UNIVERSIDAD
COMPLUTENSE
MADRID

Proyectos de Innovación

Convocatoria 2016/2017

Nº de Proyecto: **184**

Título del Proyecto

Desarrollo de una herramienta para la evaluación de los examen tipo test y sus aplicaciones en la mejora de la calidad y en los criterios de evaluación de estos exámenes

Nombre del responsable del proyecto

Diego García Pinto

Centro

Facultad de Medicina

Departamento

Departamento de Radiología, Rehabilitación y Fisioterapia

1. OBJETIVOS PROPUESTOS EN LA PRESENTACIÓN DEL PROYECTO

Los objetivos del presente proyecto son varios:

El principal objetivo es el desarrollo de una herramienta multiplataforma que permita el análisis de ítems de exámenes tipo test y que pueda ser utilizada por todo el personal docente. El software utilizará como entrada los resultados de la corrección de un test en un formato sencillo (CVS) que se haya obtenido de forma manual o de manera automática, mediante un software de reconocimiento de caracteres y se obtendrán los índices de dificultad, discriminación, homogeneidad, fiabilidad y validez. Se pretende que dicho software se de uso fácil e intuitivo y configurable. La información de salida podrá ser importada en cualquier hoja de cálculo (Excel, google docs, etc.).

Otro de los objetivos es el estudio del impacto que tiene en la calidad de la elaboración de las preguntas de un examen tipo test el uso de la información proporcionada por el software desarrollado. Para ello se pretende analizar los resultados de exámenes de cursos anteriores para elaborar un modelo de examen siguiendo criterios óptimos de los índices calculados. Este examen se propondrá como examen intermedio a los alumnos de 1º grado de Medicina, en la asignatura de Física Médica. Posteriormente se evaluarán los resultados del examen pudiendo detectar áreas que han sido más difíciles de asimilar, preguntas cuyos resultados no han sido los esperados, etc., de modo que esta información será utilizada como "*feedback*" hacia los alumnos. Finalmente se volverá a analizar los resultados del examen al finalizar el curso para evaluar la posible mejora en la consecución de los objetivos docentes propuestos en la asignatura.

También otro de los objetivos del proyecto será establecer pautas relativas al criterio de valoración numérica de cada una de las preguntas incluidas en el examen tipo test, con la finalidad de objetivar el resultado de dicho examen, en otras palabras, escalar a los alumnos según sus conocimientos. Para ello se deberán tener en cuenta varios aspectos que tendrán su repercusión numérica pudiéndose proponer distintos criterios de valoración. Estos criterios se aplicarán en la calificación de la asignatura de Bioestadística.

Así mismo, cómo prueba de validación intra-alumnos de los tests llevados a cabo se analizará la correlación entre los resultados de los exámenes de Física Médica y Bioestadística.

2. OBJETIVOS ALCANZADOS

El principal objetivo del presente proyecto era la creación de una herramienta para el análisis de ítems de exámenes tipo test de multi-respuesta.

Como se pretendía que fuese multi-plataforma la versión final se ha implementado como una hoja de cálculo (en formato Excel y google docs) que permite su ejecución en cualquier versión de sistema operativo. Para ello se han desarrollado una serie de macros (conjunto de instrucciones que realizan una determinada tarea) que permiten realizar todas las operaciones necesarias para obtener los índices de dificultad y discriminación (ANEXO I) de cada una de las preguntas, ambos indicadores relacionados con la calidad del examen. Así mismo, permite comprobar la distribución de las respuestas de cada una de las preguntas permitiendo detectar errores en la elaboración del examen.

La herramienta, además, permite obtener la calificación de cada uno de los alumnos que han realizado el examen y es configurable, posibilitando, por ejemplo, obtener la nota con el total de preguntas o por bloques si se desea evaluar determinadas áreas o también, modificar el peso de cada una de las preguntas y su correspondiente valor (se puede incluir penalizaciones por cada respuesta no acertada). El usuario ha de introducir la cadena con las respuestas correctas en la que es posible definir más de una posible respuesta correcta para una misma pregunta.

El software utiliza como datos de entrada los resultados de la corrección del conjunto de exámenes realizados por los alumnos en formato CVS, con una fila por cada alumno. Este es el formato de salida más común utilizado por los distintos programas OMR (*optical mark recognition*) es decir programas de reconocimiento de marcas empleados en la corrección automática de exámenes.

El resultado del análisis de las preguntas se puede exportar como hoja de cálculo si se desea para un posterior tratamiento de los datos.

Otro de los objetivos del proyecto es estudiar el impacto que tiene en la calidad de la elaboración de las preguntas de un examen tipo test el uso de la información proporcionada por el software desarrollado.

Utilizando la herramienta se analizaron exámenes de cursos anteriores y la información obtenida se utilizó en la elaboración de un examen intermedio de carácter voluntario. Con el análisis de los resultados de este examen y atendiendo a los resultados previos, se elaboró el examen final de la asignatura propuesto en diciembre de 2017. Aunque los resultados no son concluyentes ya que ambos exámenes (intermedio y final) no corresponden a la misma cantidad de materia si se ha observado cierta mejoría (ANEXO II) en los resultados obtenidos por aquellos alumnos que obtuvieron peores calificaciones en el primer examen. Además, la metodología se ha mostrado ser muy útil para la elaboración de este tipo de pruebas.

Además, otro de los objetivos era el establecer pautas relativas al criterio de valoración numérica de cada una de las preguntas incluidas en el test. Aunque no se han analizado en profundidad los criterios de valoración numérica y su influencia en la nota del alumno, se han comparado dos modos distintos de puntuar las preguntas para dos exámenes distintos, los correspondientes a la asignatura de Física Médica y el de la asignatura de Bioestadística ambas pertenecientes a 1º de Grado de Medicina. Si bien el resultado no es concluyente, ya que existen muchas variables a tener en cuenta que podrían afectar el resultado (distinta materia, elaboración de las preguntas y los distractores, etc.) se ha podido comprobar que no existen diferencias significativas entre las puntuaciones medias y medianas de ambas asignaturas (ANEXO III).

3. METODOLOGÍA EMPLEADA EN EL PROYECTO

La metodología empleada durante el desarrollo del presente proyecto puede dividirse en tres fases, correspondientes a los distintos objetivos planteados en el proyecto.

1ª Fase: Desarrollo del software.

En esta primera fase se ha llevado a cabo la implementación de todo el código necesario para la obtención de los índices de dificultad y discriminación utilizando como entrada el resultado de un examen tipo test. Para su cálculo hay que evaluar el número de respuestas correctas e incorrectas de cada una de las preguntas, así como la distribución en cuartiles de las notas de los alumnos.

El código ha sido creado para que pueda ser ejecutado en una hoja de cálculo (Excel o Google docs) y atendiendo a las necesidades que pudieran aparecer a la hora de evaluar los exámenes.

Posteriormente se utilizó el software en el análisis de varios exámenes para su validación.

2ª Fase: Estudio del impacto en la calidad el uso de la información proporcionada por el software

Haciendo uso de la herramienta desarrollada se han analizado los resultados de las preguntas de exámenes realizados en los cursos 2013-2014, 2014-2015 y 2015-2016, de la asignatura de Física Médica de 1º de Grado de Medicina. Dicho análisis nos ha permitido clasificar las preguntas en relación con su grado de dificultad y discriminación, así como detectar preguntas formuladas erróneamente y elaborar una librería de cuestiones.

Con esta librería de preguntas se ha elaborado un examen atendiendo a los índices de dificultad y discriminación. El examen fue propuesto en octubre de 2017 para que lo realizaran los alumnos de forma voluntaria.

Posteriormente se analizaron los resultados del examen intermedio para detectar qué preguntas resultaron más difíciles, así como descartar posibles errores. Esta información se utilizó durante las clases de seminarios para recopilar la información de los propios alumnos e identificar qué aspectos hacían que las preguntas fueran difíciles. Gracias a ello se pudo detectar los conceptos de la asignatura que necesitaban más atención y a su vez, cómo algunos enunciados podían resultar confusos para los alumnos.

La información proporcionada por los alumnos se tuvo en cuenta a la hora de la elaboración del examen final de la asignatura realizado en diciembre de 2017.

Finalmente se analizó la repercusión del método en las notas obtenidas por los alumnos.

3ª Fase: Criterio de valoración numérica de cada una de las preguntas incluidas en el examen tipo test

Con el fin de analizar la repercusión que tendría sobre la nota del alumno los distintos criterios de valoración numérica de las preguntas de un examen, se han analizado los resultados obtenidos por los mismos alumnos en dos pruebas de tipo test

correspondientes a las asignaturas de Física Médica y Bioestadística, ambas de 1º de Grado de Medicina.

En la asignatura de Física Médica los errores no puntúan negativamente (se compensa el azar aumentando el número de preguntas acertadas para aprobar), mientras que en la asignatura de Bioestadística penaliza con -1/3. El hecho de que una respuesta errónea reste podría implicar que el alumno no responda, aun sabiendo la respuesta correcta, por no arriesgarse a una puntuación negativa.

Se han analizado las posibles diferencias entre ambas notas por medio de la correlación de Pearson y Spearman. También se han comparado las medias y medianas de datos pareados de los tests paramétrico (t de Student) y no paramétrico (pruebas de Wilcoxon y signos). Los resultados de los tres tests muestran que no existen diferencias significativas entre las puntuaciones medias y medianas de ambas.

4. RECURSOS HUMANOS

Se ha contado con todos los profesores que forman la unidad de Física Médica del Departamento de Radiología y Medicina Física de la facultad de Medicina UCM: Eduardo Guibelalde, Gabriel Prieto, Víctor Delgado, Eliseo Vañó, José Miguel Fernández, Carlos Prieto, Alfonso López, Margarita Chevalier y Diego García Pinto (responsable del proyecto). Su contribución ha consistido en:

- Validación de la herramienta creada.
- Elaboración de las preguntas para la creación de los exámenes tipo test.
- Análisis de los resultados.
- Discusión con los alumnos de los resultados durante las sesiones de seminario.

Además, han contribuido al desarrollo del proyecto, Antonia García Salinero (PAS) y María Castillo García (estudiante de doctorado) en la corrección de los exámenes y clasificación de los mismos.

También ha participado Agustín Turrero Nogues, Prof. de la Sección Departamental de Estadística e Investigación Operativa de la Facultad de Medicina UCM, encargado del estudio comparativo de los resultados obtenidos en ambas asignaturas.

5. DESARROLLO DE LAS ACTIVIDADES

Septiembre-Octubre: Desarrollo de todo el código necesario para la creación del software para el análisis de los exámenes.

Octubre: Análisis de los exámenes de la asignatura de Física Médica de los cursos 2013-2014, 2014-2015, 2015-2016. Elaboración de una base de datos de preguntas utilizadas anteriormente atendiendo a criterios de dificultad y discriminación.

Octubre: Realización del examen intermedio por parte de los alumnos de 1º de Grado de Medicina.

Noviembre: Discusión de los resultados con los alumnos en las distintas sesiones de seminario.

Diciembre: Realización del examen final de la asignatura de Física Médica y análisis de los resultados.

Junio: Realización del examen final de la asignatura de Bioestadística. Análisis de los resultados.

6. ANEXOS

ANEXO I. Cálculo de los índices

Índice de dificultad: Fracción de estudiantes que han acertado la respuesta, esto es número de aciertos/ número de estudiantes. Valores bajos de este índice implican un alto grado de dificultad. Un buen ejemplo de examen sería aquel que tenga gran variedad de valores de dificultad.

Índice de discriminación: Diferencia entre el valor medio de aciertos de los alumnos del 1er cuartil y el valor medio de aciertos de los alumnos del 3er cuartil. Un valor alto de discriminación (el máximo es 1) implica que los alumnos que mejor han realizado el examen han respondido correctamente, por el contrario, un valor bajo (o incluso negativo) implica que los alumnos que han realizado mal el examen han respondido bien a la pregunta. Sería recomendable que la pregunta tuviese un valor de discriminación de al menos > 0,20.

	ID	PUNTUACION	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24
	CLAVE	22	A	D	A	B	B	A	C	D	B	C	D	B	D	B	C	C	A	D	B	D	C	C		
Ex_Int_1B	NXXXXXXXXX	19	A	D	A	B	B	A	C	D	B	C	D	A	D	B	D	C	A	D	C	D	C	C		
Ex_Int_1B	NXXXXXXXXX	17	B	D	A	B	B	A	A	D	B	C	D	B	D	B	B	C	B	D	C	D	C	C		
Ex_Int_1B	NXXXXXXXXX	20	B	D	A	B	B	A	C	D	B	C	D	B	D	B	B	C	A	D	B	D	C	C		
Ex_Int_1B	NXXXXXXXXX	17	B	D	A	B	A	A	D	D	B	B	D	B	D	B	C	C	A	D	C	D	C	C		
Ex_Int_1B	NXXXXXXXXX	15	C	D	A	B	B	A	C	D	B	B	D	B	A	C	D	C	D	D	C	D	C	C		
Ex_Int_1B	NXXXXXXXXX	11	C	D	A	B	B	D	A	A	B	B	A	A	C	B	D	D	A	D	B	C	C	C		
Ex_Int_1B	NXXXXXXXXX	19	B	D	A	B	B	A	C	D	B	C	D	B	D	B	C	B	A	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	13	B	D	A	B	B	D	A	D	B	C	A	B	C	B	D	B	A	D	B	C	A	C		
Ex_Int_1B	NXXXXXXXXX	9	B	D	A	A	C	D	D	D	D	C	A	D	C	C	D	C	C	D	C	D	C	C		
Ex_Int_1B	NXXXXXXXXX	10	A	D	A	B	B	D	A	B	B	B	C	B	C	D	D	D	A	D	C	B	C	D		
Ex_Int_1B	NXXXXXXXXX	18	A	D	A	A	B	A	C	D	B	C	D	B	D	B	D	C	C	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	17	A	D	A	B	B	A	D	D	B	C	A	B	D	B	B	D	A	D	D	D	C	C		
Ex_Int_1B	NXXXXXXXXX	20	B	D	A	B	B	A	C	D	B	C	D	B	D	B	C	C	D	D	B	D	C	C		
Ex_Int_1B	NXXXXXXXXX	14	B	D	A	B	A	A	C	D	B	D	D	A	C	B	C	C	B	D	D	C	C	C		
Ex_Int_1B	NXXXXXXXXX	15	B	D	A	B	B	A	C	D	B	B	D	B	C	B	D	D	B	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	13	B	D	A	B	B	A	D	D	B	C	B	A	D	B	A	D	C	D	D	D	B	C		
Ex_Int_1B	NXXXXXXXXX	12	B	A	A	B	A	D	C	A	B	C	A	A	C	B	D	C	A	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	17	B	D	A	B	B	A	C	D	B	C	D	B	D	B	B	A	A	D	B	C	C	B		
Ex_Int_1B	NXXXXXXXXX	11	B	D	A	B	B	A	C	A	C	C	B	A	D	D	A	C	D	B	A	B	C	C		
Ex_Int_1B	NXXXXXXXXX	11	B	D	A	B	B	D	C	D	B	D	A	A	D	A	C	B	C	D	C	C	A	C		
Ex_Int_1B	NXXXXXXXXX	14	B	D	A	B	B	D	C	D	B	C	D	B	C	B	B	A	D	D	C	A	C	C		
Ex_Int_1B	NXXXXXXXXX	21	B	D	A	B	B	A	C	D	B	C	D	B	D	B	C	C	A	D	B	D	C	C		
Ex_Int_1B	NXXXXXXXXX	13	B	D	B	B	B	D	C	D	B	D	D	B	D	B	D	C	A	C	B	C	A	A		
Ex_Int_1B	NXXXXXXXXX	19	A	D	A	B	B	A	D	D	B	C	D	B	D	B	A	C	B	D	B	D	C	C		
Ex_Int_1B	NXXXXXXXXX	14	A	D	A	B	B	A	B	D	B	D	D	A	D	B	A	D	C	D	B	B	C	D		
Ex_Int_1B	NXXXXXXXXX	13	B	D	A	B	B	A	A	A	B	C	D	C	A	B	B	C	B	D	C	B	C	C		
Ex_Int_1B	NXXXXXXXXX	0	B	X	X	D	X	X	A	X	X	X	B	X	X	B	X	X	A	X	X	C	X	X	A	
Ex_Int_1B	NXXXXXXXXX	12	B	D	A	B	B	D	C	D	B	C	C	B	A	C	D	C	D	D	D	B	A	C		
Ex_Int_1B	NXXXXXXXXX	12	D	D	A	B	B	A	D	D	B	B	C	A	D	A	B	C	C	D	A	D	A	C		
Ex_Int_1B	NXXXXXXXXX	17	B	D	A	B	B	A	C	A	B	C	D	B	D	B	B	C	B	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	3	X	X	D	B	C	D	B	B	A	B	B	X	A	A	B	B	D	D	C	B	A	C		
Ex_Int_1B	NXXXXXXXXX	13	A	D	B	B	A	A	C	A	B	C	A	C	D	D	C	C	A	D	C	C	B	C		
Ex_Int_1B	NXXXXXXXXX	13	B	D	A	B	B	D	C	C	C	C	A	A	D	B	B	C	A	D	D	C	C	C		
Ex_Int_1B	NXXXXXXXXX	15	B	D	A	B	B	A	D	D	C	C	D	B	C	A	B	C	A	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	17	B	D	A	B	B	A	C	D	B	C	D	B	A	B	D	C	B	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	15	B	D	A	B	B	A	D	D	B	C	D	B	D	B	B	D	A	D	D	C	B	C		
Ex_Int_1B	NXXXXXXXXX	17	B	D	A	B	B	A	C	D	B	B	D	B	D	B	A	C	A	D	D	A	C	C		
Ex_Int_1B	NXXXXXXXXX	19	A	D	A	B	B	A	C	D	B	C	D	B	D	B	B	C	B	D	B	C	C	C		
Ex_Int_1B	NXXXXXXXXX	9	B	D	A	A	B	A	A	D	C	D	A	A	C	C	B	C	D	D	B	D	B	D		
Ex_Int_1B	NXXXXXXXXX	15	X	D	A	B	B	A	C	D	B	B	D	B	D	A	D	C	C	D	A	D	A	C		
Ex_Int_1B	NXXXXXXXXX	14	B	A	A	B	B	A	C	D	B	C	D	D	A	B	B	D	B	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	19	A	D	A	B	B	A	C	D	B	C	D	C	D	B	C	C	C	D	A	D	C	C		
Ex_Int_1B	NXXXXXXXXX	11	B	D	A	B	B	A	D	A	B	B	C	B	D	B	A	B	B	D	D	C	D	C		
Ex_Int_1B	NXXXXXXXXX	12	X	D	A	A	B	A	D	C	C	B	D	B	D	C	A	C	A	D	B	B	C	A		
Ex_Int_1B	NXXXXXXXXX	16	B	D	A	B	A	A	D	D	B	C	D	B	A	B	B	C	C	D	B	D	C	C		
Ex_Int_1B	NXXXXXXXXX	14	B	D	A	B	B	A	C	A	B	C	D	C	D	B	B	D	B	D	C	D	D	C		
	Indice de dificultad		0,2	0,91	0,91	0,89	0,83	0,74	0,57	0,72	0,83	0,63	0,63	0,59	0,61	0,7	0,17	0,63	0,39	0,93	0,28	0,54	0,7	0,85		
	No. de Estudiantes	46																								
	Q1	12																								
	Q2	14																								
	Q3	17																								
	Discriminacion		0,3	0,23	0,15	0,25	0,26	0,62	0,44	0,56	0,46	0,57	0,86	0,57	0,56	0,77	0,23	0,27	0,19	0,15	0,15	0,5	0,54	0,25		
	Distribucion Respuestas:																									
		A	9	2	42	4	5	38	7	8	1	0	9	11	7	5	7	3	16	0	12	2	7	3		
		B	31	0	2	41	38	0	2	2	38	12	3	27	1	33	17	5	11	1	15	7	4	1		
		C	2	0	0	0	2	0	34	2	5	29	4	4	10	5	8	29	9	1	13	11	37	39		
		D	1	43	1	1	0	11	11	33	1	5	29	2	29	3	13	9	7	43	6	25	2	3		
		Blancos	3	2	1	0	1	1	0	1	1	0	1	2	0	1	1	0	1	1	0	1	1	0		

Figura 1: Ejemplo del resultado del análisis de uno de los exámenes evaluados en el presente proyecto.

ANEXO II. Comparación de los resultados entre los exámenes intermedio y final de la asignatura de Física Médica

Con el Objetivo de analizar el impacto que tiene en la calidad del examen se han comparado los resultados obtenidos entre los exámenes intermedio y final. Como se puede apreciar en la figura 2, se aprecia una mejora en el resultado del examen para aquellos alumnos que pero hicieron la prueba intermedia.

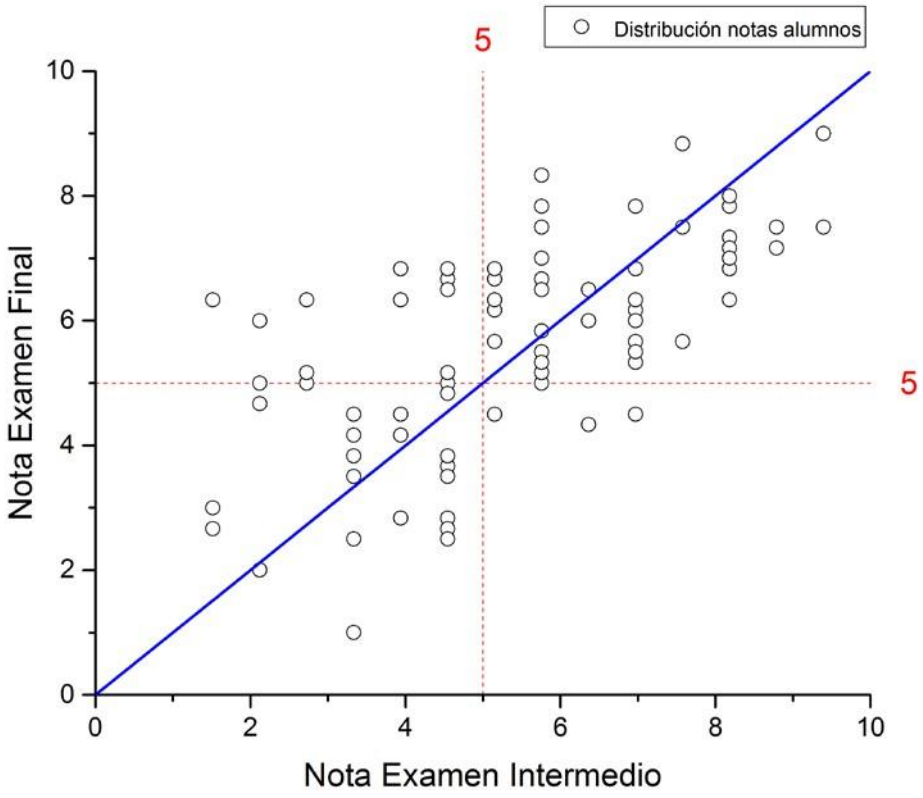

Figura 2. Distribución de notas de los alumnos obtenidas en los dos exámenes

ANEXO III. Comparación de los exámenes de las asignaturas de Física Médica y Bioestadística.

A continuación se muestran los resultados de la comparación estadística de los exámenes de las dos asignaturas.

Estadísticos

		Estadística	Física
N	Válido	64	64
	Perdidos	0	0
Media		6,627	6,255
Mediana		7,300	6,300
Desviación estándar		2,1135	206
Varianza		4,467	2,018
Asimetría		-,432	-,568
Error estándar de asimetría		,299	,299
Rango		8,5	7,3
Mínimo		1,5	2,2
Máximo		10,0	9,5
Percentiles	25	4,800	5,800
	50	7,300	6,300
	75	8,300	7,075

Las notas medias son 6,627 y 6,255 para Estadística y Física respectivamente. Las desviaciones típicas son 2,113 y 1,421 para Estadística y Física, mostrando la mayor heterogeneidad de las notas la asignatura de Estadística. Por esta última razón el percentil 25 o primer cuartil es inferior en Estadística, el 75% de los alumnos obtiene más de 4,8 en Estadística y más de 5,8 en Física. La Mediana (percentil 50) y el percentil 75 muestran notas más altas en Estadística, 7,3 frente a 6,3 y 8,3 frente a 7,07 respectivamente.

Finalmente, se incluyen los resultados de las comparaciones de medias y medianas de datos pareados de los tests paramétrico (t de Student) y no paramétrico (pruebas de Wilcoxon y signos).

Prueba t de muestras emparejadas

		Diferencias emparejadas			t	gl	Sig. (bilateral)
		Media	Desviación estándar	Media de error estándar			
Par 1	Estadística - Física	,3719	2,0255	,2532	1,469	63	**,147**

Prueba de Wilcoxon de los rangos con signo

Estadísticos de prueba[a]

	Física - Estadística
Z	-1,407[b]
Sig. asintótica (bilateral)	**,159**
Significación exacta (bilateral)	**,161**
Significación exacta (unilateral)	,080
Probabilidad en el punto	,001

a. Prueba de Wilcoxon de los rangos con signo

b. Se basa en rangos positivos.

Prueba de los signos

Estadísticos de prueba[a]

	Física - Estadística
Z	-,756
Sig. asintótica (bilateral)	**,450**
Significación exacta (bilateral)	**,450**
Significación exacta (unilateral)	,225
Probabilidad en el punto	,068

a. Prueba de los signos

Los resultados de los tres tests muestran que no existen diferencias significativas entre las puntuaciones medias y medianas de ambas asignaturas (los valores de los diferentes p-valores, exactos y aproximados, se destacan en azul en las tres tablas precedentes).