UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ANÁLISIS ESTADÍSTICOS PARA EL CONTROL DE CALIDAD DE LAS PRUEBAS DE CERTIFICACIÓN

Julia Zabala
juzadel@upv.es

CENTRO DE LENGUAS
Universitat Politècnica de València
www.upv.es

Contenidos del taller

Estadística para lingüistas
¿Qué buscamos obtener?
¿Qué herramientas tenemos?

Análisis estadístico de pruebas I

Análisis a nivel de prueba:
Representación de resultados
Estadística descriptiva
Histogramas
Calculo de fiabilidad: Alfa de Cronbach

Análisis estadístico de pruebas II
Análisis a nivel de ítems:
Índice de facilidad
Índice de discriminación

Fiabilidad en la corrección de tareas subjetivas
Fiabilidad inter e intra-corrector

¿Qué queremos obtener mediante el uso de la estadística? ¿Cuál es el objetivo de estas sesiones?

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.

¿Qué información podemos obtener??

Podemos averiguar si nuestra prueba es útil para nuestro propósito (prueba de nivel, de competencia, de aprovechamiento, etc.)

Podemos averiguar datos sobre el comportamiento de los alumnos

Podemos averiguar si nuestra prueba es consistente (fiable)

Podemos saber si una pregunta tiene el nivel de dificultad que queremos

Podemos detectar problemas con la redacción de las preguntas

Podemos ajustar la pregunta para asegurarnos de que discrimina

Podemos comparar los comportamientos de distintos grupos de alumnos

Podemos comparar nuestras correcciones con las de el estándar

Además, como profesionales de la evaluación, usamos la enseñanza para asegurar que nuestras pruebas cumplen con los principios de evaluación:

Practicidad – si un ítem funciona bien, obtendremos más información de la capacidad del candidato

Fiabilidad – si un ítem funciona bien de manera estable, medirá lo que queremos

Validez – una prueba fiable no siempre es válida, pero una prueba válida siempre tiene que ser fiable

Autenticidad – para calcular el difícil equilibrio entre la fiabilidad y la autenticidad

Efecto colateral – si nuestro examen mide lo que queremos, impactará positivamente en el aula y fuera de ella

Fundamentalmente …

Podemos sustentar nuestras decisiones sobre datos empíricos

Podemos aportar datos para un argumento de validez

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

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)

Herramientas disponibles

Teoría clásica: **SPSS, TiaPlus, TAP, Excel, CITAS**

Teoría de respuesta al ítem: **Winsteps, Facets, R**

Preparación de un pilotaje

Materiales:

Grupo de candidatos identificados

Tareas a pilotar en electrónico identificadas

Cuadernillo de tareas

Instrucciones para el profesor en el aula encargado del pilotaje

Cuestionario de opinión para el profesor en el aula encargado del pilotaje

Hoja Excel e instrucciones ¿introduciremos los datos nosotros? ¿El profesor?

Preparación de un pilotaje

Consideraciones:

Nuestro grupo de candidatos debe tener el mismo perfil

Si pilotamos en niveles limítrofes, obtendremos información más amplia

La administración del piloto debe ser lo más similar posible a la administración real de la prueba

Guardaremos el formato electrónico de nuestra prueba identificado y siempre con clave de respuestas en documento aparte

El cuadernillo impreso tendrá siempre el mismo formato

Tendremos en cuenta los tiempos (uniformes o no según objetivo)

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

La calidad de nuestro resultados dependerá de la calidad de los datos recabados

CENTRO DE LENGUAS
Universitat Politècnica de València
www.upv.es

Julia Zabala Delgado

Preparación de un pilotaje

Consideraciones:

Tendremos en cuenta la fatiga del candidato

Tendremos en cuenta la motivación del candidato, sobretodo en la interpretación de resultados

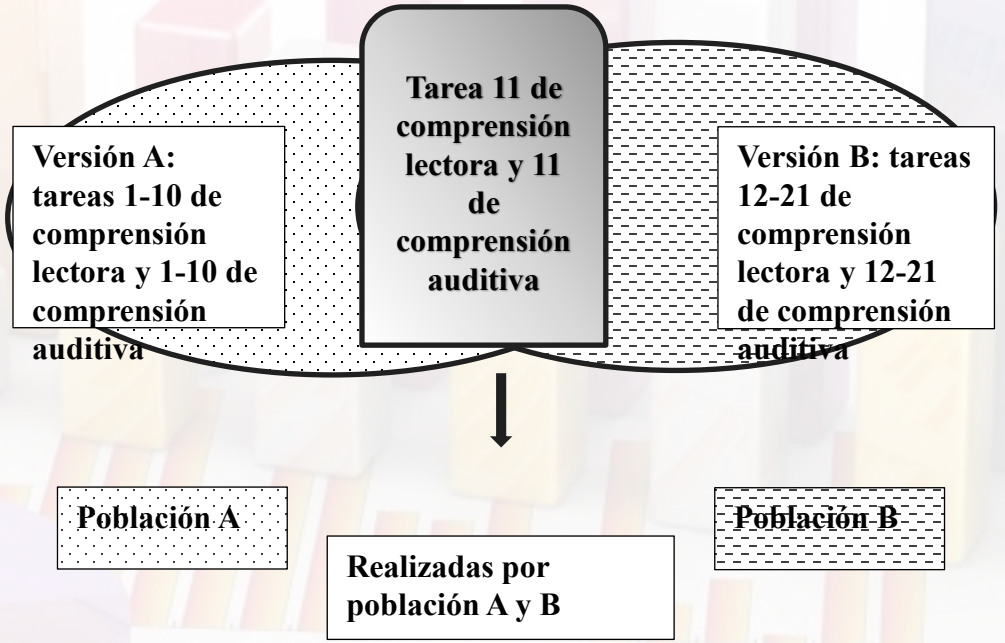
La seguridad de la prueba es clave

Planificaremos la corrección para asegurar la seguridad de la prueba

Guardaremos los resultados con el mismo identificador de la tarea y la clave de respuesta

Preparación de un pilotaje

Selección de la población



Julia Zabala Delgado

Preparación de un pilotaje

Datos cualitativos a recabar

Concepto a evaluar	Si/No	Consideraciones o acciones de mejora
¿Los candidatos han tenido tiempo suficiente para completar las tareas?		
¿Son el tipo de tareas familiares para los candidatos?		
¿Han comprendido los candidatos el mecanismo de la tarea?		
¿Existe en las tareas, o en los textos o temas de las tareas, algún sesgo que pueda beneficiar a unos candidatos sobre otros?		
¿Hay vocabulario o formulaciones ambiguas que pudieran dar lugar a error?		
En las tareas de destrezas productivas, ¿Es la tarea capaz de obtener del candidato un texto (oral o escrito) del nivel requerido?		
En las tareas de destrezas productivas, ¿Queda claro el objetivo de las tareas y el registro que deben utilizar?		
¿El nivel de dificultad de la tarea corresponde al nivel de los candidatos?		
¿Los candidatos han considerado las tareas demasiado fáciles o difíciles?		
¿El maquetado de los ejercicios (¿disposición en el folio, tamaño de letra, etc. resultaba apropiado?		
¿Los candidatos consideraron el tema interesante y/o motivador?		
¿La clave de respuestas era correcta e incluía todas las respuestas posibles?		

Julia Zabala Delgado

Preparación de un pilotaje

Organización de datos

Nombre de la tarea o de la prueba

Instrucciones de administración y recogida de datos cualitativos

Tarea formateada con datos identificativos

Clave de respuestas

Plantilla de análisis

Julia Zabala Delgado

Análisis a nivel de prueba:

Representación de resultados

Estadística descriptiva

Histogramas

Calculo de fiabilidad: Alfa de Cronbach

Julia Zabala Delgado

Conceptos básicos

Media: la media es la respuesta “media” de los candidatos a la prueba calculada dividiendo el numero de respuestas correctas por el número de estudiantes

Mediana: el centro de las puntuaciones ordenadas de menor a mayor

Moda: es la puntuación que más veces se repite (en una distribución puede haber varias modas)

Rango: nos indica la puntuación menor y la mayor en un grupo de notas. Útil para observar comportamiento de la población

Desviación estándar: la desviación estándar es una medida de dispersión de las notas de los candidatos en una prueba y nos indica como de diversa es nuestra población.

Curtosis: nos indica la densidad de la distribución de notas alrededor de la media. Valores positivos nos indicarán que hay muchos mas resultados cerca de la media que lejos de la media, mientras que valores negativos indicarán un mayor número de resultados alejados de la media.

Coeficiente de asimetría: nos indica hacia que lado se inclinan los resultados con respecto a la media. En el caso de una distribución normal “perfecta” el coeficiente de asimetría sería de 0. Valores negativos de este coeficiente indican que la población se inclina hacia la derecha, es decir hacia notas más altas. Valores positivos indican que la población se indica a la izquierda de la media, es decir, notas más bajas.

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Cómo lo representamos gráficamente? Histogramas

CENTRO DE LENGUAS
Universitat Politècnica de València
www.upv.es

Julia Zabala Delgado

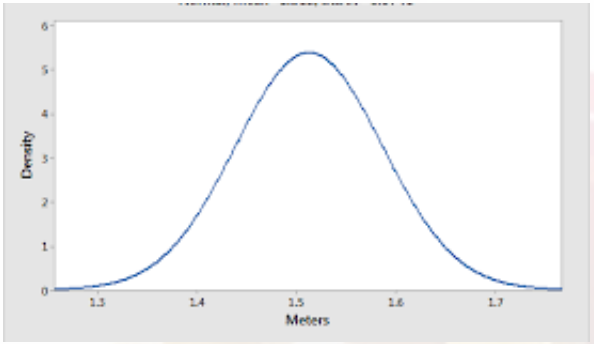
¿Qué buscamos lograr?

Buscamos una distribución normal:
media, mediana y moda se acerquen o sean iguales, y que la curtosis y coeficiente de asimetría este entre -2 y 2

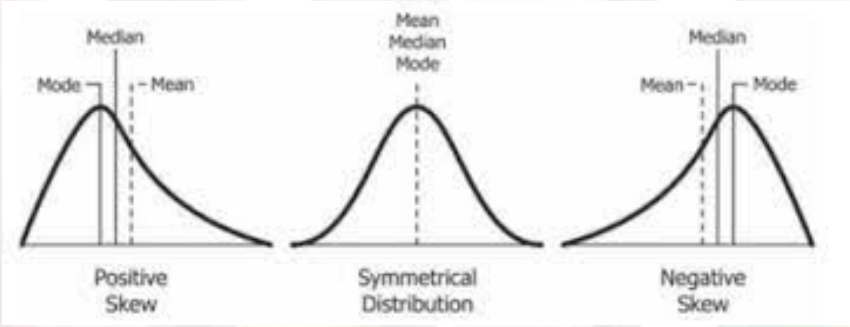
¿Por qué buscamos una distribución normal? Porque una distribución normal nos da ayuda a saber si nuestro examen está bien dirigido a nuestra población.

Buscaremos también evitar el efecto suelo y efecto techo

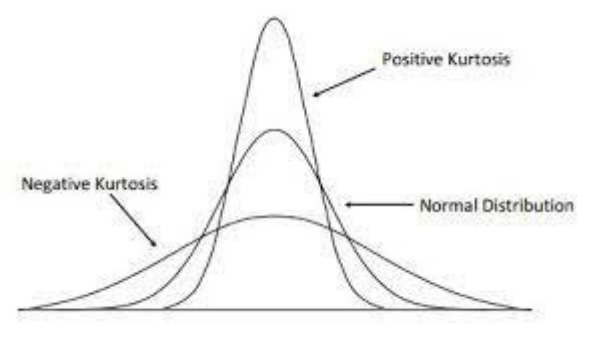
Veámoslo representado



Distribución normal



Coeficiente de asimetría



Curtosis

Julia Zabala Delgado

Análisis a nivel de prueba. Descriptivos e histogramas

Tarea 1.1 Cálculo de notas totales

Abrid archivo Tarea 1_empecemos.xlsx

Comprobad vuestros resultados abriendo archivo

Tarea 1_solucion paso 1

Julia Zabala Delgado

Análisis a nivel de prueba. Descriptivos e histogramas

Tarea 1.2 Estadística descriptiva de la prueba

Abrid archivo Tarea 1_solucion paso 1.xlsx

Comprobad vuestros resultados abriendo archivo

Tarea 1_solucion paso 2

Julia Zabala Delgado

Análisis a nivel de prueba. Descriptivos e histogramas

Tarea 1.3 Histogramas

Abrid archivo Tarea 1_solucion paso 2.xlsx

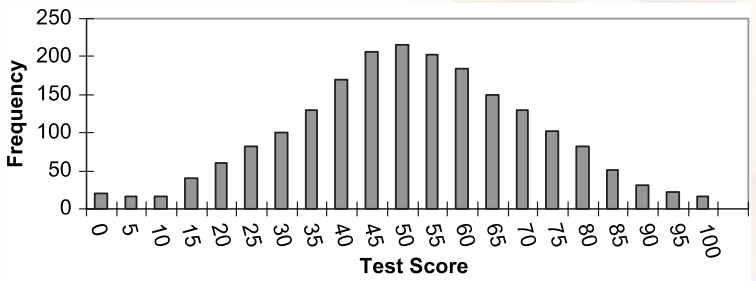
Comprobad vuestros resultados abriendo archivo

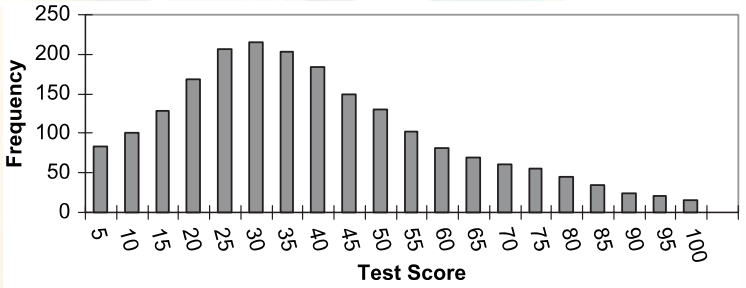
Tarea 1_solucion paso3 Histograma

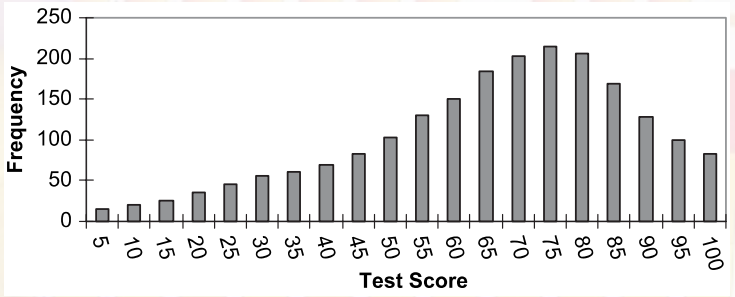
Julia Zabala Delgado

Ejercicio 1

Si el resultado de estas pruebas ha sido el esperado, ¿qué tipo de pruebas podría ser en cada uno de los casos? ¿Qué nos dicen los histogramas?





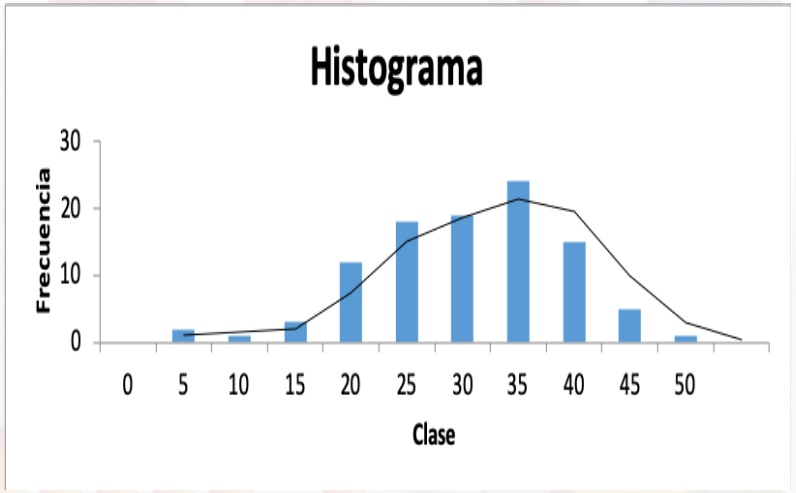


Ejemplos de Carr (2008)

Julia Zabala Delgado

Ejercicio 2

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
Coeficiente de asimetría	-0,44
Rango	45,00
Mínimo	4,00
Máximo	49,00



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

Julia Zabala Delgado

Análisis a nivel de prueba. Alfa de Cronbach

Tarea 2. Alfa de Cronbach

Abrid archivo Tarea 2_calcular alfa de Cronbach.xlsx

Comprobad vuestros resultados abriendo archivo

Tarea 2_solucion calcular alfa de Cronbach

Julia Zabala Delgado

Análisis a nivel de ítems

Índice de facilidad

Índice de discriminación

Funcionamiento de distractores

Análisis a nivel de ítem

- **Índice de facilidad del ítem:** porcentaje de alumnos que contestó correctamente a la respuesta. i.e. 33 de 45 estudiantes contestaron correctamente I.F (33/45) =0,73
- **Índice de discriminación del ítem:** capacidad del ítem para diferenciar a los candidatos entre aquellos que son más y menos competentes
- **Funcionamiento de distractores:** nos permite saber si el distractor funciona como tal.

<7% (no funciona)

(puede usarse10% con ítems de 4 opciones)

.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

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

Julia Zabala Delgado

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

Julia Zabala Delgado

Análisis a nivel de ítem

Análisis a nivel de ítem - índices de discriminación I

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

Julia Zabala Delgado

Análisis a nivel de ítem

Análisis a nivel de ítem - índices de discriminación II

Julia Zabala Delgado

Análisis a nivel de ítem

Análisis a nivel de ítem - índices de discriminación III

Solución: Abrid archivo Tarea 3_Analisis a nivel de item ID_solucion.xlsx

Julia Zabala Delgado

Señales de alarma 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

Análisis a nivel de ítem

Análisis a nivel de ítem – análisis de distractores

Abrid archivo Tarea 5_Analisis de distractores.xlsx

Solución: Abrid archivo Tarea 5_Analisis de distractores_solucion.xlsx

Julia Zabala Delgado

Análisis de correctores: fiabilidad intercorrector e intracorrector

Análisis de correctores

Abrid archivo Tarea 7_ correlaciones.xlsx

Solución: Abrid archivo Tarea 7_correlaciones_solucion

Interpretación de coeficientes de correlación

Débil: .0.1-0.3
moderada: 0.4-0.6
Fuerte: 0.7-0.9

Buscaremos un coeficiente de correlación de al menos .80 con las notas del comité experto y una media que no se desvíe en exceso de la media del comité

¡Cuidado!

- Que exista correlación no quiere decir que esta sea estadísticamente significativa
- La correlación no indica causalidad, el hecho de que dos variables estén correlacionadas no quiere decir que una cause la otra o viceversa
- ¡Es importante ver las medias! Un corrector que pone notas mucho mas bajas puede correlacionar con uno que las pone mucho mas altas si los rangos son los mismos

Si no obtenemos la correlación esperada, debemos preguntarnos:

- ¿Hemos puntuado revisando los criterios de corrección (las escalas)?
- ¿Hemos estado igual de atentos a los criterios en cada uno de los candidatos?
- ¿Hemos puntuado únicamente basándonos en la producción, sin que nos afectaran factores externos?
- ¿Hemos releído las producciones para evitar errores?
- ¿Si hemos "ajustado" el criterio a mitad corrección, nos hemos asegurado de revisar las anteriores corregidas y hacerlo en todas?
- ¿Nos hemos asegurado que el cansancio no afectaba a nuestra corrección?

Julia Zabala Delgado
juzadel@upv.es

www.upv.es