
total		3				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	5	2	7	21DPR2903L	14A
	no	0	18			
total		5				

		tríos			información	
		si	no	total	escuela	grupo
pares	si	3	3	6	21EPR0002G	14A
	no	1	18			
total		4				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	0	3	21EPR0307Z	24A
	no	4	18			
total		7				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	12	0	12	21EPR0455H	14B
	no	4	9			
total		16				

		tríos		total	información		mal e igual
		si	no		escuela	grupo	
pares	si	0	0	0	21DPB0514G	14A	min= 14.4 %
	no	22	3				max= 31.2 %
total		22					

Cuadro 3.8: Grupos con 25 alumnos, grupos que están en el conjunto de "Pares: no sospechosos y Tríos: sospechosos" [optimización]

		tríos		total	información	
		si	no		escuela	grupo
pares	si	0	0	0	21DPR0573X	14B
	no	3	32			
total		3				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	2	0	2	21DPR2380W	24A
	no	1	32			
total		3				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	2	2	4	21DPR2420G	14A
	no	5	26			
total		7				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	5	8	21DPR2818O	14A
	no	14	12			
total		17				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	2	6	8	21DPR2921A	14A
	no	1	26			
total		3				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	1	4	5	21DPR3694C	24A
	no	5	25			
total		6				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	18	0	18	21EPR0325O	14C
	no	11	6			
total		29				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	0	3	21EPR0380H	14B
	no	2	30			
total		5				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	3	2	5	21EPR0399F	14C
	no	3	27			
total		6				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	5	0	5	21EPR0629H	24B
	no	21	9			
total		26				

		tríos		total	información	
		si	no		escuela	grupo
pares	si	19	1	20	21DPR0291P	14B
	no	6	9			
total		25				

Cuadro 3.9: Grupos con 35 alumnos, grupos que están en el conjunto de "Pares: no sospechosos y Tríos: sospechosos" [optimización]

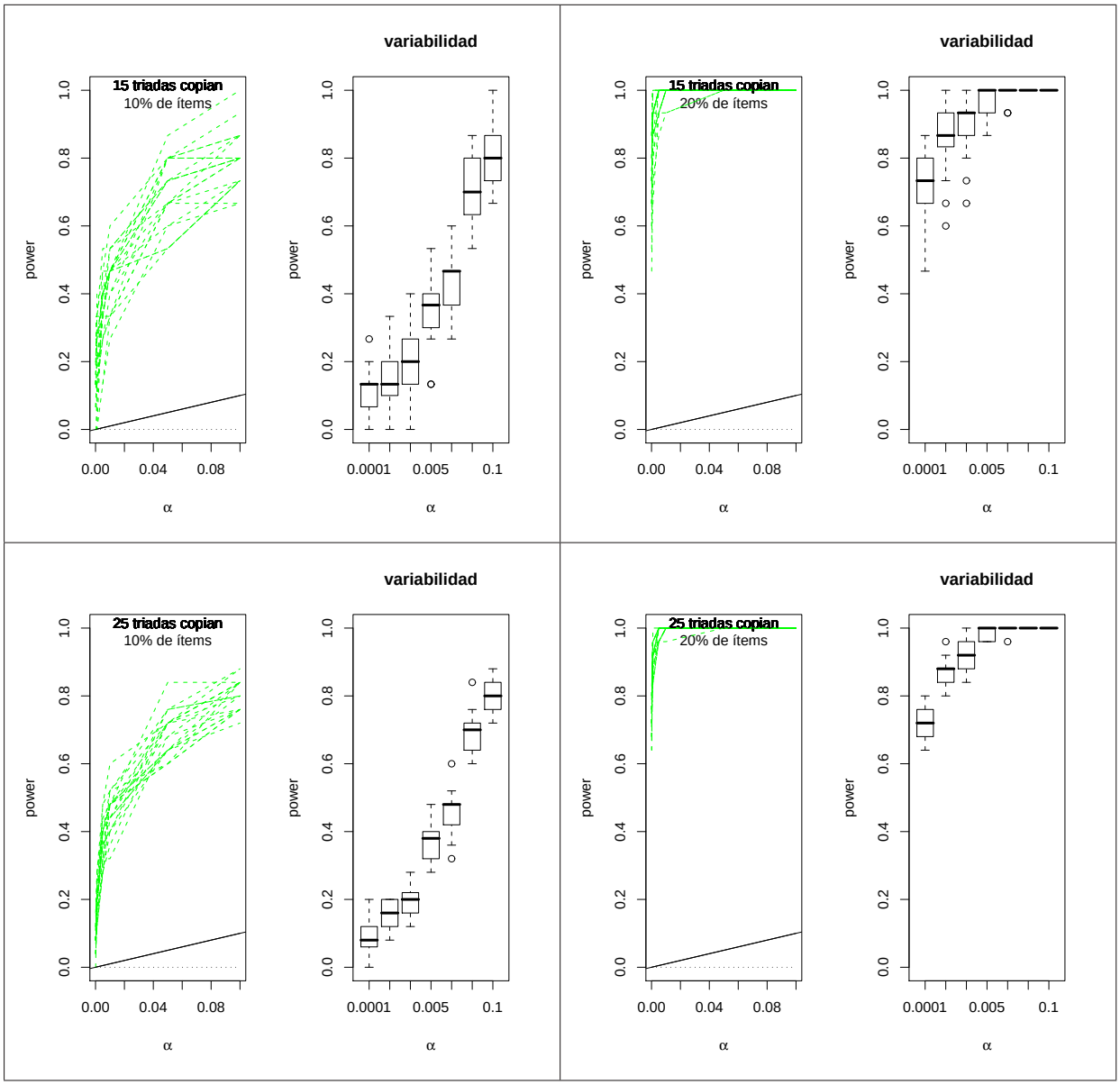
3.4. Evaluación de poder

Para esta sección se simulan trenes de respuesta para 20 grupos. Cada grupo con 150 alumnos y 80 ítems, bajo distintos parámetros de copia. Las respuestas al examen se han simulado bajo la opción 1 de la sección 1.1.1 y la copia se simula siguiendo la idea de la opción 7 de la sección 1.1.2.

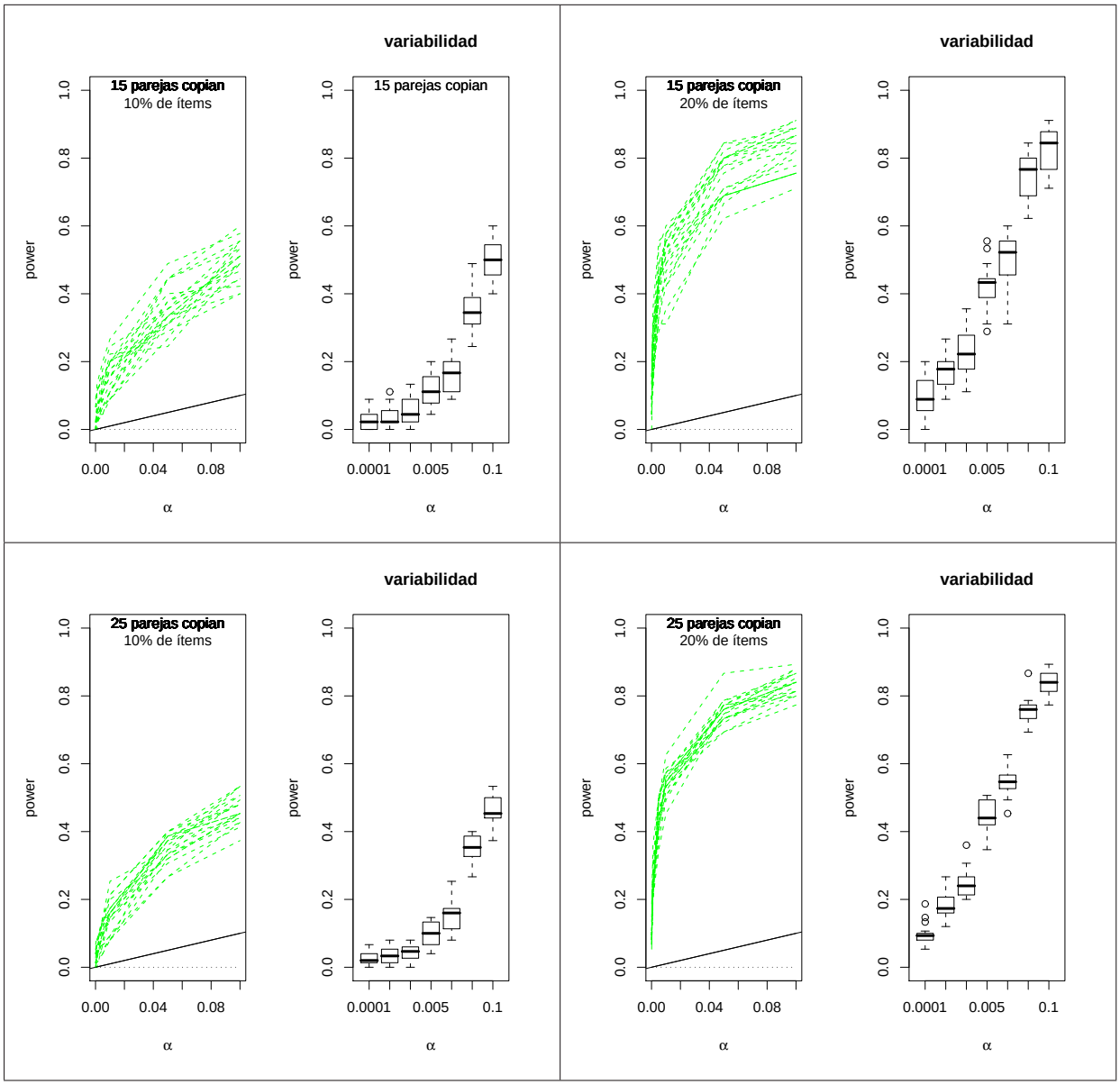
El cuadro 3.10 muestra las graficas del poder de *scrutiny* con tríos, para 10 % y 30 % de ítems copiados, y 30 % y 50 % de tríos que copian. Mientras que el cuadro 3.11 muestra las graficas del poder de *scrutiny* con pares, para 10 % y 30 % de ítems copiados, y 30 % y 50 % de parejas que copian. En ambos casos el error tipo II se calculó según el primer camino mostrado en la sub-sección 2.9.1.

Para la evaluación del poder a nivel de pares, se toman los elementos del trío $\{s, c_1, c_2\}$ con los cuales se forman las parejas $\{(s, c_1), (s, c_2), (c_1, c_2)\}$. Lo anterior es con la finalidad de comparar el poder a nivel de pares contra el poder a nivel de tríos, haciendo uso de los mismos datos.

De las graficas se puede observar que el poder de *scrutiny* crece conforme el número de ítems copiados aumenta. También se alcanza a preciar que, acorde el número de tríos que copian crece, la variabilidad en los valores del poder disminuye. Esto último ocurre tanto en las graficas del poder de *scrutiny* en tríos como en las pertenecientes al poder de *scrutiny* en parejas.



Cuadro 3.10: Graficas del poder de *scrutiny* con tríos, en datos simulados.



Cuadro 3.11: Graficas del poder de *scrutiny* con pares, en datos simulados.

Capítulo 4

Evaluación por grupos

4.1. Introducción

Como se expuso en capítulos anteriores, los índices para detección de copia se basan en parejas de ciertos alumnos, a lo que se bautizó como *búsqueda a nivel micro de copia*. Sin embargo queda por explorar el comportamiento global del grupo en el proceso de copia en un cierto examen, cuando por ejemplo entre muchos sujetos pertenecientes a un cierto grupo se intercambian respuestas. En este capítulo se desea explorar el comportamiento grupal ante la copia, lo que se señalará como *búsqueda a nivel macro de copia*.

Para lograr esto, se consideran los trenes de respuesta del grupo como un archivo de texto. Se buscan patrones que se repiten. Para aprovechar la duplicidad de cadenas/secuencias de caracteres se hace uso de algoritmos de compresión de datos/textos. Si después de aplicar el algoritmo de compresión el tamaño del texto se reduce considerablemente entonces, se puede inferir que existen varias secuencias duplicadas. Para la *búsqueda a nivel macro de copia* se sigue la misma línea de todos los índices de detección de copia estudiados: se trabaja sólo con las respuestas mal contestadas.

Con esta idea a seguir se presentan dos variantes, una llamada "compresión básica" donde se muestra un pequeño modelo de compresión basado en secuencias, mientras que un segundo modelo llamado "compresión usando LZW" muestra un modelo basado en compresión de datos.

4.2. Compresión básica

Por compresión de datos se entiende cualquier algoritmo que reciba un conjunto de datos de entrada y que sea capaz de generar datos de salida cuya representación ocupa menos espacio de almacenamiento.

Como primer acertamiento al problema, se desarrolla un método de compresión basado en las secuencias de caracteres. Por ejemplo, el cuadro 4.1 muestra una serie de trenes de respuestas donde se resaltan las secuencias repetidas. Cabe señalar que para

que una cadena sea tomada en cuenta debe de hacer referencia a los mismos ítems que aquella a la cual se parece, en otras palabras, no se admiten secuencias desfasadas.

		Preguntas								
		1	2	3	4	5	6	7	8	9
Alumnos	1	A	B	C	D	D	A	D	C	B
	2	A	B	C	D	A	A	C	B	D
	3	A	B	C	D	C	B	B	D	A
	4	A	A	A	A	B	B	A	C	C
	5	D	D	C	C	B	D	B	D	A

Cuadro 4.1: Ejemplo de un archivo de texto.

Como se puede apreciar, existe información repetida en el archivo, por lo tanto al eliminar las secuencias extras, dejando sólo una como representante, se tendrá un archivo de texto mucho más chico. Siguiendo esta idea, se cuenta el número de secuencias repetidas así como también la longitud de la misma. Por otro lado para tomar en cuenta una secuencia de caracteres, la longitud mínima de está debe ser de almenos 5 % de M.

Sea V el número de secuencias diferentes presentes en el archivo de texto, D el número de veces que aparece cada secuencia y L la longitud de cada secuencia (número de caracteres que la conforman). El valor de compresión básica esta dada por

$$k = \frac{\sum_{i=1}^{V} (D_i - 1) * L_i}{N * M} \tag{4.1}$$

El valor de compresión, k, toma valores entre $[0, 1]$. Si $k \to 0$ se entiende que el archivo se puede comprimir muy poco, por otro lado, si $k \to 1$ se entiende que el archivo se puede comprimir bastante.

Siguiendo el ejemplo del cuadro 4.1, se tiene que $V = 2$,

$$k = \frac{(3-1) * 4 + (2-1) * 3}{5 * 9}$$

de donde se observa que $k \approx 0.244$, por lo tanto se deduce que el archivo se puede comprimir un poco.

Dado que se trabaja con respuestas erróneas, se tienen dos alternativas:

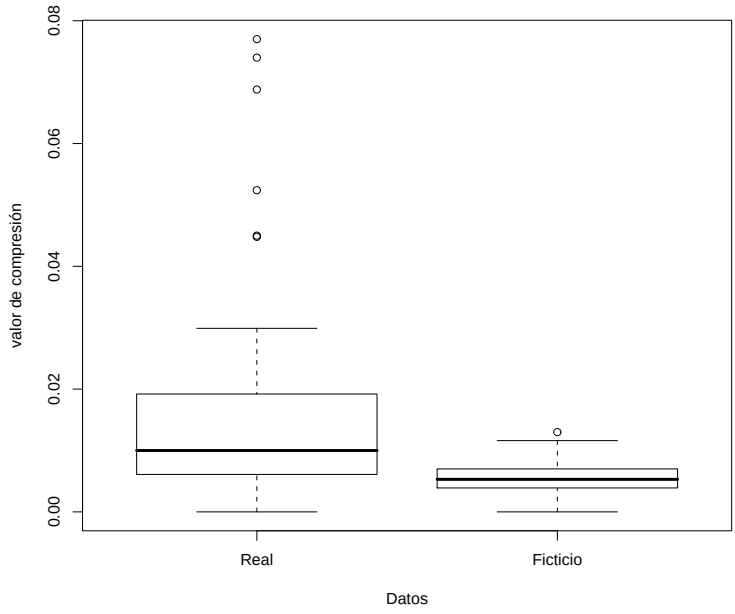
- tipo 1: la secuencia pertenece a ítems mal contestados estrictamente consecutivos,
- tipo 2: la secuencia pertenece a ítems bien y mal contestados.

La diferencia entre los dos tipos radica en el valor del parámetro L, de la ecuación (4.1). Bajo el tipo 1, L es la longitud de la secuencia, si la secuencia tiene una longitud de 10 caracteres entonces, los 10 caracteres deben referirse a ítems mal contestados. Bajo el tipo 2, L es el número de ítems incorrectos de la secuencia, esto es, si la secuencia tiene una longitud de 10 caracteres pero sólo 4 refieren a ítems mal contestados entonces, $L = 4$.

4.2.1. Experimentos

Con la finalidad de corroborar la eficiencia del método, se desarrollaron un par de experimentos, con grupos ficticios y grupos reales, ver sub-sección 1.1.1. Los experimentos son sobre 50 grupos, donde $N = 40$ y $M = 127$.

Tipo 1.- Para grupos ficticios, se espera que no exista copia y por lo tanto los valores de k muy pequenõs. Mientras que para grupos reales, se espera que exista un grado de copia mayor. El cuadro 4.2 muestra los resultados para la compresión tipo 1, donde se puede ver que efectivamente los valores de compresión para los datos ficticios son mucho más chicos que los valores de compresión para los datos reales.

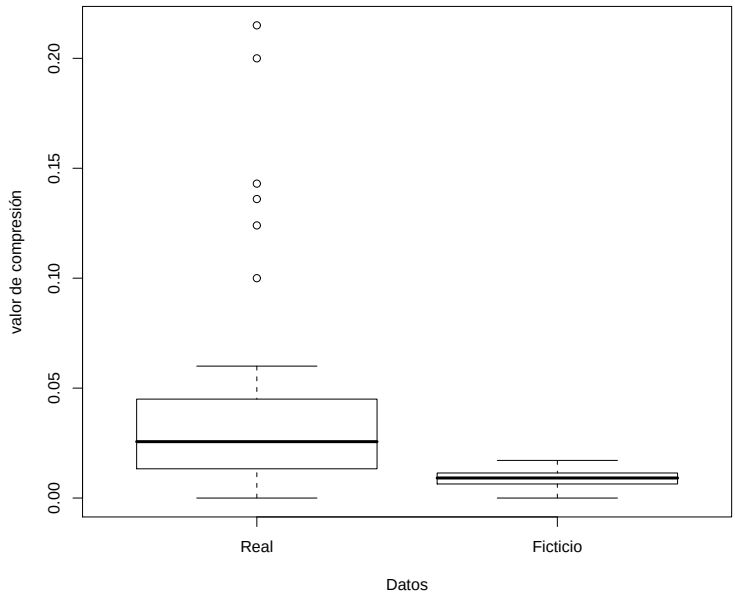


Cuadro 4.2: Compresión básica tipo 1.

Tipo 2.- El cuadro 4.3 muestra los resultados para la compresión tipo 2, donde al igual que en compresión tipo 1, los valores de compresión para los datos ficticios son mucho más chicos que los valores de compresión para los datos reales.

Como se puede observar el máximo valor de compresión tipo 1 es ≈ 0.08 mientras que el valor máximo para compresión tipo 2 es ≈ 0.2. La gran diferencia radica en que la compresión tipo 1 es mucho más estricta que la tipo 2.

En general este método da una idea general de cómo están compuestos los datos. Sin embargo existen métodos más especializados para la compresión de datos, los cuales se describen a continuación.



Cuadro 4.3: Compresión básica tipo 2.

4.3. Técnicas de compresión usando diccionarios

Un enfoque muy razonable para la codificación de archivos de datos es mantener una lista o diccionario [8] de los patrones que ocurren con frecuencia. Cuando estos patrones aparecen en el archivo de salida, se codifican haciendo uso de una referencia al diccionario. Para que esta técnica sea eficaz, el tamaño del diccionario, debe ser mucho menor que el número de todos los posibles patrones.

Los métodos de compresión sin pérdida de información (*lossless*) se caracterizan en que la tasa de compresión que proporcionan está limitada por la entropía de la señal original, que es una medida de compresión, lo que se traduce en la mejor compresión asintótica posible por cualquier algoritmo. Entre estas técnicas destacan las que emplean métodos estadísticos basados en la teoría de Shannon, que permiten la compresión sin pérdida. Por ejemplo: codificación de Huffman, codificación aritmética y Lempel Ziv. Son métodos idóneos para la compresión dura de archivos.

Existen dos tipos de diccionarios, los cuales se describen a continuación.

4.3.1. Diccionario estático

Si se tiene información a priori de las palabras que van a ocurrir con mayor frecuencia entonces, el uso de un diccionario estático es el más apropiado. En esté tipo de diccionarios se almacenan sólo los patrones más frecuentes.

Durante la compresión; si ocurre un patrón que está en el diccionario entonces, se entrega la palabra código almacenada previamente en el diccionario. Si por el contrario el patrón no se encuentra en el diccionario, entonces se codifica separadamente cada símbolo del patrón.

4.3.2. Diccionario adaptativo

La mayoría de las técnicas basadas en diccionarios adaptativos tienen sus raíces en artículos publicados por Jacob Ziv y Abraham Lempel en 1977 y 1978 [8]. Proporcionan dos enfoques diferentes a la construcción de diccionarios y cada enfoque ha dado lugar a una serie de variaciones. Los enfoques basados en el artículo del '77 se dice que pertenecen a la familia **LZ77**, mientras que los enfoques basados en el artículo del '78 se dice que pertenecen a la familia **LZ78**.

La técnica LZ77

Se basa en el uso de una ventana deslizante sobre el archivo de texto a comprimir. Moviendo una ventana de búsqueda sobre los datos y codificando tripletas con elementos como la longitud de coincidencias, desplazamiento y el siguiente símbolo de la cadena. Si un símbolo no aparece se codifica con longitud 0 y desplazamiento 0, seguido del símbolo no codificado.

Si en el texto aparecen varios caracteres poco frecuentes entonces, el uso de codificación por tripletas del tipo $< o, l, c >$, donde o es el desplazamiento, l es la longitud y c el símbolo a codificar, es ineficiente. Una simple modificación como el agregar una bandera, que indique si lo que sigue es el código de sólo un símbolo. Al utilizar está bandera también se deshace la necesidad del tercer elemento de la tripleta.

La técnica LZ78

Mientras que en **LZ77** se asume que los patrones iguales ocurren en estrecha colaboración (dentro del alcance de una ventana). En **LZ78**, se eliminó la suposición de recurrencia cercana, sustituyendo el uso de una ventana deslizante con la construyendo explícitamente de un diccionario de palabras. La modificación más conocida, que inicialmente provocó la mayor parte del interés de los algoritmos **LZ**, es una modificación introducida por Terry Welch conocida como **LZW**.

Durante el proceso de codificación se va construyendo una tabla o diccionario con la información que se va manejando. El codificador sólo envía el índice del diccionario; para ello, el diccionario ha de ser pre-cargado con todas las letras del alfabeto del archivo.

La entrada para el codificador se acumula en un patrón p que es almacenado en el diccionario. Si la adición de otra letra tiene como resultado el patrón $p*a$ (donde $*$ denota

la concatenación) que no está en el diccionario, entonces el índice de p se transmite al receptor, el patrón $p*a$ es agregado al diccionario y se inicia un nuevo patrón con la letra a.

Con **LZW** es posible crear sobre la marcha, de manera automática y en una única pasada un diccionario de cadenas de caracteres que se encuentren dentro del texto, que se desea comprimir, mientras al mismo tiempo se codifica. Dicho diccionario no se transmite con el texto comprimido, dado que el descompresor puede reconstruirlo usando la misma lógica que usa el compresor.

Ejemplo del algoritmo LZW - codificación

A continuación se muestra uno de los ejemplos más comunes sobre **LZW** tomado de [8]. Se tiene la secuencia de entrada: wabba/wabba/wabba/wabba/woo/woo/woo.

Se asume que el alfabeto esta dado por {/, a, b, o, w}, el diccionario inicial de **LZW** se puede observar en el cuadro 4.4.

Índice	Entrada
1	/
2	a
3	b
4	o
5	w

Cuadro 4.4: Diccionario inicial.

El codificador primero encuentra la letra w. Este "patrón" ya está en el diccionario por lo tanto al concatenarlo con la siguiente letra se forma el patrón wa. Este patrón no existe en el diccionario al añadirlo forma el sexto elemento del diccionario el siguiente patrón se comienza con la última letra obtenida a. Como la letra a existe en el diccionario, al concatenarla con el próximo elemento b se obtiene el nuevo patrón ab. Este patrón no está en el diccionario se añade el patrón ab al diccionario como el séptimo elemento del mismo se empieza a construir un nuevo patrón principiando con la letra b. Siguiendo con la construcción de patrones de dos letras hasta llegar a la letra w en el segundo $wabba$. Hasta este punto la salida del codificador se compone enteramente de los índices a partir del diccionario inicial: {5, 2, 3, 3, 2, 1}. El diccionario hasta este punto aparece que el cuadro 4.5.

El siguiente símbolo en la secuencia es a, concatenando con w, tenemos el patrón wa. Este patrón ya existe en el diccionario, índice 6, de modo que hay que leer el siguiente símbolo, el cual es b. Concatenando con wa, se tiene el patrón wab. Este patrón no existe en el diccionario, por lo que se incluye como la 12va entrada del diccionario y se empieza un nuevo patrón con el símbolo b. Observe que después de una serie de entradas de dos letras, ahora se tiene una entrada de tres letras. Al avanzar la codificación, la longitud de las entradas del diccionario sigue aumentando. El diccionario final del

Índice	Entrada
1	/
2	a
3	b
4	o
5	w
6	wa
7	ab
8	bb
9	ba
10	a/
11	/w
12	w...

Cuadro 4.5: Construcción de la 12va entrada.

Índice	Entrada	Índice	Entrada
1	/	14	a/w
2	a	15	wabb
3	b	16	ba/
4	o	17	/wa
5	w	18	abb
6	wa	19	ba/w
7	ab	20	wo
8	bb	21	oo
9	ba	22	o/
10	a/	23	/wo
11	/w	24	oo/
12	w...	25	/woo
13	bba		

Cuadro 4.6: Diccionario final.

proceso de codificación se muestra en el cuadro 4.6. Se observa que las entradas desde la 12va hasta la 19va son todas de tres o cuatro letras.

La secuencia de salida codificada esta dada por {5, 2, 3, 3, 2, 1, 6, 8, 10, 12, 9, 11, 7, 16, 5, 4, 4, 11, 21, 23, 4}.

4.3.3. Algoritmos LZW: compresión

El algoritmo **LWZ** puede iniciarse con 256 caracteres simples predeterminados, sin embargo la implementación más común de **LWZ** es que el diccionario inicie vació y se valla creando poco a poco.

Bajo esta idea el funcionamiento del compresor se muestra en el algoritmo 9. El método crea una cadena, llamase *s* y un diccionario *D*, ambos inicialmente vacíos.

Algorithm 9: Compresión **LZW**

```
Input: s= vació, D= vació
repeat
    leer un caracter c;
    if sc ya está en D then
        s= sc;
    else
        enviar a la salida el código de s;
        agregar sc a D;
        s= c;
until fin de archivo ;
```

4.4. Compresión usando LZW

Como un segundo acercamiento al problema de *búsqueda a nivel macro de copia*, se aplica la compresión sin perdida usando **LZW**. Siguiendo la idea básica de ver los trenes de respuesta de un grupo con un archivo de texto, al cual se le aplica una compresión de textos, en este caso **LZW**.

Antes de iniciar con el proceso de compresión es necesario realizar los pasos:

- Limitarse a datos mal contestados. Para utilizar este método como un índice de copia para todo el grupo, es preciso que las respuestas que se desean comprimir sean sólo respuestas mal contestadas, de acuerdo a la *búsqueda a nivel micro de copia*.

  En el cuadro 4.7 se muestran once trenes de respuestas, la longitud del examen es de veinte preguntas. Se descartan las respuestas bien contestadas, esto es, si para el *i*-ésimo ítem todos los alumnos contestaron bien, entonces dicho ítem no es tomado en cuenta en el proceso de compresión. Suponiendo que el tren de

respuestas correctas es: {A, A, C, B, B, D, C, C, B, A, C, A, D, A, C, C, C, C, C, D}, al eliminar las respuestas bien contestadas, del cuadro 4.7, se obtiene el cuadro 4.8.

- Codificación de trenes de respuestas. Hay que tomar en cuenta que no es lo mismo observar como respuesta una *D* en el primer ítem a observar como respuesta una *D* en el segundo ítem. Ya que los ítems uno y dos representan preguntas distintas, no es valido comprimir ambas respuestas como si fueran iguales. Por lo tanto, cada respuesta se redefine según el número del ítem al que se esta haciendo referencia. Haciendo uso del código **ASCII** y tomando en cuenta que son cuatro opciones de respuesta por ítem, se ha utilizado $opc * M + j$, para M= número de ítems, $j = \{1, ..., M\}$ y opc= $\{0, 1, 2, 3\}$ que representa las posibles opciones de respuesta $\{A, B, C, D\}$.

  Al redefinir los valores del cuadro 4.8 se obtiene como resultado el cuadro 4.9. Dentro del mismo cuadro, para los ítems 10 y 11, donde en ambos casos la respuesta más recurrente es la opción *D*, si se comprimieran dichas columnas sin redefinir las respuestas, entonces se fundirían en una sola respuesta, sin embargo al aplicar la redefinición se logra una diferencia entre ellas. Con esto al aplicar compresión **LWZ** no se confundirían las respuestas.

- Casos especiales los valores 7 y 8. Existen preguntas que no han sido contestadas (esto se ve en datos reales), en el examen se representan con un 7 u 8 según sea la razón por la cual no se contesto (7 representa que no contesto, tal vez, por que no sabía la respuesta, mientras que 8 representa que no se presento al examen, por lo tanto no contesto). Para la redefinición se toma el valor **ASCII** 253 para un 7 y el valor 254 para un 8.

- Agregar una última columna con respuestas permutadas; de tal forma que no se presenten sub-cadenas similares entre el fin de un tren de respuesta y principio del siguiente.

4.4.1. Experimentos

Como parte de los ejercicios realizados se probó con dos grupos de datos: datos simulados y datos reales, estos últimos obtenidos de los resultados de la prueba **ENLACE**.

Datos simulados

Las figuras mostradas en los cuadros 4.10, 4.11 y 4.12 contienen dos partes. En la primera columna se exponen dos histogramas; el primer histograma muestra el resultado de la compresión de grupos donde se ha copiado, el segundo histograma muestra el resultado de la compresión de grupos donde no se ha realizado copia. En la segunda columna se muestran los boxplots correspondientes.

alum	ítems																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	B	A	C	D	B	A	C	A	D	A	C	A	D	A	C	C	B	C	B	A
2	B	A	C	D	A	D	C	A	D	D	D	A	D	A	C	C	B	C	B	A
3	B	A	C	D	A	D	C	A	B	D	D	A	D	A	C	C	B	C	B	A
4	B	A	C	D	B	D	C	A	B	D	D	A	D	A	C	C	C	C	B	A
5	B	A	C	D	A	D	C	D	B	D	D	A	D	A	C	C	C	C	B	A
6	B	A	C	D	A	D	C	D	B	D	D	A	D	A	C	C	C	C	B	A
7	B	A	C	D	B	D	D	C	B	A	C	A	D	A	B	C	A	C	B	A
8	B	A	C	D	B	A	D	A	B	D	A	C	D	A	C	B	A	D	B	A
9	B	A	C	D	B	A	C	A	D	A	C	A	D	A	C	C	B	C	B	A
10	B	A	C	D	A	B	C	A	D	B	C	A	C	D	B	C	A	C	B	A
11	B	A	C	D	A	D	C	D	B	D	D	A	D	A	C	C	C	C	B	A

Cuadro 4.7: Trenes de respuestas de ejemplo.

alum	ítems																	
	1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	B	D	-	A	-	A	D	-	-	-	-	-	-	-	B	-	B	A
2	B	D	A	-	-	A	D	D	D	-	-	-	-	-	B	-	B	A
3	B	D	A	-	-	A	-	D	D	-	-	-	-	-	B	-	B	A
4	B	D	-	-	-	A	-	D	D	-	-	-	-	-	-	-	B	A
5	B	D	A	-	-	D	-	D	D	-	-	-	-	-	-	-	B	A
6	B	D	A	-	-	D	-	D	D	-	-	-	-	-	-	-	B	A
7	B	D	-	-	D	-	-	-	-	-	-	-	B	-	A	-	B	A
8	B	D	-	A	D	A	-	D	A	C	-	-	-	B	A	D	B	A
9	B	D	-	A	-	A	D	-	-	-	-	-	-	-	B	-	B	A
10	B	D	A	B	-	A	D	B	-	-	C	D	B	-	A	-	B	A
11	B	D	A	-	-	D	-	D	D	-	-	-	-	-	-	-	B	A

Cuadro 4.8: Limitandose sólo a las respuestas mal contestadas.

alum	ítems																	
	1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	21	64	-	6	-	8	69	-	-	-	-	-	-	-	37	-	39	20
2	21	64	5	-	-	8	69	70	71	-	-	-	-	-	37	-	39	20
3	21	64	5	-	-	8	-	70	71	-	-	-	-	-	37	-	39	20
4	21	64	-	-	-	8	-	70	71	-	-	-	-	-	-	-	39	20
5	21	64	5	-	-	68	-	70	71	-	-	-	-	-	-	-	39	20
6	21	64	5	-	-	68	-	70	71	-	-	-	-	-	-	-	39	20
7	21	64	-	-	67	-	-	-	-	-	-	-	35	-	17	-	39	20
8	21	64	-	6	67	8	-	70	11	52	-	-	-	36	17	78	39	20
9	21	64	-	6	-	8	69	-	-	-	-	-	-	-	37	-	39	20
10	21	64	5	26	-	8	69	30	-	-	53	74	35	-	17	-	39	20
11	21	64	5	-	-	68	-	70	71	-	-	-	-	-	-	-	39	20

Cuadro 4.9: Resultado después de aplicar redefinición.

Para la generación de datos se utilizaron las dos técnicas descritas en la sub-sección 1.1.1. El proceso de copia es a nivel grupal por medio de *cluster*, ver opción 6 de la sub-sección 1.1.2. A continuación se describen un par de experimentos usando estos procesos de generación de datos.

a) **Siguiendo una distribución uniforme.-** El cuadro 4.10 muestra los resultados para mil repeticiones, las preguntas a copiar son seleccionadas según las respuestas mal contestadas por cada alumno dentro del *cluster*.

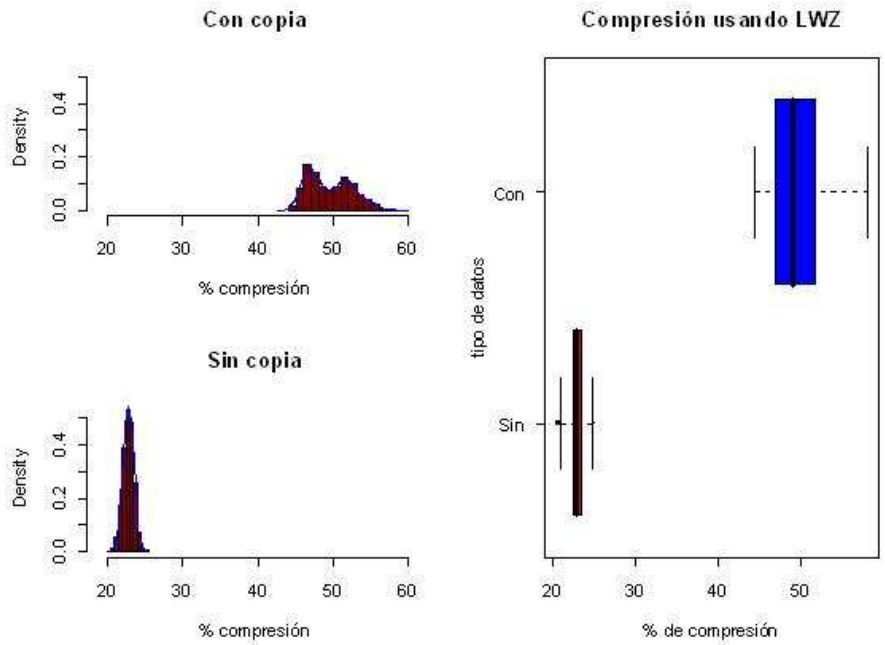
   La clara separación entre histogramas, representantes de grupos, con copia y sin copia; es una muestra a favor de la idea que se pueden discriminar los grupos que han copiado de los que no.

b) **Siguiendo una distribución IRT.-** Los cuadros 4.11 y 4.12 muestran los resultados para mil repeticiones. Los cuadros muestran resultados para grupos simulados con los mismos parámetros, excepto el porcentaje de ítems copiados el cual es distinto para cada experimento, PC= {50, 60, 70, 80, 90, 100} %.

   Los resultados muestran que entre mayor sea PC, los histogramas correspondientes a grupos con y sin copia, se separan cada vez más. Otro punto a resaltar es el hecho que, para PC muy grandes como 90 y 100 los histogramas de grupos con copia no cambian mucho.

Datos ficticios (evaluación por días)

La descripción de la generación de grupos ficticios se puede encontrar en la subsección 1.1.1. Para un mejor análisis se han divido los resultados del examen por días, esto es, para la aplicación de la prueba **ENLACE** se eligen dos días a nivel nacional para aplicar el examen; en el primer día se aplica la primera mitad del examen, mientras que



Cuadro 4.10: Ejemplo con 40 alumnos, 50 ítems, 2 clusters y 50 % de ítems copiados por los alumno dentro de los *clusters*. La generación de respuestas fue con una distribución uniforme.

la segunda mitad del mismo se aplica en el transcurso del segundo día.

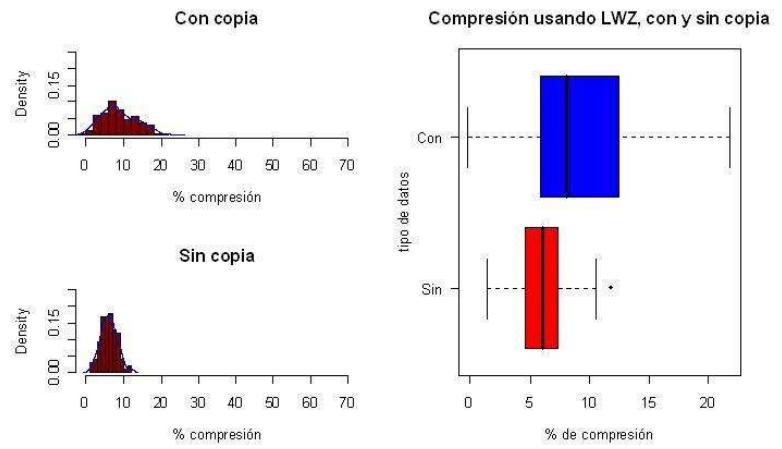
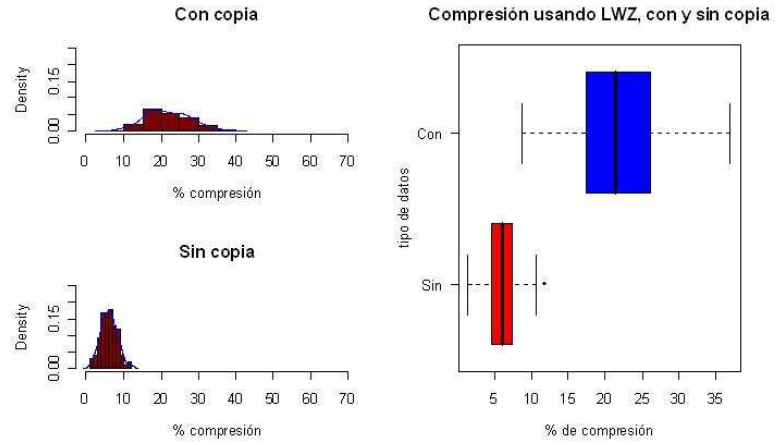
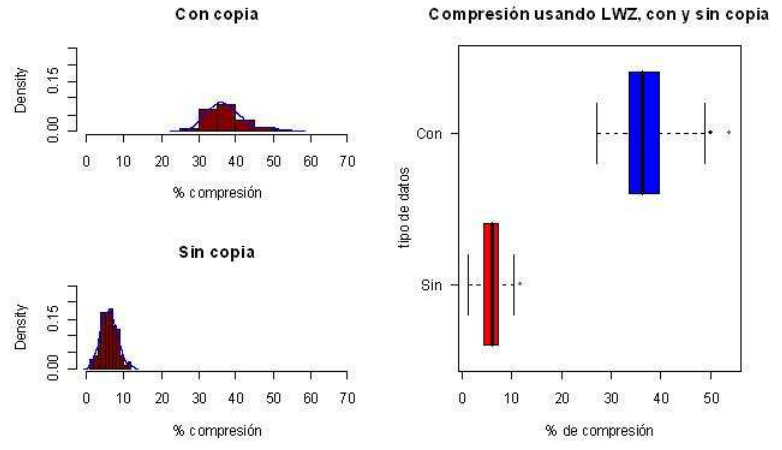
El cuadro 4.13 muestra resultados de 100 grupos ficticios. En ambos días, 1 y 2, el porcentaje de compresión es relativamente pequeño y sin mucha variabilidad, lo cual se esperaba ya que se sabe a priori que no existe copia entre los alumnos de cada grupo. Aun así, en el segundo día se alcanza un mayor porcentaje de compresión que en el primer día.

Datos reales (evaluación por días)

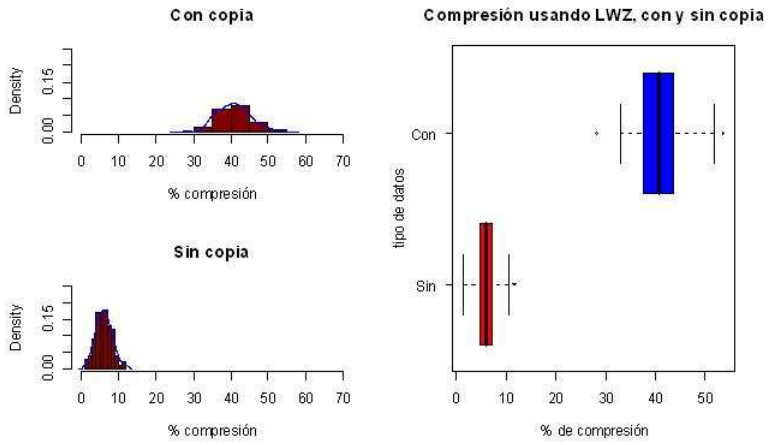
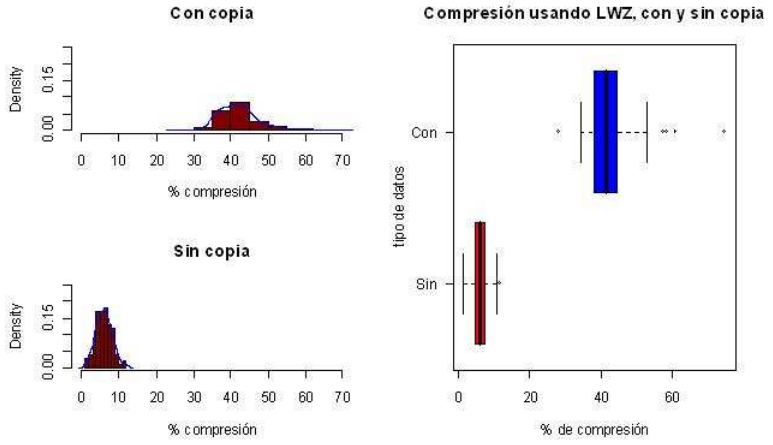
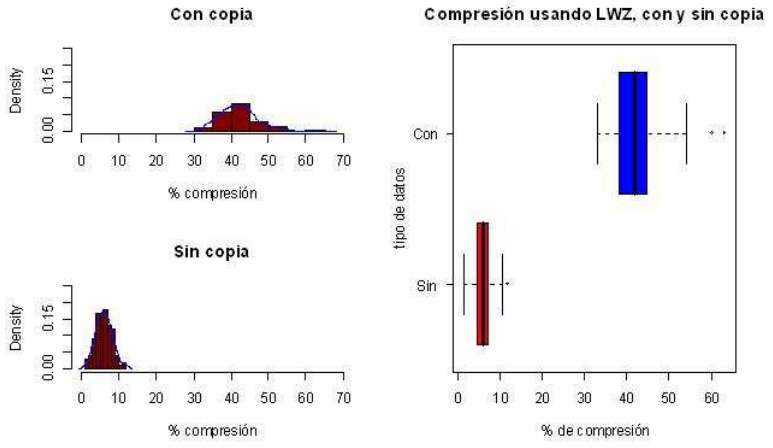
Al igual que en la sección anterior, los trenes de respuesta a exámenes provenientes de datos reales, ver sub-sección 1.1.1, han sido divididos por días. El cuadro 4.14 muestra los resultados obtenidos para grupos con 25, 35 y 45 alumnos del estado de Puebla, donde todos los integrantes de cada grupo contestaron completamente sus respectivos exámenes.

A continuación se enlistan un par de observaciones sobre dicho cuadro, en general las observaciones son sobre *outliers* detectados en los resultados.

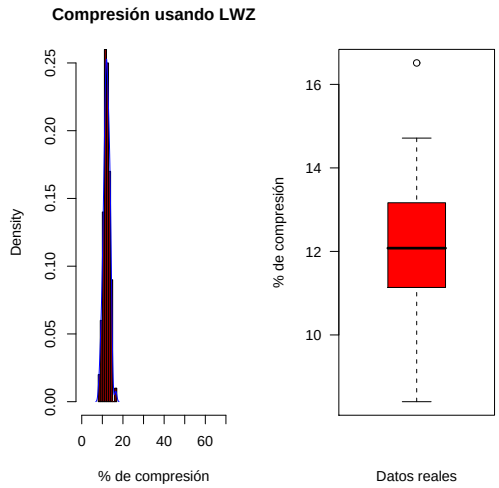
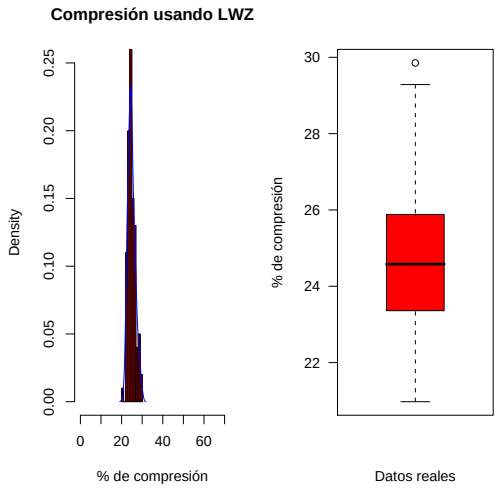
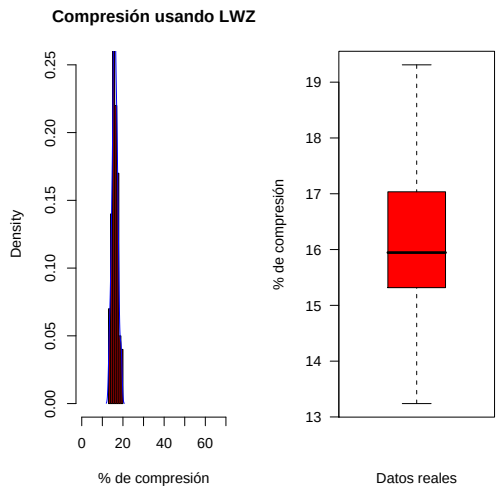
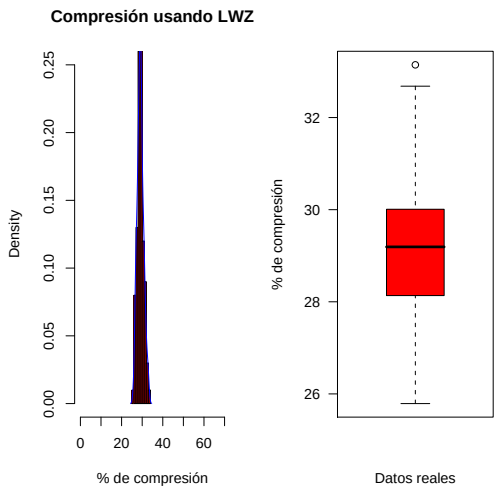
- En grupos con 35 alumnos, se observa que existe un par de *outliers*, uno por día, que vale la pena investigar ya que surge la duda ¿pertenecen al mismo grupo? la respuesta es afirmativa. Al investigar dichos *outliers* se descubre que pertenecen

PC	resultado
50 %	
60 %	
70 %	

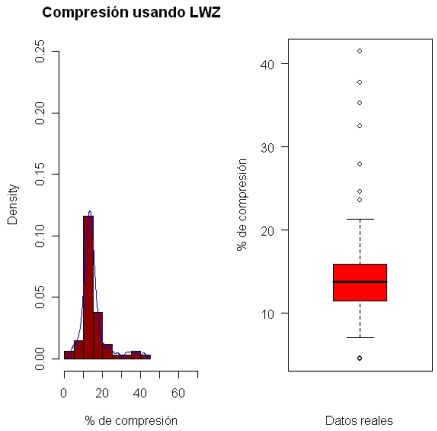
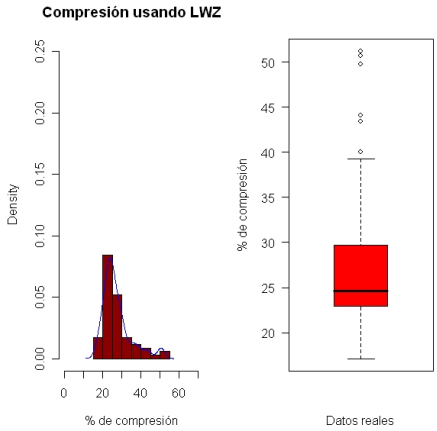
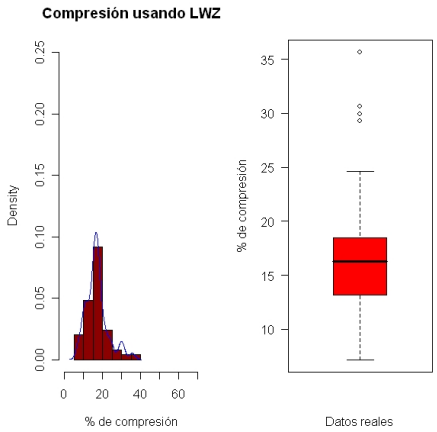
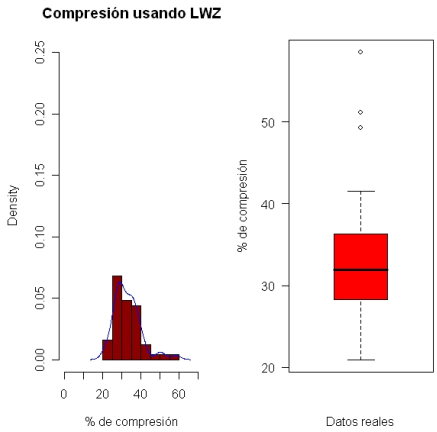
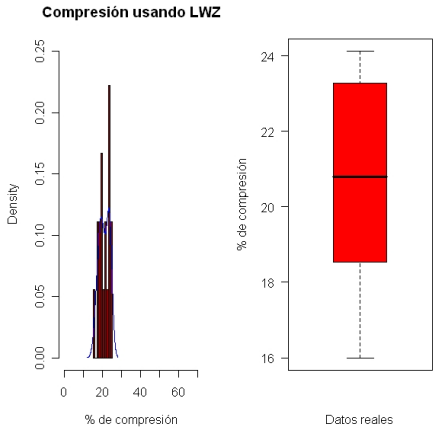
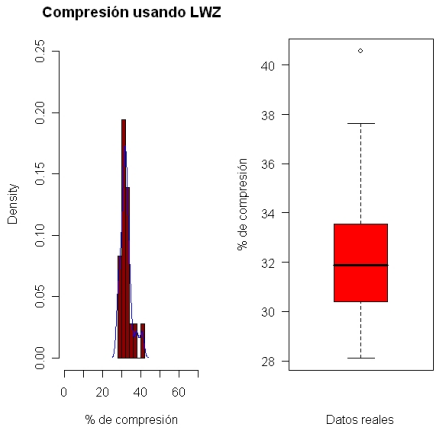
Cuadro 4.11: Ejemplo con 40 alumnos, 50 ítems, 2 clusters y 50 %, 60 % y 70 % de ítems copiados por los alumno dentro de los *clusters*. La generación de respuestas fue con una distribución **IRT**.

PC	resultado
80 %	
90 %	
100 %	

Cuadro 4.12: Ejemplo con 40 alumnos, 50 ítems, 2 clusters y 80 %, 90 % y 100 % de ítems copiados por los alumno dentro de los *clusters*. La generación de respuestas fue con una distribución **IRT**.

N	día 1	día 2
30		
40		

Cuadro 4.13: Compresión de datos ficticios por días.

N	día 1	día 2
25		
35		
45		

Cuadro 4.14: Compresión de datos reales por días.

al mismo grupo de alumnos, dado que, para ambos días el porcentaje de compresión es alto (el *outlier* del día 1 da una compresión de ≈ 35.62 % y el *outlier* del día 2 da una compresión de ≈ 58.48 %) vale la pena indagar más en el grupo en cuestión.

Al investigar en el conjunto de datos de esté grupo en particular, sobre todo el examen (las 127 preguntas), se encontró que existen parejas que tienen entre 67 y 125 respuestas iguales. Este ejemplo en particular esta a favor de la idea de utilizar la metodología de compresión de datos para explorar a nivel grupal si existe copia.

- En los *outliers* de los resultados pertenecientes a aquellos grupos con 25 alumnos, se tienen dos en particular, un primer *outlier* que muestra para el día 1 una compresión de ≈ 41.46 % y para el día 2 una compresión de ≈ 50.69 % y un segundo *outlier* que muestra para el día 1 una compresión de ≈ 35.24 % y para el día 2 una compresión ≈ 51.17 %. Al explorar en los trenes de respuesta de ambos grupos se encontró que efectivamente son todas las respuestas muy parecidas y mal contestadas. Al mismo tiempo se revisaron los resultados de la **SEP** sobre estos grupos, se encontró que para el primer grupo sólo siete estudiantes no fueron detectados como sospechosos de copia y en el segundo grupo se encontró que sólo dos alumnos no fueron detectados como sospechosos de copia por ninguno de los tres índices de copia utilizados en la prueba **ENLACE** 2007.

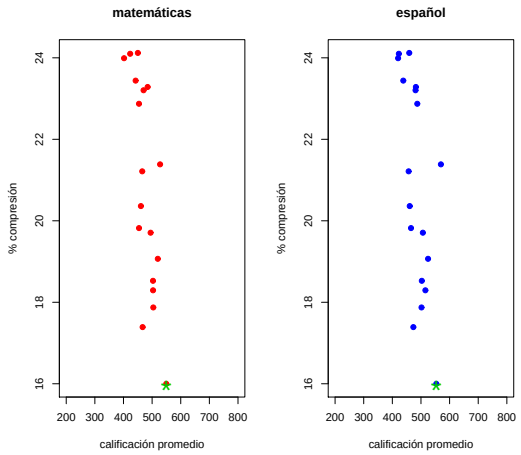
- Otro grupo investigado fue aquel *outlier*, de los resultados pertenecientes a grupos con 25 alumnos, que dio el menor valor de compresión en el día 1, una compresión de ≈ 4.53 %, mientras que para el día 2 dio una compresión de ≈ 40.06 % al indagar en los trenes de respuesta de dicho grupo se observa que son muy parecidos los trenes de respuesta sin embargo en la primera parte (la correspondiente al día 1) las respuestas en su mayoría son correctas y en la segunda parte en su mayoría incorrectas. Al revisar los resultados de la **SEP** sobre esté grupo, se encontró que ningún alumno ha sido catalogado como sospechoso de copia, esto es porque en la **SEP** se trabaja con todo el tren de respuesta y no lo dividen por días. Este ejemplo es un punto a favor de la idea de explorar los trenes de respuesta por días y no como un todo.

- Al indagar el *outlier* perteneciente a grupos con 45 alumnos, se tiene en particular uno que surge en el día 2 con una compresión de ≈ 40.55 %. Con la finalidad de indagar en los datos se realizó el cuadro 4.15, donde se resalta con una estrella el grupo de interés. En dicho cuadro se muestran tres graficas: a) primer día de aplicación, b) segundo día de aplicación y c) ambos días.

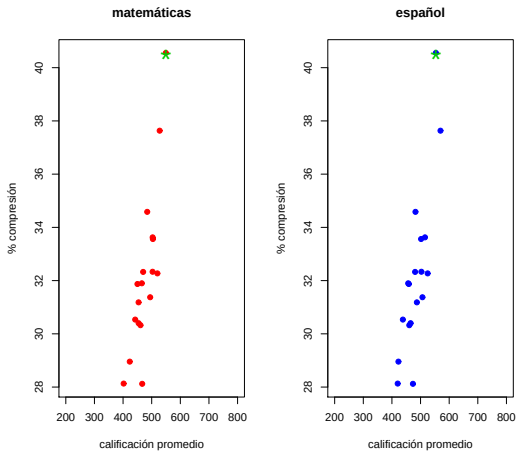
  La grafica (a) muestra que el grupo de interés es el que tiene un menor grado de compresión ≈ 15.99 %, no así en la grafica (b) donde el grupo de interés reporta el valor máximo de compresión ≈ 40.55 %. Sin embargo al realizar la compresión de ambos días, como un todo, la compresión es de ≈ 35.67 %. Como se puede observar, al comprimir todo el examen se pierde información importante sobre

los resultados. Ya que aparentemente es durante el segundo día cuando los niños copian. Complementariamente, al revisar los resultados de la **SEP** se tiene que ningún alumno, de este grupo en particular, fue detectado como sospechoso de copia.

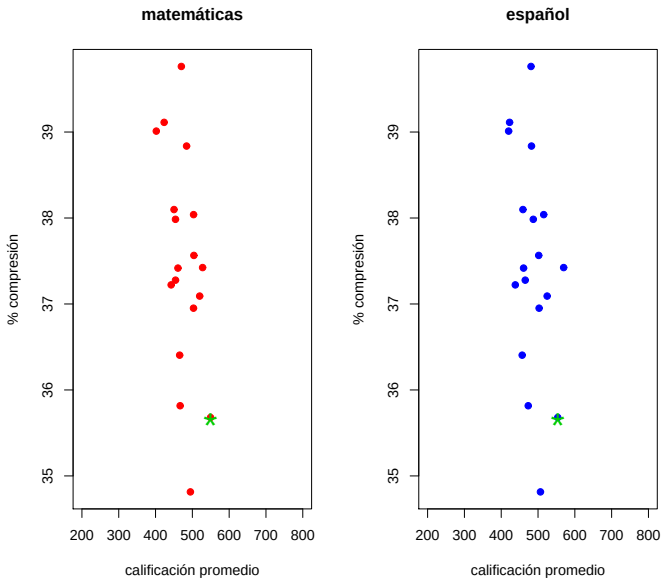
a) Grafica perteneciente al día 1



b) Grafica perteneciente al día 2



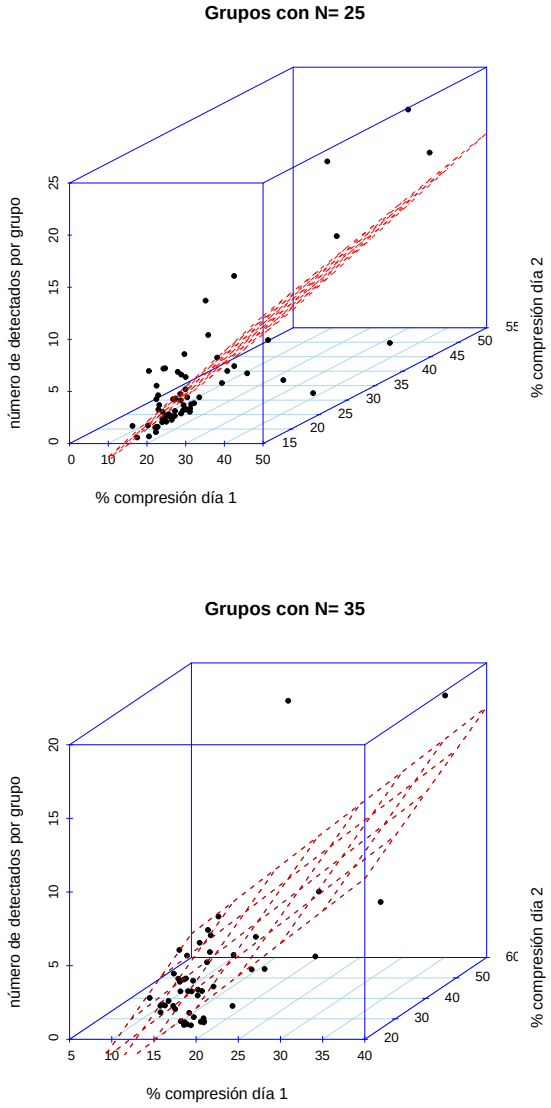
c) Grafica perteneciente a ambos días



Cuadro 4.15: Compresión de datos reales por días para grupos con 45 alumnos.

Las graficas mostradas en el cuadro 4.16, muestran el porcentaje de compresión por días y el número de alumnos detectados como sospechosos de copia por alguna de las técnicas utilizadas por la **SEP**. Algo que vale la pena resaltar es el hecho de que en aquellos grupos donde se detectaron varios alumnos sospechosos de copia, se alcanzo niveles de compresión grande en alguno de los dos días ó en ambos. Cabe resaltar

el hecho de que existen muchos grupos donde no se detecto copia y sin embargo el porcentaje de compresión muestra un comportamiento sospechoso en los trenes de respuesta. Anexamente los cuadros 4.17 y 4.18 enlistan los identificadores de las escuelas y grupos evaluados. Para grupos con $N = 25$ y $N = 35$ respectivamente.



Cuadro 4.16: Número de alumnos detectados por la **SEP** y el porcentaje de compresión por días.

	escuela	grupo		escuela	grupo
1	21DPR0464Q	24B	36	21DPR2401S	14B
2	21DPR0511K	14A	37	21DPR2481U	14B
3	21DPR0608W	14A	38	21DPR2525A	14A
4	21DPR0611J	14B	39	21DPR2535H	14A
5	21DPR0622P	14A	40	21DPR2820C	14A
6	21DAI0013D	14B	41	21DPR2903L	14A
7	21DPR0633V	14A	42	21DPR3457A	14A
8	21DPR0697F	14A	43	21DPR3535O	14A
9	21DPR0711I	14B	44	21DPR3535O	14B
10	21DPR0770Y	14A	45	21EPR0002G	14A
11	21DPR0772W	24C	46	21EPR0012N	14B
12	21DPR0782C	14C	47	21EPR0045E	14A
13	21DPR0809T	14A	48	21EPR0101G	14A
14	21DPR0834S	14A	49	21EPR0106B	14B
15	21DPR0851I	14A	50	21EPR0116I	14D
16	21DPR0913E	14A	51	21EPR0123S	14B
17	21DPR0940B	14B	52	21EPR0307Z	24A
18	21DPR1086M	14A	53	21EPR0455H	14B
19	21DPR1095U	14A	54	21EPR0475V	14B
20	21DPR1539X	14B	55	21EPR0540E	24B
21	21DPR1540M	14B	56	21EPR0643A	24A
22	21DPR1644H	14A	57	21EPR0710I	24B
23	21DPR1668R	14A	58	21PPR0296P	14A
24	21DPR1777Y	14A	59	21PPR0320Z	14A
25	21DPR1785G	14A	60	21PPR0320Z	14B
26	21DPR1907A	14A	61	21PPR0347F	14B
27	21DPR2116X	14A	62	21DPB0514G	14A
28	21DPR2122H	14B	63	21DPB0707V	14A
29	21DPR2130Q	14B	64	21DPR0175Z	14A
30	21DPR2199W	14B	65	21DPR0232Z	14B
31	21DPR2223F	14A	66	21DPR0241H	24A
32	21DPR2227B	14A	67	21DPR0263T	14B
33	21DPR2257W	14B	68	21DPR0359F	14A
34	21DPR2285S	24A	69	21DPR0359F	14B
35	21DPR2321G	14A			

Cuadro 4.17: Información sobre grupos con $N = 25$.

	escuela	grupo		escuela	grupo
1	21DPR0462S	14B	26	21EPR0325O	14C
2	21DPR0556G	14C	27	21EPR0380H	14B
3	21DPR0573X	14B	28	21EPR0399F	14C
4	21DPR0660S	14A	29	21EPR0415G	14B
5	21DPR0730X	14B	30	21EPR0420S	14A
6	21DPR0792J	14A	31	21DPB0391N	14A
7	21DPR0943Z	14C	32	21DPB0394K	14A
8	21DPR0949T	14A	33	21EPR0541D	14B
9	21DPR0959Z	14A	34	21EPR0629H	24B
10	21DPR1273G	14A	35	21EPR0708U	24B
11	21DPR1531E	14A	36	21EPR0710I	24A
12	21DPR1662X	14B	37	21EPR0729G	14B
13	21DPR2013A	14B	38	21EPR1501J	14C
14	21DPR2230P	14B	39	21DPB0472Y	14A
15	21DPR2380W	24A	40	21PPR0266V	14A
16	21DPR2420G	14A	41	21PPR0298N	14A
17	21DPR2541S	14A	42	21PPR0573B	14A
18	21DPR2818O	14A	43	21DPR0044G	24A
19	21DPR2921A	14A	44	21DPR0057K	14A
20	21DPR2977C	14A	45	21DPR0234Y	14A
21	21DPR3025M	24B	46	21DPR0254L	14B
22	21DPR3694C	24A	47	21DPR0276X	14B
23	21EPR0003F	14C	48	21DPR0291P	14B
24	21EPR0064T	14A	49	21DPR0316H	14B
25	21DAI0036O	14A	50	21DPR0390P	14B

Cuadro 4.18: Información sobre grupos con $N = 35$.

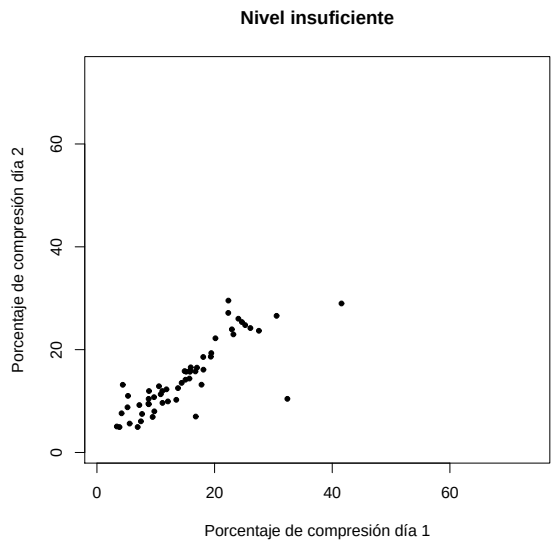
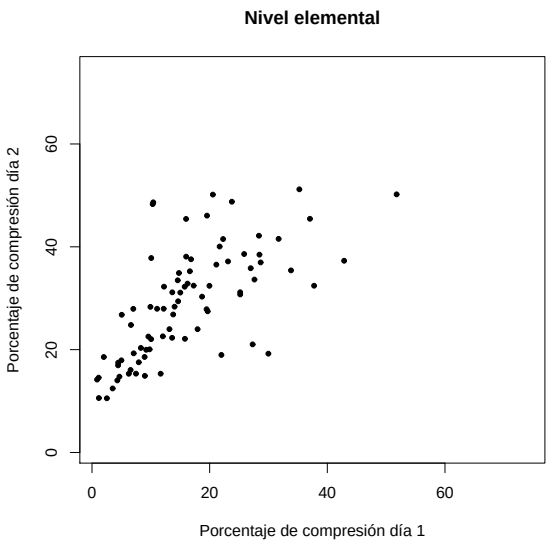
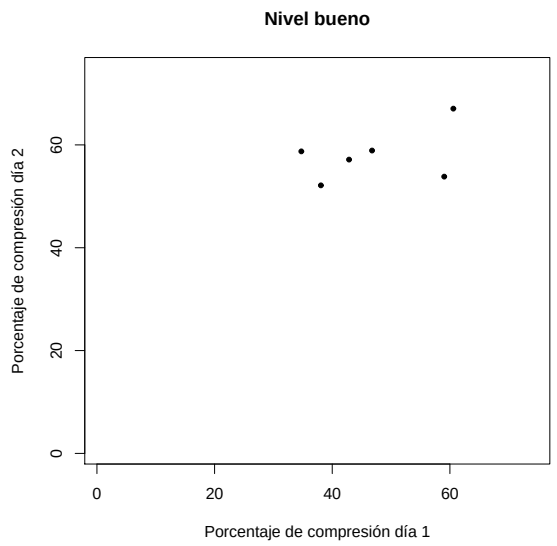
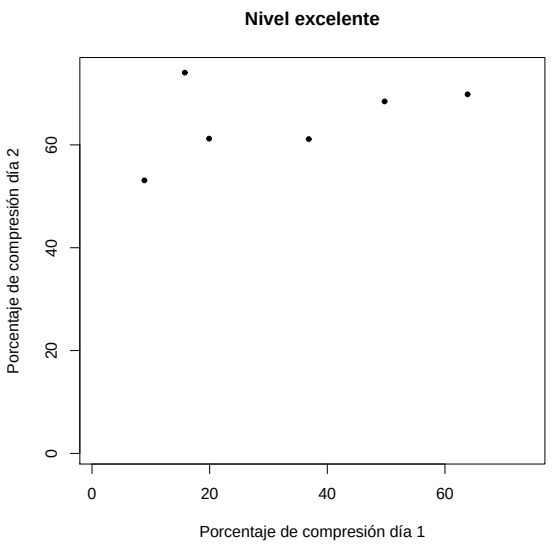
Datos reales (evaluación por días, grupos comparables en nivel de logro)

Siguiendo la idea de estudiar los resultados por días, se toman grupos comparables entre si. Se seleccionan grupos donde al menos 90 % de sus alumnos tienen el mismo nivel de logro en la sección de matemáticas, los niveles son {0: insuficiente, 1: elemental, 2: bueno, 3: excelente}. Adicionalmente el tamaño de los grupos es mayor a 10 alumnos.

De los 5,755 grupos del estado de Puebla de cuarto grado de primaria, el cuadro 4.19 muestra los resultados por días de los niveles 0, 1, 2 y 3. El nivel de logro se asigna conforme la calificación obtenida en la sección de interés, cabe señalar que dicha categoría es independiente de si el alumno fue o no detectado como sospechoso de copia. Se puede observar que para el nivel 0 no cambia mucho los niveles de compresión por días. Para el nivel 1 en el día 1 existen grupos con 0 % de compresión, mientras que en el día 2 el mínimo valor de compresión es de $\approx$ 10 %. Para el nivel 2, aunque en ambos días se alcanzan porcentajes de compresión considerablemente altos, los resultados del segundo día son más grandes que los del primer día. En el caso del nivel 3 el resultado de compresión en el primer día es un poco sorprendente ya que se puede decir que existe un poco de todo, sin embargo en el segundo día los porcentajes de compresión son en general muy grandes.

En este caso los resultados por días no son tan dramáticos como en la sub-sección anterior, donde es obvio el contraste del porcentaje de copia alcanzado en ambos días. Sin embargo resaltan los casos del nivel 1; donde la media es de $\approx$ 15 % en el día 1 y de $\approx$ 30 % en el día 2 y del nivel 3; donde la media es de $\approx$ 30 % en el día 1 y de $\approx$ 65 % en el día 2.

Si se desea comprender a fondo el comportamiento de los datos, es valido el estudio de evaluación por días en grupos comparables.

Nivel insuficiente	Nivel elemental
	
Nivel bueno	**Nivel excelente**
	

Cuadro 4.19: Compresión de datos reales por días en grupos similares. Niveles {0: insuficiente, 1: elemental, 2: bueno, 3: excelente}.

Capítulo 5

Conclusiones y Aportaciones

En general, la evidencia muestra que se está produciendo copia durante la aplicación de la prueba **ENLACE** y es razonable concluir que el problema no desaparecerá. Sin embargo, debe abordarse con el fin de garantizar la integridad, equidad y validez de los resultados de la prueba. A continuación se enlistan las conclusiones a las cuales se ha llegado con la realización de este proyecto de tesis, así como también las aportaciones realizadas para un mejor entendimiento del problema.

Conclusiones:

1. Los métodos tradicionales no abordan la posibilidad de que los alumnos no sólo copien en parejas, tampoco toman en cuenta el tamaño del grupo.
2. Por la naturaleza del proceso de copia, existen varios trucos o formas de copiar en un examen, y aunque puede ser una practica imposible el emular todos estos trucos, existe la posibilidad de reproducir un par de ellos.
3. Una *búsqueda de copia a nivel macro* es una opción novedosa, ya que se ha descartado esta posibilidad de análisis. El índice de copia a nivel de pares puede no reportar sospechosos de copia, sin embargo al analizar el grupo como un todo se pueden obtener porcentajes de compresión elevados lo cual da a entender que existe copia entre los trenes de respuesta.
4. Se detectó más copia de la que actualmente ha detectado la **SEP**.

Aportaciones:

1. Extensión del índice *scrutiny* de pares a tríos de alumnos.
2. Exploración del comportamiento a nivel grupal ante la copia.
3. Uso de técnicas de compresión de archivos para la búsqueda de copia.
4. Descripción y revisión de los diferentes métodos de detección copia expuestos en la literatura consultada, así como el detección de algunos errores en ellos.

5. Algoritmo para manipular los datos de datos en **ENLACE** y simulación de distintos métodos de copia, con los cuales se trata de emular el comportamiento de los alumnos al momento de realizar copia.

6. Análisis exploratorio del fenómeno de copia en la prueba **ENLACE**.

Apéndice A

Prueba ENLACE

Información tomada parcialmente de http://enlace.sep.gob.mx y del reporte
"Automatización y optimización para procesos de calificación y detección de copia"
por J. Van Horebeek, D. de la Rosa y M. Tapia

A.1. Introducción

La **E**valuación **N**acional de **L**ogro **A**cadémico en **C**entros **E**scolares (**ENLACE**) es una prueba del **S**istema **E**ducativo **N**acional que se aplica a planteles públicos y privados del país.

- En Educación Básica: a niños y niñas de tercero a sexto de primaria y jóvenes de tercero de secundaria, en función de los planes o programas de estudios oficiales en las asignaturas de español, matemáticas y ciencias.
- En Educación Media: a jóvenes que cursan el último grado de bachillerato para evaluar conocimientos y habilidades básicas adquiridas a lo largo de la trayectoria escolar para hacer un uso apropiado de la lengua (comprensión lectura) y las matemáticas (habilidad matemática).

La prueba se aplica en las diferentes escuelas primarias a nivel nacional: CONAFE, general, indígena y particular. Así como también en las distintas escuelas secundarias: general, particular, telesecundaria y técnica.

Las diferentes pruebas evalúan los contenidos establecidos en los planes y programas de estudios oficiales vigentes de la **S**ecretaría de **E**ducación **P**ública (**SEP**).

A.1.1. Metodología de calificación

Toma como base una escala subyacente, basada en la Teoría de Respuesta al Item (**IRT**, por sus siglas en inglés), ver sección A.1.2, utilizando el modelo de tres parámetros: adivinación, dificultad y discriminación.

El modelo **IRT** especifica que la probabilidad de que la persona s conteste el j-ésimo ítem correctamente (i.e. $P(X_{sj} = 1)$) depende de:

- La habilidad de la persona, usualmente denotada por θ_s
- La característica del j-ésimo ítem.

Se le asigna a cada alumno un valor en puntaje no sólo por la cantidad de respuestas correctas sino de cuáles respondió. Este valor se presenta de forma estandarizada, en una escala con media 500 y desviación estándar de 100 para cada grado-asignatura. Dado que la escala se establece para cada grado-asignatura, es incorrecto hacer comparaciones de estos puntajes entre niveles, asignaturas y grados diferentes.

De acuerdo al grado de dificultad de las preguntas, se establecen cuatro niveles de logro. Se determina el conocimiento mínimo que debe tener un alumno para cada grado y se establece el primer nivel de dificultad (bajo). A partir de la pregunta más difícil, se determinan las características de los alumnos sobresalientes y se establece el segundo nivel de dificultad (alto). El grupo de preguntas que quedan entre estos dos puntos de corte, determinan el nivel medio de dificultad.

Establecidos estos tres niveles, se determina el punto medio de cada uno, para definir el cuarto nivel de dificultad (insuficiente, elemental, bueno y excelente), de esta forma los alumnos en el nivel insuficiente responden al menos al 50 % de los reactivos de dificultad baja y los alumnos en el nivel excelente responden al menos al 50 % de los reactivos de dificultad alta.

A.1.2. Item Response Theory

Los modelos **IRT** son muy utilizados en evaluación de la educación, por ejemplo las calificaciones obtenidas de exámenes como las aplicadas por **ETS** ó **ENLACE** de la **SEP** están basadas en una modelación estadística usando modelos **IRT**.

Estos modelos aplican funciones matemáticas que especifican la probabilidad de un resultado discreto, como una respuesta correcta, en términos de los parámetros de la persona evaluada y de la pregunta en cuestión. Dichos parámetros son,

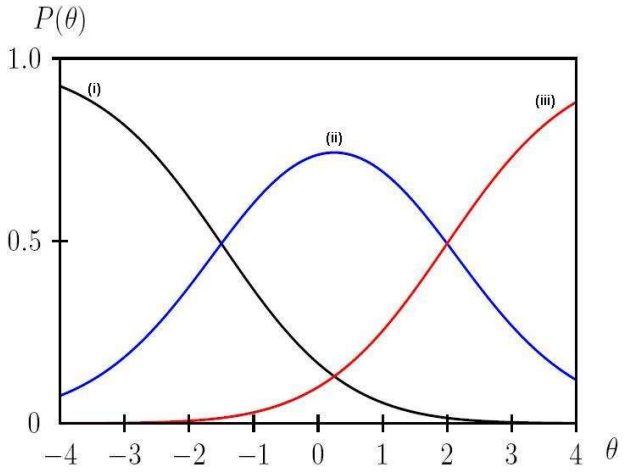
- a nivel del sujeto: habilidad de contestar bien,
- a nivel de la pregunta: dificultad, discriminación y la probabilidad de adivinar la respuesta correcta.

La mayoría de los modelos **IRT** se han desarrollado para preguntas dicotómicas, una respuesta incorrecta y otra correcta. No obstante también existen técnicas para preguntas politómicas, llamadas de opción múltiple.

Como ejemplo de **IRT** se tiene la pregunta: *¿Cuál es el área de un círculo con un radio de 3 cm?* y las respuestas posibles son: *(i)* 9.00 cm^2, *(ii)* 18.85 cm^2 y *(iii)* 28.27 cm^2. La primera de las tres opciones es posiblemente la peor, la segunda está mal; pero implica un poco de conocimiento (el área se confunde con la circunferencia) y la tercera es la

respuesta correcta.

Si la pregunta sobre el área del círculo ha sido calibrada, como se muestra en el cuadro A.1, entonces se tienen tres curvas en la grafica, una para cada posible respuesta. La habilidad de la persona se denota con θ y se traza a lo largo del eje horizontal. El eje vertical indica la probabilidad $P(\theta)$ de cada persona dado su nivel de habilidad. Debido que cada alumno sólo puede dar una respuesta a la pregunta y las tres opciones son mutuamente excluyentes, la suma de las tres probabilidades para cada valor θ es 1.



Cuadro A.1: Calibración de la pregunta.

La curva de la opción *(i)* es alta en los más bajos niveles de habilidad y disminuye gradualmente al par que la gente se vuelve más "inteligente"; para las personas con habilidad por debajo de -1.5, esta es la opción más probable. La curva de la opción *(ii)* tiene el aspecto de una campana de gauss, como resultado, esta opción tiene la mayor probabilidad de ser seleccionada en los niveles de habilidad entre -1.5 y 2.0, en comparación con las otras dos opciones. La probabilidad de la respuesta correcta, opción *(iii)*, es muy pequeña para una baja habilidad, pero al aumenta la habilidad esta también aumenta y se convierte en la mayor probabilidad en los niveles superiores a 2.0. Aun así las personas en cualquier nivel de habilidad aún tienen un no-cero de probabilidad de seleccionar cualquiera de las tres opciones, por lo tanto, incluso aquellas personas con un valor de θ alto tienen una pequeña probabilidad de seleccionar la opción (i) y una probabilidad ligeramente mayor de seleccionar la opción (ii).

Una gran parte de **IRT** es acerca de los diversos modelos posibles para $P(\theta)$. Además, se escribe $P(\theta)$ para demostrar que la probabilidad de una respuesta correcta es una función de la habilidad, θ. Sin embargo, P también depende de las propiedades de la pregunta, sus parámetros. Para preguntas dicotómica, se examinan los modelos **IRT** que tengan una, dos o tres parámetros y la probabilidad predicha por los modelos se denotan como $P(X_{ij} = 1|\theta_i, b_j)$, $P(X_{ij} = 1|\theta_i, a_j, b_j)$ o $P(X_{ij} = 1|\theta_i, a_j, b_j, c_j)$, donde a_j, b_j y c_j

son los parámetros de la pregunta.

Una gran parte de **IRT** es acerca de los diversos modelos posibles para el cálculo de la probabilidad de que el alumno con una habilidad θ, conteste bien una pregunta con ciertos parámetros, $P(\theta)$.

Para preguntas dicotómicas:

- **One parameter logistic (1PL) model**
  El modelo más simple de **IRT**, para respuestas dicotómicas, tiene un solo parámetro. Y esta dada por

$$P(X_{ij} = 1|\theta_i, b_j) = \frac{exp[\theta_i - b_j]}{1 + exp[\theta_i - b_j]},$$

  donde la expresión $exp[\theta_i - b_j]$, en el numerador, denota que el modelo predice la probabilidad de una respuesta correcta, con la interacción entre la habilidad individual θ_i del alumno y el parámetro b_j del ítem, llamado parámetro de dificultad.

- **Two parameter logistic (2PL) model**
  El modelo 2PL predice la probabilidad de obtener una respuesta correcta en cualquier examen con dos parámetros para las preguntas. La curva característica toma la forma de una distribución logística de 2-parámetros:

$$P(X_{ij} = 1|\theta_i, a_j, b_j) = \frac{exp[a_j(\theta_i - b_j)]}{1 + exp[a_j(\theta_i - b_j)]},$$

  donde b_j es el parámetro de dificultad y a_j es llamado el parámetro discriminante.

- **Three parameter logistic (3PL) model**.
  Esta basado en el modelo 2PL agregándole un tercer parámetro, denotado por c_j.

$$P(X_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{exp[a_j(\theta_i - b_j)]}{1 + exp[a_j(\theta_i - b_j)]}$$

  donde b_j es el nivel de dificultad de la pregunta j, a_j es el parámetro de discriminación de la pregunta j, por último c_j representa la probabilidad de que alumnos con baja habilidad contesten bien una pregunta j.

Para preguntas politómicas:

- **Graded Response Models (GRM)**
  La probabilidad que el alumno i con habilidad θ_i seleccione al azar la opción k del ítem j, esta dado por

$$P(X_{ij} = 1|\theta_i, a_j, b_{jk}) = \frac{exp[a_j(\theta_i - b_{jk})]}{1 + exp[a_j(\theta_i - b_{jk})]},$$

  donde b_{jk} es la probabilidad de seleccionar la opción k en pregunta j.

- **Nominal Response Model (NRM)**
  La probabilidad que el alumno i con habilidad θ_i seleccione al azar la opción k del ítem j, esta dado por

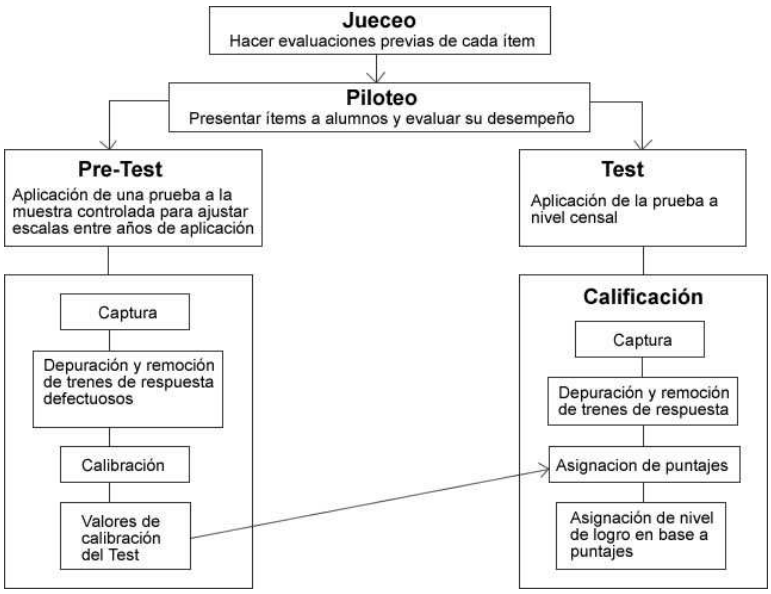
$$P(X_{ij} = 1|\theta_i, \zeta_{jk}, \lambda_{jk}) = \frac{\exp\{\zeta_{jk} + \lambda_{jk}\theta_i\}}{\sum_{v=1}^{V} \exp\{\zeta_{jv} + \lambda_{jv}\theta_i\}}$$

  donde ζ_{jk} y λ_{jk} son parámetros del ítem; intercepción y discriminación del j-ésimo ítem en la k-ésima opción de respuesta, respectivamente. El valor de θ y los parámetros ζ y λ pueden ser estimados usando el método *maximum likelihood*.

En los modelos de **IRT**, la probabilidad de una respuesta correcta depende de la habilidad del examinado y los parámetros del ítem. Lo que se conoce son las respuestas de los alumnos en la prueba. El problema es determinar el valor de θ para cada alumno y los parámetros para cada ítem.

A.2. Proceso de calificación

Existen cinco etapas en la prueba **ENLACE**, estas son: jueceo, piloteo, test, pre-test y calificación. Las cuales se muestran en el diagrama del cuadro A.2.



Cuadro A.2: Etapas en la prueba **ENLACE**.

La última etapa, la de calificación, es la única de interés en esta tesis, la cual consta de tres partes: captura, depuración y asignación de puntajes. A continuación se describen estas partes.

- Captura

1. Participantes. Únicamente personal de la Unidad de Plantación y Evaluación de Políticas Educativas de la **SEP**.
2. Propósito. Captura de las hojas de respuesta y ponerlas en formato electrónicos.
3. Ejecución. Se leen las boletas por medio de hardware especializado y se ponen en formato electrónico obteniendo respuestas en formato crudo. El formato de salida puede ser fácilmente traducido a un formato de *software* comercial de análisis de datos.

   A través de la plantilla de respuestas se dicotomiza el tren de respuesta de cada alumno. Se hacen estadísticos de las respuestas de los alumnos. En el proceso de captura también se incluye el grupo al cual pertenece el alumno.
4. Resultados. Trenes de respuesta dicotomizados en formato electrónico. Estadísticos de los ítems que se analizaran posteriormente.

- Depuración

  1. Participantes. Únicamente personal de la Unidad de Plantación y Evaluación de Políticas Educativas de la **SEP**.
  2. Propósito. Limpiar los trenes de respuesta para su posterior calificación. Marcar trenes con defectos y tratarlos de acuerdo a ellos.
  3. Ejecución. El proceso de depuración es exactamente el mismo usado en el pre-test con la diferencia que no se eliminan los trenes de respuesta. A los trenes mascados como defectuosos se les agrega una leyenda en la cual se le indica al alumno que los resultados no son confiables ya sea porque fue marcado como una prueba con copia o por alto grado de incertidumbre en la calificación debido a pocos ítems respondidos. Los trenes con todos los ítems correctos se les asignara la mayor puntuación. Los trenes con todos los ítems incorrectos se les asignara la menor puntuación.
  4. Resultados. Trenes marcados de acuerdo a sus características.

- Asignación de puntajes

  1. Participantes. Únicamente personal de la Unidad de Plantación y Evaluación de Políticas Educativas de la **SEP**.
  2. Propósito. Calificar las pruebas previamente depuradas. Obtener la variable latente de cada alumno. Entregar los resultados de la prueba y dar a cada nivel de logro en cada asignatura evaluada.
  3. Ejecución. Con los parámetros calibrados por la muestra controlada se califica toda la población por lotes, los parámetros son fijos y son únicos para toda la población. Los alumnos en la muestra entran también el proceso de calificación.

El resultado del *software* de calificaron se encuentra estandarizado (distribución normal $m = 0$, $r = 1$) y se reescala a una normal de 200 a 800 con media 500 con una transformación lineal. A cada alumno se le asigna un puntaje en esta escala. Una vez obtenido el puntaje se hacen los puntos de corte para determinar los niveles de logro. Los niveles de logro son los siguientes: Insuficiente, Elemental, Bueno y Excelente. La manera en que se hizo en 2006 fue:

Se ordenan los reactivos de mayor a menor dificultad de acuerdo al parámetro b del modelo **IRT**. Un grupo de expertos determina los reactivos que corresponden al nivel bajo que debe saber el evaluado así como para el nivel alto. Estos dos puntos sirven de referencia superior e inferior en la escala dividiéndola en tres regiones. La manera en que se establecen los puntos de corte se explica a continuación con ayuda del cuadro A.3.

Cuadro A.3: Puntos de corte.

De las tres regiones previamente divididas se hace una consideración extra: para tener un nivel de logro debe de al menos contestar el 50 % de una región. Los niveles de logro están calculados de acuerdo a esta premisa y de esta forma quedan delimitados los cuatro niveles de logro. Por ejemplo para tener un nivel de logro bueno es necesario haber contestado al menos el 50 % de la región de nivel medio. Los puntos de corte de cada nivel de logro quedaran fijos para todas las subsecuentes aplicaciones de **ENLACE**.

4. Resultados. Pruebas calificadas con valor theta (variable latente de habilidad) tanto de matemáticas como de español en escala normal de 200 a 800 con media 500.

   Se entrega el puntaje de la prueba y dependiendo de este se le cataloga en uno de los cuatro niveles de logro. Se añade un análisis personal de las fortalezas y debilidades del evaluado en el contenido temático a través de la página Web de **ENLACE** [3].

Bibliografía

[1] Robert L. Brennan Bradley A. Hanson, Deborah J. Harris. A comparison of several statistical methods for examining allegarions of copying. *ACT Research Report Series*, September 1987.

[2] Dorothy T. Thayer Charles Lewis. The power of the k-index to detect copying. *Educational Testing Service*, December 1998. RR-98-49.

[3] Secretaria de Educación Publica. http://enlace2007.sep.gob.mx/.

[4] Paul W. Holland. Assessing unusual agreement between the incorrect answers of two examinees using the k-index: statistical theory and empirical support. *ETS Research Report*, (96-7), March 1996.

[5] G. Karabatsos. Comparing the aberrant response detection performance of thirty-six person-fit statistics. 2003.

[6] Rob R. Meijer Leonardo Sotoridona. Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement*, 39(2):115–132, Summer 2002.

[7] Wim J. van der Linder Leonardo Sotoridona. Detecting answer copying using the kappa statistic. *Applied Psychological Measurement*, 30(5):412:431, September 2006.

[8] Khalid Sayood. *Introduction to data compression*. Morgan Kaufmann Publishers, 2000.

[9] CIMAT Soluciones Matemáticas Avanzadas. Automatización y optimización para procesos de calificación y detección de copia. Octubre 2007.

[10] James A Wollack. A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4):307–320, December 1997.

[11] James A Wollack. Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40(3):189:205, Fall 2003.