A derivation of the optimal answer-copying index and some applications

Mauricio Romero[1,2] Álvaro Riascos[2,3] Diego Jara[2]

[1]UC San Diego

[2]Quantil

[3]Universidad de los Andes

June 17 2014

Table of Contents

Introduction

- In Colombia every student has to take a multiple-choice exam in the 5th, 9th and 11th grades.
- The Instituto Colombiano para la Evaluación de la Educación (ICFES) is in charge of making, distributing and applying these exams.
- Between 2010 and 2011, the ICFES used the κ index for answer-copy detection, based on the work of Sotaridona, van der Linden, and Meijer (2006)
- Jara, Riascos, and Romero (2010) showed that in practice the κ index has a larger type-I error rate than predicted by theory

How to choose an answer-copy index?

- All the indices in the literature are ad-hoc and there are no theoretical results justifying the use of one index over the other.
- We could use empirical or simulated type-I and type-II calculations... such as Wollack (2003) but the set of indices compared is not comprehensive (e.g. Wesolowsky (2000) index)

What we do...

- Provide theoretical foundations that justify the use of indices that reject the null hypothesis of no cheating for large values of the number of identical answers
- Comapres empirical type-I and type-II error rates using Monte Carlo simulations for the indeces developed by Wollack (1997) and Wesolowsky (2000) , both based on the work of Frary, Tideman, and Watts (1977).
- Compare results with two strategies to control cheating: stricter proctoring and diversification of questions.
- Outline a procedure to detect massive cheating.

Table of Contents

Background

- Multiple choice exams are frequently used as an efficient and objective way of evaluating knowledge.
- But... they are more vulnerable to answer-copying than tests based on open questions
- It is normal for two answer patterns to have similarities
- Answer-copy indices try to detect similarities so unlikely that answer-copying becomes a more natural explanation than chance

Setting up of the problem

- Test if individual supected of cheating (denoted by c) copied from the individual who is suspected of being the source of answers (denoted by s)
- Assume that there are N questions and n alternatives for each questions
- $\gamma_{cs} \rightarrow$ number of questions that s copied from c.

$$H_0 : \gamma_{cs} = 0$$
$$H_1 : \gamma_{cs} > 0$$

Setting up of the problem

- Let I_{csi} be equal to one when individuals c and s have the same answer to question i and zero otherwise

$$M_{cs} = \sum_{i=1}^{N} I_{csi}. \tag{1}$$

- Under H_0 $M_{cs} \sim B(\pi_1, ..., \pi_N)$ (i.e is distributed Poisson binomial)
- Let $f_N(x; \pi_1, ..., \pi_N)$ be the probability mass function (pmf) at x
- If $\pi_1 = \pi_2 = ... = \pi_N$ then we have a binomial distribution

Setting up of the problem

- Let A denote the set of questions that student c copied from s.
- If $|A| = k$, it means that $\gamma_{cs} = k$
- M_{cs} has the following probability mass function (pmf) $\hat{f}_N(x; \pi_1, ..., \pi_N, A)$,

$$\hat{f}_N(x; \pi_1, ..., \pi_N, A) \doteq f_N(x, \pi'_1, .., \pi'_N)$$

s.t.

$$\pi'_i = \begin{cases} 1 & \text{if } i \in A \\ \pi_i & \text{if } i \notin A \end{cases}$$

Neyman-Pearson Lemma

Theorem

Neyman-Pearson Lemma

Consider testing $H_0 : \theta = \theta_0$ *against* $H_1 : \theta = \theta_1$ *where the pmf is* $f(\mathbf{x}|\theta_i)$, $i = 0, 1$, *using a test with rejection region* R *that satisfies*

$$\begin{aligned} \mathbf{x} \in R \text{ if } f(x|\theta_1) > f(x|\theta_0)k \\ \mathbf{x} \in R^c \text{ if } f(x|\theta_1) < f(x|\theta_0)k \end{aligned} \tag{2}$$

for some $k \geq 0$, *and*

$$\alpha = P_{H_0}(\mathbf{X} \in R) \tag{3}$$

Then

1. *(Sufficiency) Any test that satisfies 2 and 3 is a UMP level* α *test.*
2. *(Necessity) If there exists a test satisfying 2 and 3 with* $k > 0$, *then every UMP level* α *test is a size* α *test (satisfies 3) and every UMP level* α *test satisfies 2 except perhaps on a set A satisfying* $P_{H_0}(\mathbf{X} \in A) = P_{H_1}(\mathbf{X} \in A) = 0$.

the test is the uniformly most powerful (UMP) level α *test.*

Simple hypothesis

- Simple hypothesis test $H_0 : A = \emptyset$ and $H_1 : A = A_1$

$$\lambda^A(x) = \frac{\hat{f}_N(x; \pi_1, ..., \pi_N, A)}{f_N(x; \pi_1, ..., \pi_N)}$$

- To find the critical value of the test we need the greatest value c such that:

$$1 - P_{H_0}\left(\frac{\hat{f}_N(x; \pi_1, ..., \pi_N, A)}{f(x; \pi_1, ..., \pi_N)} < c\right) = P_{H_0}\left(\frac{\hat{f}_N(x; \pi_1, ..., \pi_N, A)}{f_N(x; \pi_1, ..., \pi_N)} > c\right) \leq \alpha$$

- How to fid then UMP for more complex alternative hypothesis (In particular $H_1 : \{A : |A| \geq 1\}$) ?

Useful results

Theorem (Theorem 2 in (Wang, 1993))

The pmf of a poisson binomial satisfies the following inequality:

$$f_N(x;\pi_1,\pi_2,...,\pi_N)^2 > C(x) f_N(x+1;\pi_1,\pi_2,...,\pi_N) f_N(x-1;\pi_1,\pi_2,...,\pi_N)$$

where $C(x) = \max\left(\frac{x+1}{x}, \frac{N-x+1}{N-x}\right)$

Lemma

$\frac{\hat{f}_N(x;\pi_1,...,\pi_N,A)}{f_N(x,\pi_1,...,\pi_N)}$ *is increasing in* $x \in \{0,...,N\}$ *for all* A.

Complex Hypothesis

Given that $\frac{\hat{f}(x;\pi_1,\ldots,\pi_N;A)}{f(x;\pi_1,\ldots,\pi_N)}$ is increasing in x for all $A \Rightarrow$ for every c there exists a k^* such that

$$P_{H_0}\left(\frac{\hat{f}_N(x;\pi_1,\ldots,\pi_N,A)}{f(x;\pi_1,\ldots,\pi_N)} < c\right) = \sum_{w=0}^{k^*} f(w,\pi_1,\ldots,\pi_N)$$

Remark

Notice that the rejection region is the same for all A, thus if we reject the null hypothesis when $M_{cs} > k^*$, we get the UMP for all A such that $|A| \geq 1$.

But...

However, π_j must be estimated somehow (it was taken as known in this section and thus in empirical applications we don't really have the UMP) and different ways to go about this yield different results

Table of Contents

Definitions

- $\pi_{iv}^{j} \rightarrow$ probability student j answers option v on question i.
- $\pi_{i} \rightarrow$ probability two students have the same answer on question i
- $\pi_{iv_s}^{c} \rightarrow$ probability that individual c answered the option v_s which was chosen by s in question i.

Conditional vs Unconditional

- Assuming independent answers: $\pi_i = \sum_{v=1}^{n} \pi_{iv}^{c} \pi_{iv}^{s}$.
- Conditional on the answers of s, the probability that individual c has the same answer as individual s: $\pi_i = \pi_{iv_s}^{c}$

Critical Values: Poisson Binomial vs Normal

- M_{cs} is the sum of n Bernoulli variables and has mean $\sum_{i=1}^{N} \pi_i$ and variance $\sum_{i=1}^{N} \pi_i(1-\pi_i)$.
- $\frac{M_{cs}-\sum_{i=1}^{N} \pi_i}{\sqrt{\sum_{i=1}^{N} \pi_i(1-\pi_i)}} \rightarrow_d N(0,1)$

How to calculate π^j_{iv}?

The main difference between indices is how they calculate π^j_{iv}

ω index

The ω index is based on the work of Wollack (1997)

$$\pi_{iv}(\theta_j) = \frac{e^{\xi_{iv}+\lambda_{iv}\theta_j}}{\sum_{h=1}^{m} e^{\xi_{ih}+\lambda_{ih}\theta_j}}, \tag{4}$$

1. ω_1: unconditional
2. ω_2: conditional
3. ω_1^s: unconditional standardized
4. ω_2^s: conditional standardized

γ index

- The γ index is based on the work of Wesolowsky (2000)
- r_i is the proportion of students that got the right answer in question i
- c_j is the proportion of questions answered correctly by individual j
- $p_i = (1 - (1 - r_i)^{a_j})^{1/a_j} \rightarrow$ Probability that student j has the correct answer in question i

γ index

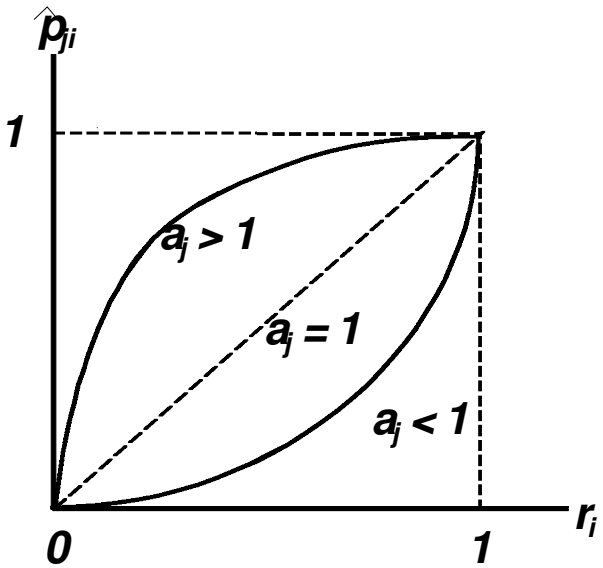


FIG. 3. Expression (2).

γ index

- a_j is estimated by solving the equations $\frac{\sum_{i=1}^{N} p_i}{n} = c_j$
- To estimate the probability of incorrect options we find the proportion of students that answered each incorrect option.

1. γ_1: unconditional
2. γ_2: conditional
3. γ_1^s: unconditional standardized
4. γ_2^s: conditional standardized

Table of Contents

Data

- 5th and 9th grade tests for 2009
- 12 different exams that can be classified in three broad fields, Science, Mathematics and Language for each grade (5th and 9th) for two different dates (May and September, 2009)
- Answer patterns for each student and answer keys

Summary Statistics

Test	Subject	Grade	Month	Questions	Students	Schools
PBA5041F1	Math	5th	May	48	60,099	3,421
PBA5041F2	Math	5th	Oct	48	403,624	31,827
PBA5042F1	Language	5th	May	36	60,455	3,441
PBA5042F2	Language	5th	Oct	36	402,508	31,642
PBA5043F1	Science	5th	May	48	60,404	3,432
PBA5043F2	Science	5th	Oct	48	405,537	31,833
PBA9041F1	Math	9th	May	54	44,577	1,110
PBA9041F2	Math	9th	Oct	54	303,233	9,059
PBA9042F1	Language	9th	May	54	44,876	1,110
PBA9042F2	Language	9th	Oct	54	302,781	9,044
PBA9043F1	Science	9th	May	54	44,820	1,107
PBA9043F2	Science	9th	Oct	54	30,3723	9,053

Source: ICFES. Calculations: Authors.

Note: The number of schools corresponds to the number of examination rooms.

Table of Contents

Methdolody

1. 100,000 pairs are picked in such a way that for each couple, the individuals performed the exam in different examination rooms.
2. The answer-copy methodology is applied to these pairs, and the proportion of pairs accused of cheating is the empirical type-I error rate estimator.
3. To calculate the power of the index, the answer-pattern for individual c is changed by replacing k of his answers to correspond to those of the individual s.
    1. The level of copy, k, is set, and is defined as the number of answers transferred from s to c.
    2. k questions are selected randomly.
    3. Individual c's answers for the k questions are changed to replicate exactly those of individual s. Answers that were originally identical count as part of the k questions being changed.
4. We apply the answer-copy methodology to the tampered couples. The proportion of pairs accused of cheating is the power of the index for a copying level of k.

Type-I error: γ index

Exam	Subject	Grade	Month	γ_1	γ_2	γ_1^s	γ_2^s
PBA5041F1	Mathematics	5th	May	0.66	2.20	0.41	0.76
PBA5041F2	Mathematics	5th	October	0.87	2.44	0.59	1.11
PBA5042F1	Language	5th	May	1.20	2.18	0.77	1.16
PBA5042F2	Language	5th	October	1.21	2.36	0.92	1.49
PBA5043F1	Science	5th	May	1.05	2.59	0.73	1.38
PBA5043F2	Science	5th	October	0.74	1.81	0.61	1.24
PBA9041F1	Mathematics	9th	May	1.38	1.97	0.96	1.26
PBA9041F2	Mathematics	9th	October	2.15	2.14	1.69	1.53
PBA9042F1	Language	9th	May	0.85	2.24	0.56	1.04
PBA9042F2	Language	9th	October	0.84	1.92	0.59	1.34
PBA9043F1	Science	9th	May	1.32	2.06	0.93	1.42
PBA9043F2	Science	9th	October	1.02	1.70	0.74	1.37

Source: ICFES. Calculations: Authors.
Number of copy-free couples accused of copying (for every 1000 pairs)
at a 99.9% confidence level

Type-I error: ω index

Exam	Subject	Grade	Month	ω_1	ω_2	ω_1^s	ω_2^s
PBA5041F1	Mathematics	5th	May	0.42	1.28	0.23	0.52
PBA5041F2	Mathematics	5th	October	0.61	1.38	0.31	0.78
PBA5042F1	Language	5th	May	0.80	1.61	0.46	0.73
PBA5042F2	Language	5th	October	0.86	1.51	0.55	0.95
PBA5043F1	Science	5th	May	0.79	1.37	0.47	0.87
PBA5043F2	Science	5th	October	0.82	1.47	0.57	0.88
PBA9041F1	Mathematics	9th	May	0.89	1.53	0.58	0.89
PBA9041F2	Mathematics	9th	October	1.22	1.53	0.99	1.07
PBA9042F1	Language	9th	May	0.55	1.44	0.31	0.65
PBA9042F2	Language	9th	October	0.86	1.47	0.63	0.97
PBA9043F1	Science	9th	May	0.78	1.46	0.59	0.98
PBA9043F2	Science	9th	October	0.78	1.36	0.63	1.03

Source: ICFES. Calculations: Authors.
Number of copy-free couples accused of copying (for every 1000 pairs)
at a 99.9% confidence level

Power

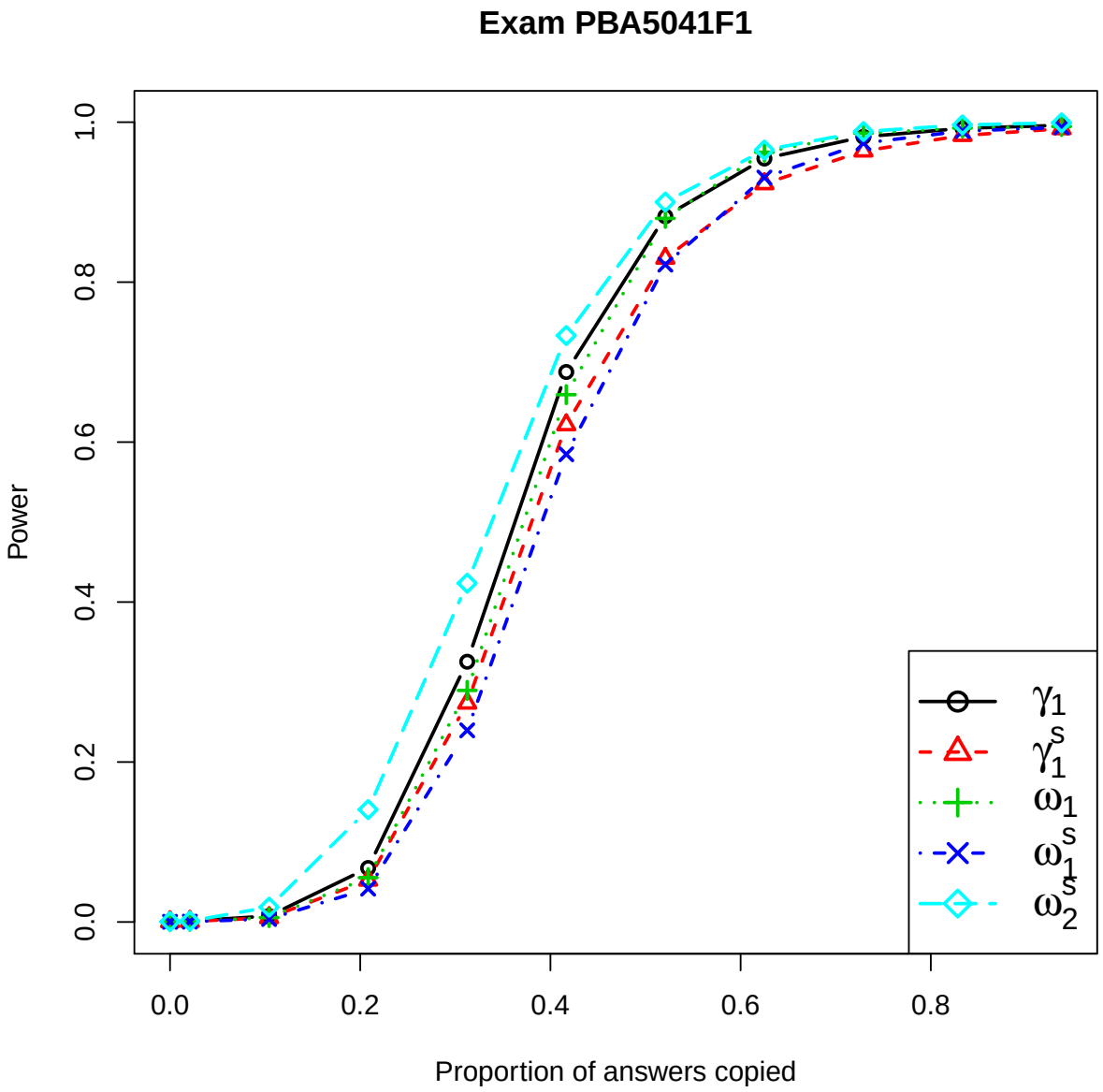
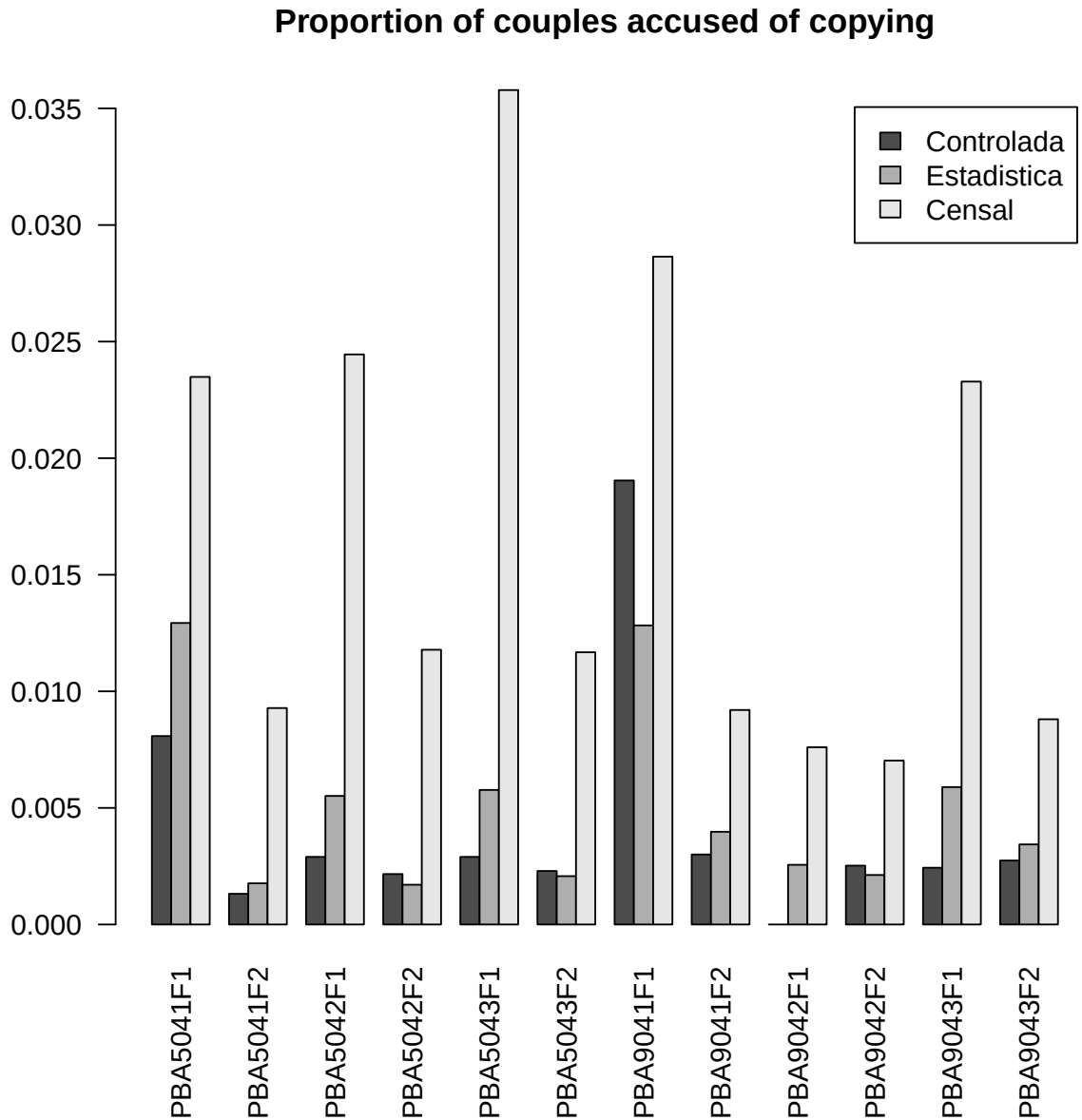

Exam PBA5041F1
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Proportion of answers copied
γ_1
γ_1^s
ω_1
ω_1^s
ω_2^s

Table of Contents

- The ICFES *randomly* assigns schools to three different samples that have different levels of proctoring
- Most of the schools are assigned to the *censal* sample where the ICFES distributes the exams to the schools and the schools perform the proctoring.
- In the *controlada* the proctoring is done by the central government (i.e. the ICFES)
- In the *estadistica* the proctoring is done by the regional government (Secretarias de Educacion).

	Controlada	Estadistica	Censal
	5th Grade May		
No. Students	1,413	7,648	60,099
No. of Schools	141	680	3,421
Students/School	10.02	11.25	17.57
	(0.88)	(0.47)	(0.46)
	5th Grade October		
No. Students	3,830	26,393	403,624
No. of Schools	958	654	31,827
Students/School	4.00	40.36	12.68
	(0.13)	(1.36)	(0.11)
	9th Grade May		
No. Students	1,150	6,690	44,577
No. of Schools	75	351	1110
Students/School	15.33	19.06	40.16
	(1.62)	(1.08)	(1.44)
	9th Grade October		
No. Students	3,106	24,387	303,233
No. of Schools	495	487	9,059
Students/School	6.27	50.08	33.47
	(0.25)	(1.74)	(0.35)


Proportion of couples accused of copying
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Controlada
Estadistica
Censal
PBA5041F1
PBA5041F2
PBA5042F1
PBA5042F2
PBA5043F1
PBA5043F2
PBA9041F1
PBA9041F2
PBA9042F1
PBA9042F2
PBA9043F1
PBA9043F2

- The SABER tests are administrated over three sessions, wherein students answer a different subject in each session.
- In May, every student took the same subject at the same time
- In October, only one third of the students took the same subject in each session

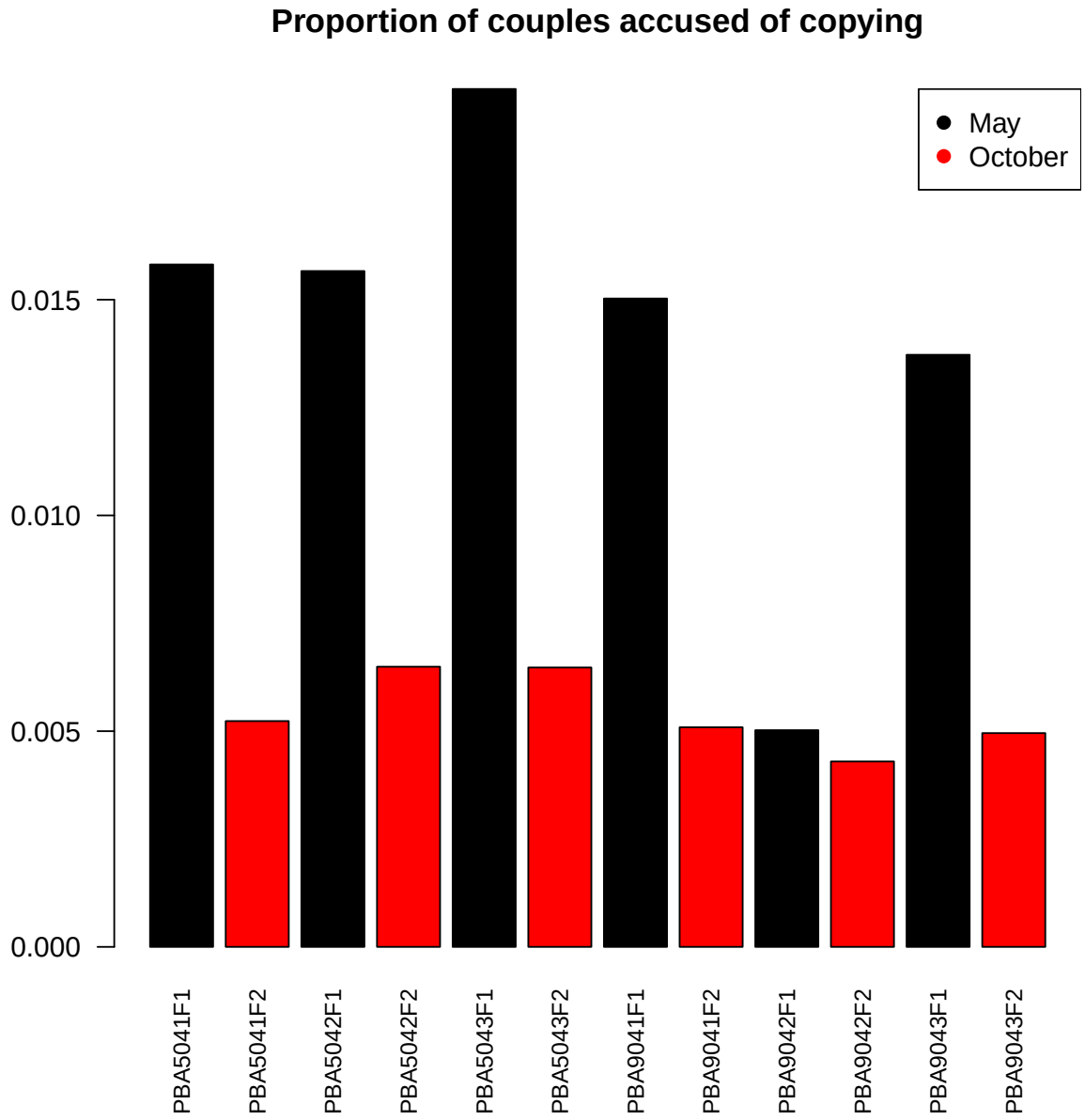

Proportion of couples accused of copying
May
October
0.015
0.010
0.005
0.000
PBA5041F1
PBA5041F2
PBA5042F1
PBA5042F2
PBA5043F1
PBA5043F2
PBA9041F1
PBA9041F2
PBA9042F1
PBA9042F2
PBA9043F1
PBA9043F2

Table of Contents

Correction

- To determine whether massive cheating has occurred in an examination room, multiple hypotheses must be tested.
- The significance level for a multiple test $\alpha_{MT} = 1 - (1 - \alpha_I)^n \leq \alpha_I \cdot n$ (assuming independence)
- If this "correction" is made, in most cases the power of the test is severely diminished
- Solution: Bonferroni-Type procedures that control the false positive rate

Correction

Suppose there are $H_1, ..., H_m$ hypotheses to be tested, ordered such that their $p - values$ follow $P_1 \leq P_2 \leq ... \leq P_m$, where P_i is the $p - value$ of hypothesis H_i. Let k be the greatest integer i, such that:

$$P_i \leq \frac{i}{m} p^*. \tag{5}$$

H_i is then rejected for every $i \in \{1, ..., k\}$. This controls for the false positive rate to a maximum of p^* (Benjamini & Hochberg, 1995)

Results

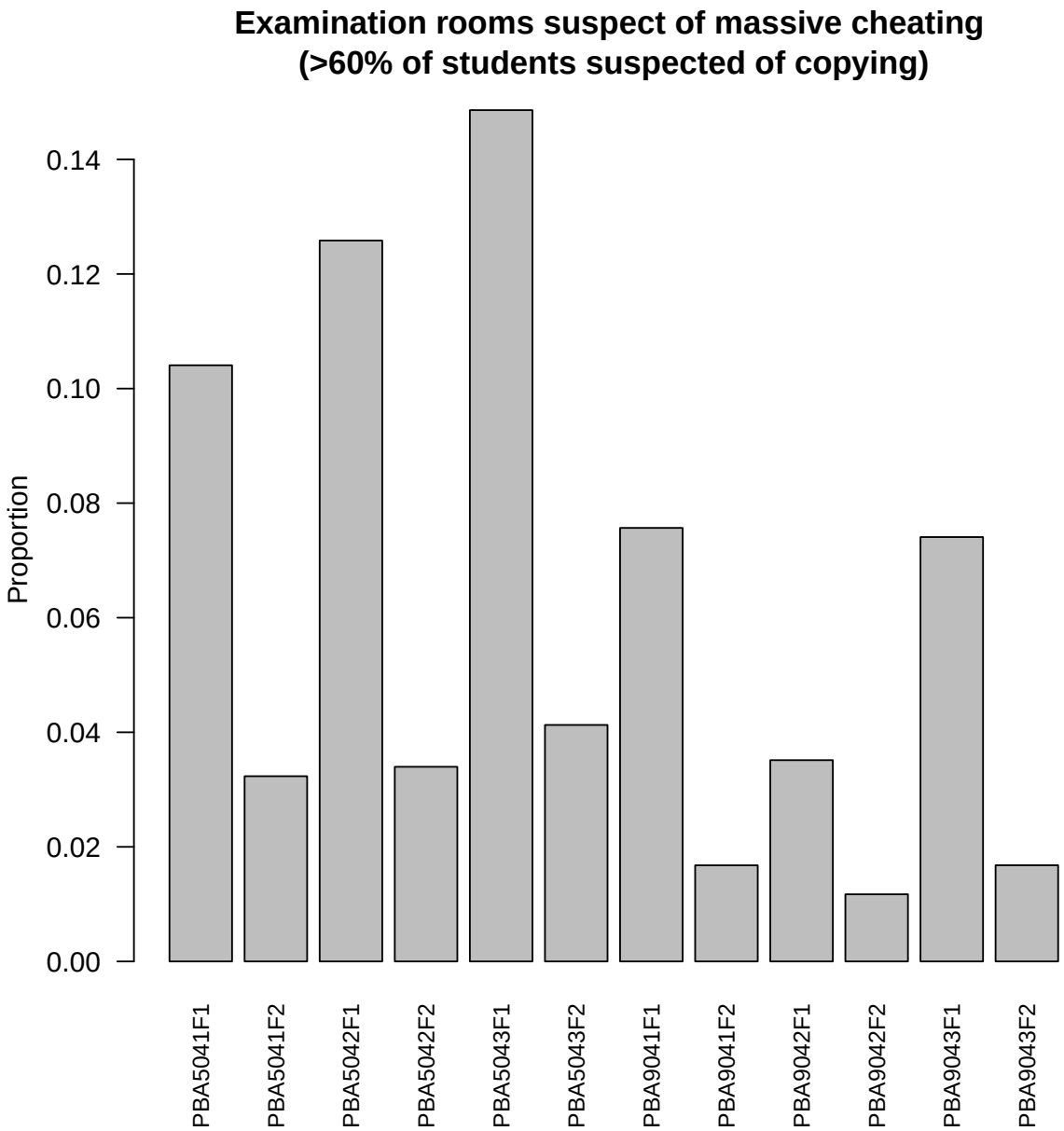

Examination rooms suspect of massive cheating
(>60% of students suspected of copying)
Proportion
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
PBA5041F1
PBA5041F2
PBA5042F1
PBA5042F2
PBA5043F1
PBA5043F2
PBA9041F1
PBA9041F2
PBA9042F1
PBA9042F2
PBA9043F1
PBA9043F2

Table of Contents

Conclusions

- Theoretical justification of the use of a variety of statistical test found in the literature to detect answer-copying.
- The most powerful test is a conditional test that models student behavior using a nominal response model and relies on the central limit theorem to find critical values.

Conclusions

- Increasing the level of proctoring in this setting can halve the prevalence of cheating.
- Randomizing the time at which each student takes each portion of the test can reduce the level of cheating by 50%.
- Methodology for detecting massive cheating while controlling for the false positive rate using a Bonferroni type procedure.

Thank you

- Gracias
- Asante Sana
- Merci
- Obrigado
- Grazie

Bibliography I

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological).*, *57*, 289-300.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *2*(4), 235-256.

Jara, D., Riascos, A., & Romero, M. (2010). Detección de copia en pruebas del estado. *Documento CEDE*, *15*.

Bibliography II

Sotaridona, L., van der Linden, W., & Meijer, R. (2006). Detecting answer copying uisng kappa statistic. *Applied Psychology Measurement*, 412-431.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, *3*, 295-312.

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*(7), 909-921.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, *21*(4), 307-320.

Bibliography III

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, *40*, 189-205.