

# Estadística para lingüistas: cómo realizar e interpretar resultados estadísticos para mejorar tus pruebas.

Dra. Julia Zabala Delgado  
Centro de Lenguas. Universitat Politècnica de València  
juzadel@upv.es

## 1. Introducción

Usos de la estadística para

## 2. Tipos de análisis

Teoría clásica del test y t

## 3. TCT: Análisis a nivel de la prueba

Estadística descriptiva y

Fiabilidad de la prueba: A

*Ejercicios prácticos: ¿qué*

## 4. TCT: Análisis a nivel de ítems

Cálculo e interpretación d

análisis de descriptores

*Ejercicios prácticos: ¿qué*

## 5. Teoría de respuesta al ítem: Mapas de

*Ejercicios prácticos: inter*

## 6. Conclusiones, práctica y dudas

¿Qué queremos obtener mediante el uso de la estadística? ¿Cuál es el objetivo de este taller?

**Desmitificar los análisis estadísticos** y empezar a verlos como medio para conseguir la mayor información posible sobre la competencia de nuestros alumnos de una forma:

- ✓ Práctica
- ✓ Eficiente
- ✓ Fiable

Para:

- ✓ Mejorar nuestra docencia
- ✓ Conocer las necesidades y comportamientos de los alumnos
- ✓ Evaluar nuestro programas
- ✓ **La estadística es una herramienta muy potente para obtener datos pero no sustituye el conocimiento experto del lingüista, lo complementa.**

¿Quiénes  
sois?

¿Dónde  
trabajáis?



¿cuál es  
vuestra  
experiencia  
con la  
estadística?

¿Qué  
esperáis  
obtener de  
este taller?



# Usos de la estadística en la evaluación

## Validez



## Fiabilidad



## Practicidad



## Autenticidad



## Efecto colateral



## Impacto





# Tipos de análisis



## TEORIA CLÁSICA DEL TEST

### VENTAJAS

Descriptiva

Accesible

Requiere un menor número de candidatos  
>40 (idealmente >100)

### DESVENTAJAS

Dependiente de la muestra

No permite generalizar para poblaciones  
diversas

Software: SPSS, TiaPlus, TAP, Excel, CITAS

## TEORIA DE RESPUESTA AL ÍTEM

### VENTAJAS

Independiente de la muestra

Nos permite generalizar los resultados

Imprescindible para anclar ítems

### DESVENTAJAS

Menos accesible

Requiere un mayor número de candidatos  
(>200)

Software: Winsteps, Facets, R



# ¿De qué depende la elección del tipo de análisis?



**OBJETIVO DEL ANÁLISIS**

**PERFIL DE LA POBLACIÓN DEL PILOTAJE O ANÁLISIS**

**NÚMERO DE CANDIDATOS**

**RECURSOS HUMANOS Y ECONÓMICOS**



# 1

## TCT: Análisis a nivel de la prueba



Análisis a nivel de prueba

- Media
- Mediana
- Moda
- Desviación estándar

Alpha de Cronbach

Histogramas





## Análisis a nivel de prueba

- **Media:** es la respuesta “media” de los candidatos. Se calcula como el número de respuestas correctas por el número de preguntas.
- **Desviación estándar:** es una medida de dispersión que indica cuánto se alejan los resultados de la media.
- **Mediana:** es el centro de las puntuaciones ordenadas.
- **Moda:** es la puntuación que más veces se repite (puede haber varias modas)

Distribución normal: media, mediana y moda se acercan o son iguales. Curtosis y coeficiente de asimetría entre -2 y 2 (-1 y 1 siendo estrictos)

Buscamos una media cercana a la media de nuestra prueba ¿Por qué? Porque indica si nuestro examen está bien dirigido a nuestra población. Idealmente buscaríamos que no hubiera resultados ni de 0% ni de 100% de aciertos para evitar el efecto suelo y el efecto techo (queremos medir a todos los candidatos, a los menos hábiles y a los más hábiles)

## Tarea 1. Análisis a nivel de prueba. Descriptivos e histogramas



Representación de resultados  
Abrid archivo Tarea 1.xlsx

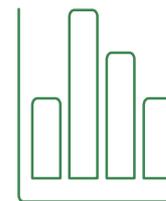


Estadística descriptiva de la prueba  
Abrid archivo Tarea 1.xlsx



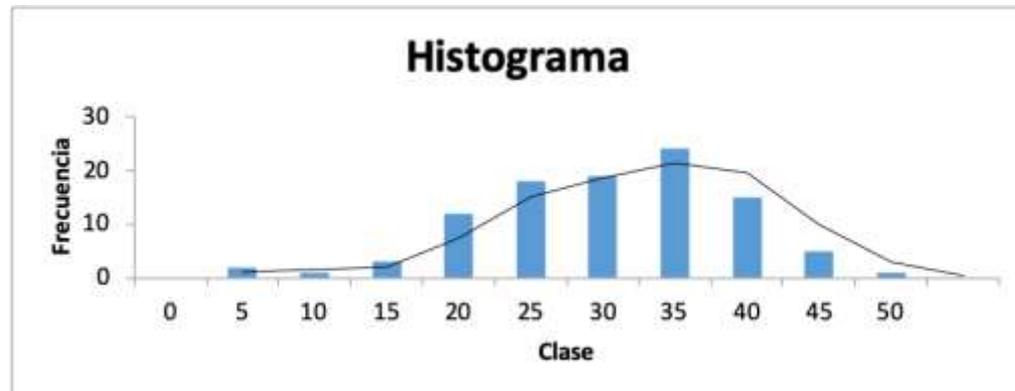
Realización de un histograma  
Abrid archivo Tarea 1\_solucion paso 1.xlsx

Resumen de resultados  
Tarea 1 \_solución paso 2Histograma.xlsx



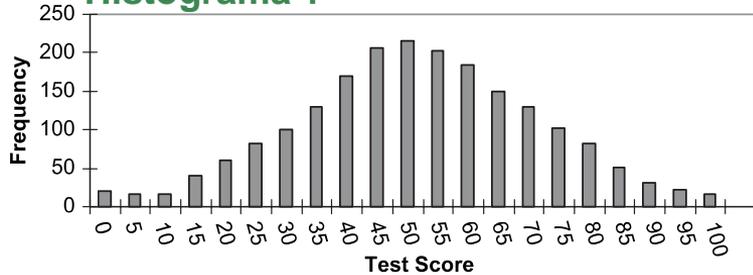


Media	28,43
Error típico	0,87
Mediana	29,00
Moda	35,00
Desviación estándar	8,70
Varianza de la muestra	75,62
Curtosis	0,21
Coefficiente de asimetría	-0,44
Rango	45,00
Mínimo	4,00
Máximo	49,00
Suma	2843,00
Cuenta	100,00
Nivel de confianza(95,0%)	1,73



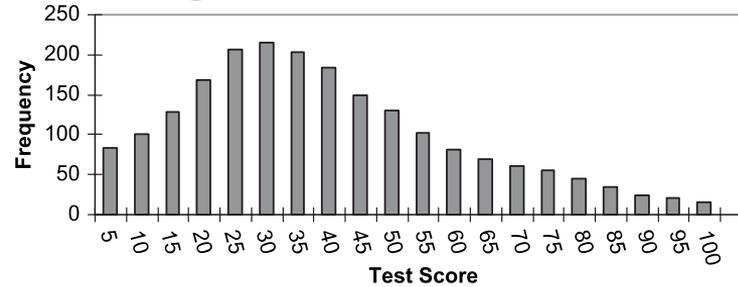


### Histograma 1



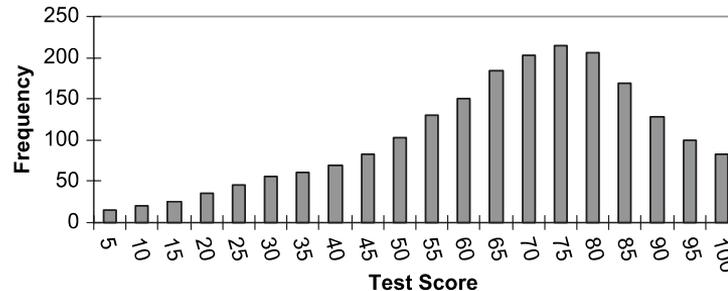
**Distribución normal:** deseable para pruebas de certificación

### Histograma 3



**Sesgada a la derecha:** test difícil para los candidatos, característico de una prueba con referencia a un criterio (a principio de curso)

### Histograma 2



**Sesgada a la izquierda:** test fácil para los candidatos, deseable para una prueba de aprovechamiento

Ejemplos de Carr (2008)

Julia Zabala Delgado



# Análisis a nivel de prueba. Alfa de Cronbach



**La fiabilidad de la prueba hace referencia a la cantidad de error que afecta a sus resultados**

**Fiabilidad (Alfa de Cronbach):** mide la consistencia interna de la prueba en un rango de 0 a 1,00.

A partir de 0,7 se consideraría aceptable para una prueba, aunque para pruebas de certificación se prefiere  $>0,8$ .

¿Por qué? un alfa de .80 indica que el 80% de la variabilidad es sin error, y solo hay un 20% debido a error de medida

Alfa de Cronbach	Consistencia interna
$\alpha \geq 0.9$	Excelente
$0.9 > \alpha \geq 0.8$	Buena
$0.8 > \alpha \geq 0.7$	Aceptable
$0.7 > \alpha \geq 0.6$	Pobre

¿Que afecta al alfa de Cronbach?

- El número de ítems (cuantos más ítems, más fiabilidad)
- El índice de discriminación de los ítems (cuanto más discriminen, mas podemos fiarnos de la medición)
- El nivel de dificultad de los ítems (los ítems muy fáciles o muy difíciles nos darán poca variabilidad)
- La homogeneidad de los ítems (a más homogeneidad, mayor fiabilidad)

(Adaptado de Green, 2013)



# Análisis a nivel de prueba. Alfa de Cronbach

Tarea 1. Análisis a nivel de prueba. Alfa de Cronbach



Calcular Alfa de Cronbach

Abrid archivo Tarea 2 \_ calcular alfa de Cronbach



Solución:  
Alfa de Cronbach=0,88

# 2

## TCT: Análisis a nivel de ítems

### Análisis a nivel de ítems o tareas

- Índice de facilidad del ítem
- Índice de discriminación del ítem
- Funcionamiento de distractores



## Análisis a nivel de ítem

- **Índice de facilidad del ítem:** porcentaje de alumnos que dan una respuesta. **i.e. 33 de 45 estudiantes contestaron**
- **Índice de discriminación del ítem:** capacidad del ítem para discriminar entre aquellos que son más y menos competentes
- **Funcionamiento de distractores:** nos permite saber

### IF

.30 -.70	Ítems óptimos
.20-.80	ítems buenos atendiendo a la discriminación (el índice de discriminación tiene que ser bueno) y a que aporten consistencia a la prueba

### ID

>40	óptima
.30-.39	Razonablemente buena
.20-29	Marginal (debe mejorarse el ítem)
<19	A mejorar o descartar

### Análisis de distractores

<7% (no funciona)

(puede usarse 10% con ítems de 4 opciones)



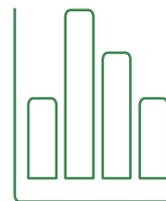
## Análisis a nivel de ítem



### Análisis a nivel de ítem – índices de facilidad

Abrid archivo Tarea 3\_Analisis a nivel de item IF.xlsx

*Solución: Abrid archivo Tarea 3\_Analisis a nivel de item IF\_solucion.xlsx*



### Análisis a nivel de ítem - índices de discriminación

Abrid archivo Tarea 3\_Analisis a nivel de item ID.xlsx

*Solución: Abrid archivo Tarea 3\_Analisis a nivel de item ID\_solucion.xlsx*





### **Consideraciones generales**

*Los ítems demasiado fáciles pueden no discriminar pero no necesariamente serán ítems problemáticos más allá de su facilidad.*

*Según el tipo de examen pueden considerarse aptos o no.*



### **Señales de alarma**

*Discriminación negativa*

*IF adecuado pero discriminación  $<.2$*

¿Qué ítems de los que hemos analizado presentan estas características?

## Ejercicio 1



Ítem 27

	IF	ID
Ítem 27	0,0	-0,1

Ítem 87

	IF	ID
Ítem 87	0,4	0,0

Ítem 103

	IF	ID
Ítem 103	0,7	0,1

We don't have \_\_\_\_\_ biscuits left, but I can offer you some cake.

- a) some
- b) any
- c) many

He's never \_\_\_\_\_ to London

- a) gone
- b) went
- c) been

I haven't got my purse! I must have forgotten \_\_\_\_\_ when we were in the restaurant.

- a) to pick it up
- b) picking it up
- c) having picked it up

## Ejercicio 2



Listen to an interview with a woman, Julia, who has been studying in Dublin.

Choose the answer A, B or C, which fits the questions (1-7).

The first one (0) has been done for you as an example.

Write your answers on the ANSWER SHEET provided.

0. Why did she end up going to Dublin?

It was where she'd always wanted to go.

There were fewer people from her country.

She wanted to go to a capital city.

Índice de facilidad	Índice de discriminación
0.9	0.11

Q3. What does she say about food in Dublin?

You can find healthy food but you pay more for it.

It's very difficult to find healthy food in Dublin.

You can ask local people where to find healthy food.

**La pregunta 3, tiene un índice de facilidad demasiado alto y además está por debajo de los niveles de discriminación aceptables.**

**En Pilotaje dio un nivel de dificultad de 0.7 y una discriminación de 0.2.**

**1. ¿A que puede deberse? Justifica tu respuesta**

**3. ¿Qué medidas tomarías si este fuera el resultado en uno de vuestros exámenes? Justifica tu respuesta**



### Posibles explicaciones:

En la prueba real suele mejorarse el rendimiento de los alumnos pues hay más en juego. Si una pregunta resulta fácil en el pilotaje es muy probable que el índice de facilidad aumente en la administración real.

En este caso la facilidad puede deberse al tema (comida y precios) que es algo sencillo para B2 y al formato de entrevista que pausa cada respuesta y permite al oyente centrarse en la pregunta.

### Decisión a tomar:

Si no detectamos ningún problema, puede mantenerse porque no afecta al candidato, pero tenemos que saber que es una pregunta inútil desde el punto de vista de la evaluación.

## Ejercicio 3



Q7. What does she recommend to other students going to Dublin?

- A. Try to be happy and the time will go quickly.
- B. Enjoy your short time there.
- C. Full year scholarships are better.

Q3. IF 0.5 ID 0.10

- A: 0.94
- B: 0.06
- C: 0

***¿Qué está ocurriendo aquí?***

***La pregunta 7 ofreció un índice de facilidad óptimo, pero no discriminó de manera acorde a éste. ¿Por qué?***



La pregunta parece del nivel apropiado pues se centra en opiniones dentro del ámbito B2. El juicio de expertos coincidió en que se trataba de un ejercicio B2. Sin embargo, las respuestas A y B son muy similares “be happy” is similar to “enjoy” y “quickly” y “short time” son conceptos afines. Por otro lado, el análisis de distractores indica que la respuesta C no fue elegida por nadie, lo que se entiende al leerla, pues el tema es ligeramente diferente y la redacción parece demasiado distinta a las opciones A y B.

**CONSECUENCIAS:** Una discriminación negativa debe eliminarse del cómputo de nota para todos los candidatos y una discriminación por debajo de .2 indica un ítem que debe modificarse. En este caso se modificó el ítem

Q7. What does she recommend to other students going to Dublin?

- A. Try to be happy and the time will go quickly.
- B. Take advantage of your short time there.
- C. Make sure you stay there for the full year.



## SAVE THE PANDA, SAVE THE WORLD

At this point, let's consider the argument from the other way around. Why shouldn't conservation organisations use the marketing power of pandas to further their worthy aims? What's so bad about cashing in on a beautiful bear if by saving it, we may save thousands of other species with whom it lives in the wild? For this reason, pandas and certain other 'charismatic megafauna' are often given special value, because they can act as 'umbrella species'. The argument goes that if these animals are protected, then so too are the other creatures that live within their shared habitats. Looking after a bear or a cheetah or tigers, for instance, protects everything else – the spiders, birds, lizards, mites and frogs – that may happen to share the same habitat, but that no one really cares enough about.

### Q10. What is an umbrella species?

Animals whose protection benefits other species.

Species that cover the small animals to keep them dry.

One that benefits from the conservation of other animals.

IF	ID
0.15	-0.17

A	B	C
0,7	0,15	0,16

La pregunta no 10 discriminó negativamente a pesar de haber funcionado tal y como indicó el pilotaje inicial. ¿Es quizás por el nivel de dificultad, la redacción? Si todo parece correcto. ¿Se elimina del cómputo? Observa los resultados del pilotaje.



Solución: Cuando se analiza el texto, la pregunta y los resultados de pilotaje y no hay explicación a unos malos resultados, muy frecuentemente la solución es la más sencilla. Un error en la clave de respuestas causó unos resultados de discriminación negativos.

Una vez corregidos, el resultado es bien diferente IF: 0.7 ID: 0.3

# 3

## Teoría de respuesta al ítem: Mapas de Wright



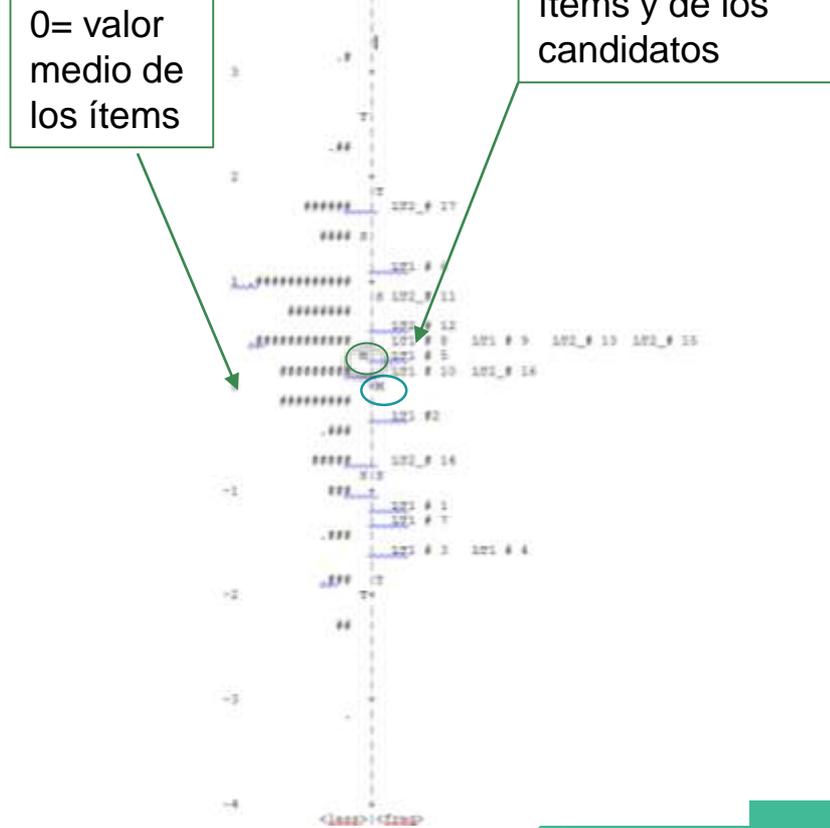
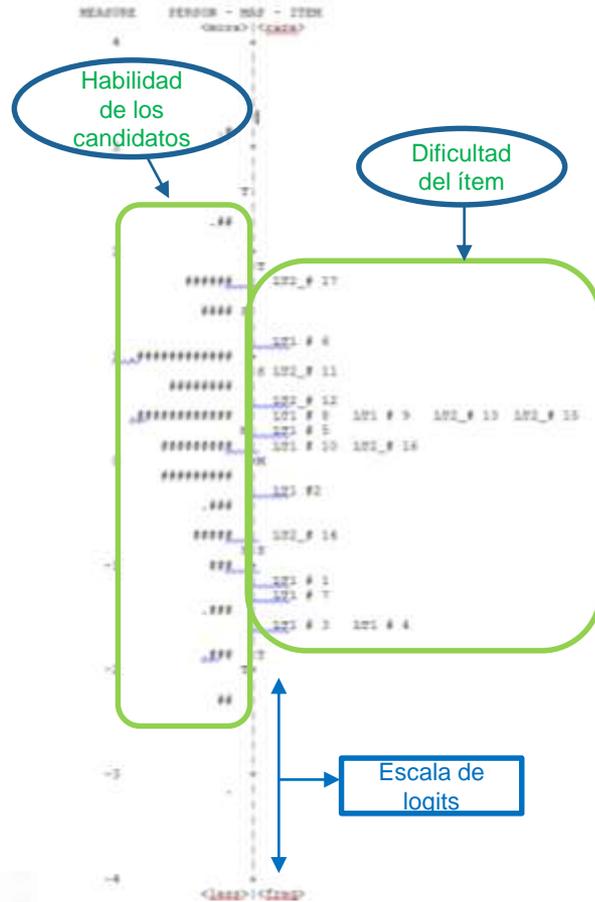
# Teoría de respuesta al ítem: modelo de Rasch



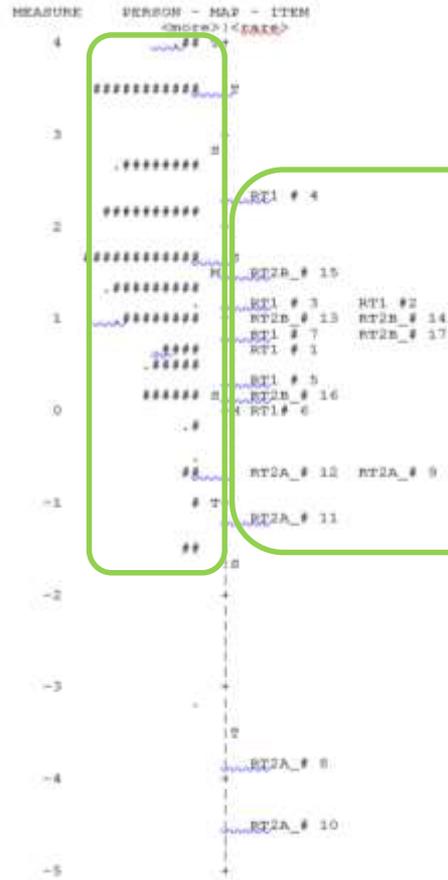
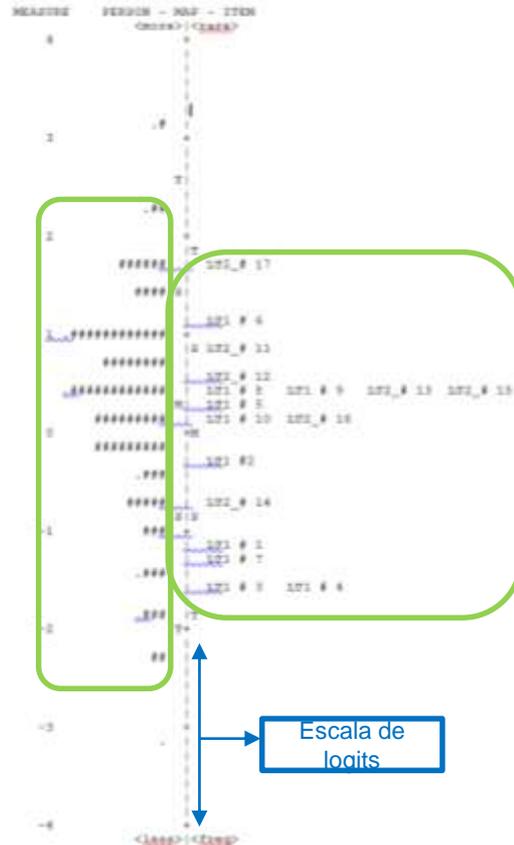
- EL modelo de Rasch mide la habilidad de los candidatos y la dificultad de los ítems de manera objetiva. La dificultad del ítem no dependerá de la habilidad del candidato y viceversa.
- El modelo de Rasch es uno de los modelos de teoría de respuesta al ítem que existen.
- Los mapas de Wright nos permiten ver el comportamiento de los ítems y los candidatos en una misma escala.
- La posición de los candidatos con respecto a los ítems nos indicara sus posibilidades de contestar correctamente a la pregunta.

-5,0	,01
-4,0	,02
-3,0	,05
-2,0	,12
-1,0	,27
0,0	,50
1,0	,73
2,0	,88
3,0	,95
4,0	,98
5,0	,99

# Representación gráfica: los mapas de Wright



## Ejercicio 1



Estos dos mapas de Wright representan el comportamiento de varios candidatos universitarios en un examen de comprensión auditiva y un examen de comprensión lectora.

¿Qué conclusiones sacas de lo apropiado de cada uno de los exámenes para esta población de candidatos?

¿Son todos los ítems útiles para medir la competencia de los candidatos?

¿La prueba ha podido medir la competencia de todos los candidatos?

¿Puedes comparar el comportamiento de los candidatos en la prueba de comprensión oral y en la prueba de comprensión escrita?



## ¿Qué más podemos aprender?

Análisis de encuestas

Correlaciones entre destrezas

Homologación de pruebas

Comportamientos de corrección

Establecimiento de estándares



¿Hemos conseguido desmitificar el uso de la estadística?

¿Vemos su utilidad en nuestro día a día?

“

*Conjeturar es barato; conjeturar  
erróneamente es caro.*

*Proverbio chino*



Julia Zabala Delgado  
*juzadel@upv.es*

**Estadística para lingüistas: cómo  
realizar e interpretar resultados  
estadísticos para mejorar tus pruebas.**