TESIS DE MAGISTER EN ESTADISTICA APLICADA

Teoría Moderna de Tests: Teoría de la Generalizabilidad

por ***José Luis Ruiz***

bajo la dirección de:

Profesor Dr. Raúl Pedro Mentz

Profesor Dr. Carlos Alberto Brailovsky

INSTITUTO DE INVESTIGACIONES ESTADISTICAS (INIE)

FACULTAD DE CIENCIAS ECONOMICAS

UNIVERSIDAD NACIONAL DE TUCUMAN

TESIS DE MAGISTER EN ESTADISTICA APLICADA

Teoría Moderna de Tests: Teoría de la Generalizabilidad

por ***José Luis Ruiz***

bajo la dirección de:

Profesor Dr. Raúl Pedro Mentz

Profesor Dr. Carlos Alberto Brailovsky

INSTITUTO DE INVESTIGACIONES ESTADISTICAS (INIE)

FACULTAD DE CIENCIAS ECONOMICAS

UNIVERSIDAD NACIONAL DE TUCUMAN

dedicado ...

a María, mi esposa y mis hijas María Victoria e Ileana Araceli, por la infinita paciencia y colaboración

a la memoria de mi padre, Alfredo, quien siempre me acompaña.

Contenidos

CAPITULO 5

CAPITULO 6

PARTE II

Agradecimientos

El haber participado del programa de postgrado y llegado a esta instancia se lo debo, en gran medida, al Profesor Dr. Raúl Pedro Mentz. Que alguien de su prestigio haya confiado permanentemente en mí significa un gran halago. Estoy en deuda, no sólo por su disponibilidad permanente para mis consultas sino también por hacerme sentir parte del INIE. desde un principio. Le agradezco la ayuda económica que me brindó y que me posibilitó seguir adelante.

Ha sido también un gran honor para mí contarme entre los alumnos del Profesor Dr. Carlos Alberto Brailovsky, quien gentilmente accedió a compartir la dirección de esta tesis y cuya gran amabilidad y paciencia sólo se compara con su enorme sapiencia.

Agradezco también el apoyo de las autoridades de la Universidad Nacional de Tucumán, en particular a su Sr. Rector, el C.P.N. Mario Marigliano y al Sr. Decano de la Facultad de Ciencias Económicas, C.P.N. Juan Carlos Cerissola, por la valiosa asistencia económica durante todo el tiempo que duró el programa y sin la cual probablemente no hubiese logrado alcanzar la meta.

Finalmente quiero agradecer también a los demás profesores del Instituto, sus secretarios Juan Carlos y Rubén y a mis compañeras de estudio, Estela, Jorgelina y Mónica, con quienes compartí muy gratos momentos.

José Luis Ruiz
Junio de 2.001

Resumen

Este trabajo se ha concebido en dos partes íntimamente relacionadas: la primera, titulada *Teoría Clásica de Tests* y que ocupa los seis primeros capítulos es una revisión de la Teoría Clásica del Score Verdadero, en sus aspectos más relevantes, mientras que la segunda parte, *Teoría Moderna de Tests: Teoría de la Generalizabilidad* que abarca los cinco últimos capítulos se dedica por completo a la presentación de esta teoría, cada vez más ampliamente utilizada por especialistas de diversas áreas del conocimiento en las que resulta necesario abrir un juicio acerca de los instrumentos de evaluación que se administran con la finalidad de recolectar mediciones.

En el Capítulo 1 se hace un brevísimo repaso de algunos conceptos básicos como el de medición, usos corrientes de un test y una clasificación de éstos en base a sus objetivos.

El Capítulo 2 presenta el modelo lineal de la Teoría Clásica, sus hipótesis y las conclusiones que se derivan y que son de gran interés en lo que sigue del trabajo.

En el Capítulo 3 se presenta en detalle el *Análisis de Confiabilidad* de un test, las medidas más frecuentes y los procedimientos para estimar el Coeficiente de Confiabilidad en diferentes situaciones. Se incluye también el Error Estándar de Estimación que permite construir intervalos de confianza para los scores verdaderos.

El Capítulo 4 se dedica por completo al *Análisis de Validez* de una prueba, los tipos más importantes y las medidas adecuadas en cada caso.

El Capítulo 5 cierra la parte teórica del enfoque clásico con el *Análisis de Item*, en el que se describen diversas técnicas para evaluar cada elemento del test, con vistas a confirmarlo, optimizarlo o excluirlo de la prueba.

El siguiente Capítulo 6 es de aplicación. Se aplican los conceptos presentados al caso de una evaluación (en Matemática y Lengua) llevada a cabo en el primer semestre del año 2.000, en los séptimos años de la Escuela *Docencia Tucumana* de la localidad de Villa Mariano Moreno, con

el fin de analizar las competencias adquiridas al término de la Educación General Básica, segundo nivel (E.G.B. II) en las asignaturas evaluadas.

La segunda parte se inicia con el Capítulo 7, que constituye una introducción a la Teoría de la Generalizabilidad. Se presentan conceptos elementales como los de Estudio G, Estudio D, Universo de Observaciones Admisibles, Universo de Generalización, los modelos y diseños posibles, los errores de medida y medidas de la precisión relativa de un procedimiento de medición.

En el Capítulo 8, *Universos de una Faceta*, se presenta la Teoría de la Generalizabilidad en el caso más simple de una única faceta (un único factor en el terminología del Análisis de la Varianza). Se explica el modelo, sus supuestos y se deriva el Coeficiente de Generalizabilidad para diferentes diseños.

El Capítulo 9 extiende la presentación al caso de *varias facetas*, pero restringiéndose a los *Modelos de Efectos Aleatorios*. Se muestran varios diseños, a pesar que algunos de ellos son poco usuales, sobre todo en aplicaciones en el área de Educación.

En el Capítulo 10 se intentan exponer procedimientos generales, válidos para cualquier caso, destinados a obtener estimaciones de las Componentes de la Varianza en los Estudios G y D y por supuesto para el Coeficiente de Generalizabilidad. Se incluye también, para cerrar el marco teórico, una breve introducción al Análisis de Generalizabilidad Multivariada, cuestión que de por sí podría ser materia de otro trabajo.

El Capítulo 11 se vuelve a las aplicaciones y se aplican los conceptos presentados al test que mide la competencia *Resolver Problemas* en Matemática. También se muestra cómo se obtienen resultados idénticos utilizando el Análisis de Generalizabilidad y el Análisis Discriminante cuando se busca establecer el subconjunto de competencias adquiridas que mejor discriminan a los grupos de alumnos clasificados de acuerdo al nivel de rendimiento exhibido. Se presenta finalmente un ejemplo de aplicación de la perspectiva multivariada. Es importante notar que se ha incorporado también, el manejo y utilización de un software específico de Análisis de Generalizabilidad, conocido como GT, en su versión para PC, con el que se *corren* todos los ejemplos.

PARTE I

TEORÍA CLÁSICA DE TESTS

CAPITULO 1

Breve Introducción a la Teoría de Medición

1.1. EL CONCEPTO DE "MEDICION"

Con el término *Medición* se hace referencia a un proceso mediante el cual se asocian números a objetos en una forma sistemática en un intento de representar ciertas propiedades de tales objetos.

La asignación de estos números se lleva a cabo en base a un procedimiento cuidadosamente establecido, en el que cada individuo (u otro objeto de medición) es expuesto a idénticas condiciones, como lo son las mismas instrucciones, preguntas y procedimientos de calificación (o puntuación). Esto hace posible interpretar y comparar los resultados que se presenten.

En particular, estaremos involucrados con la medición de *atributos psicológicos* de un individuo, como la fluidez de su vocabulario, su desarrollo de competencias en matemática, el grado de su integración social, etc.

Estos atributos psicológicos son conceptos teóricos (hipotéticos) desarrollados en un intento por construir teorías capaces de explicar el comportamiento humano. Un rasgo sobresaliente de estos *constructos*, a diferencias de atributos físicos o biológicos de un individuo, es que no pueden medirse directamente.

Esta imposibilidad de registrar directamente el grado en que una persona exhibe estos atributos deriva en el hecho en que el proceso de medición debe concretarse indirectamente, a través de comportamientos observables del individuo en diferentes escenarios. Sin dudas que esta impronta de los conceptos abstractos introduce ciertos problemas de medición como por ejemplo la posibilidad de disponer de más de una *definición operacional* para la misma variable.

Una *definición operacional* es un proceso por el cual se establecen, claramente, las relaciones de correspondencias entre el concepto hipotético y el comportamiento observado que se asocia al mismo. Un educador, por ejemplo, debería establecer en forma precisa qué aptitudes o habilidades debería exhibir un alumno como instancia previa a la medición del desarrollo de su competencia en la resolución de problemas matemáticos.

Se genera, entonces, la necesidad de construir un instrumento o *test* que permita *medir* estos constructos. En este contexto, un test puede ser pensado como un procedimiento cuya finalidad es la obtención de una muestra del desempeño típico de un individuo (por ejemplo cuestionarios en los que el individuo informa sobre sus intereses personales, sentimientos, actitudes, etc) o de su desempeño óptimo (como pruebas de logros en las que la persona intenta desempeñarse al máximo de sus capacidades).

La Teoría de Medición puede considerarse un área de la Estadística Aplicada que tiene como finalidad la descripción, categorización y evaluación de la calidad de las medidas, el mejoramiento de la precisión e interpretación de éstas y la elaboración de métodos para desarrollar instrumentos de medición cada vez más eficientes.

1.2. USOS CORRIENTES DE UN TEST

El uso de tests como herramientas útiles se ha generalizado en ámbitos que van desde la educación y la salud a la industria. Son considerados instrumentos claves en procesos de selección de candidatos (a puestos de trabajo o programas de entrenamiento, etc.), de clasificación de individuos (según su estado de salud física o mental, de acuerdo al desarrollo de ciertas aptitudes, etc.), de evaluación (con la finalidad de acreditar suficiencia para el ejercicio de profesiones o juzgar la efectividad de programas de enseñanza, etc.) o en proyectos de investigación (como procedimientos que miden diferentes comportamientos y permiten examinar sus relaciones mutuas), entre otros.

Si se hace referencia a Metodología de Investigación y Evaluación en Ciencias Sociales, debe destacarse el papel decisivo que se atribuye a la Teoría de Tests. Una vez que la o las hipótesis de investigación han sido formuladas, la Teoría de Tests debe ser considerada en diferentes instancias: la especificación de la definiciones operaciones para cada variable que establecen el modo en que éstas deben ser medidas o controladas; la selección de instrumentos de medición adecuados que prescriben la forma de

obtener y cuantificar las observaciones sobre cada variable y el análisis de la precisión o sensitividad de tales instrumentos.

1.3. TESTS REFERENCIADOS EN NORMAS Y EN CRITERIOS

La introducción de la idea de estos tests se debe a R. Glaser (1963) en su ensayo *Instructional Technology and the Measurement of Learning Outcomes: Some Questions.*

1.3.1. TEST REFERENCIADO EN UNA NORMA

En un test *referenciado en una norma*, el desempeño de una persona es juzgado teniendo en cuenta el desempeño de otros individuos en ese test.

Supongamos por ejemplo que un alumno obtuvo 70 puntos en un test cuya escala era 0 : 100 puntos. Para evaluar el desempeño relativo de este alumno puede construirse la distribución de frecuencias de los scores observados en el grupo de estudiantes. Imaginemos que los 70 puntos alcanzados se identifican con el 90° percentil de la distribución. De esto puede inferirse que el estudiante en cuestión muestra un desempeño igual o superior al noventa por ciento de los alumnos que integran este grupo. Este grupo por lo general se conoce como *norma* y usualmente se constituye por una muestra seleccionada aleatoriamente.

De esta forma un test referenciado en una norma provee información acerca de la posición relativa de una persona en la distribución de los scores observados del grupo al que pertenece. Este es el tipo de test que se construye cuando interesa llevar a cabo un *estudio comparativo* del desempeño de los individuos. Tal es el caso, por ejemplo, cuando se busca seleccionar el subconjunto de postulantes con máximos scores para un programa de entrenamiento o puestos de trabajo.

1.3.2. TEST REFERENCIADO EN UN CRITERIO

En este caso el desempeño de un individuo es evaluado desde un conjunto definido de comportamientos de referencia, un *criterio* preestablecido.

Generalmente este criterio define un *dominio específico*, como por ejemplo algún aspecto específico de la Comprensión Lectora como la operación de resumen del texto, o el conocimiento de ciertos hechos puntuales de la Historia Nacional, etc.

Usualmente se construyen seleccionando aleatoriamente una muestra de comportamientos que integran el dominio específico. Los tests elaborados de esa forma se conocen comúnmente como *formas aleatoriamente paralelas*. Esta denominación pretende expresar que tales instrumentos son equivalentes en contenidos, aunque no constituyen tests *estrictamente paralelos* que requieren de la igualdad de medias, varianzas y correlaciones con otras variables.

Un test referenciado en un criterio permite establecer si el individuo se desempeña adecuadamente o no en ese dominio específico, sin necesidad de contar con información alguna sobre los scores obtenidos por otros individuos. Tal podría ser el caso de una prueba de logros que se aplica a un alumno para decidir si éste está en condiciones de ser promovido al siguiente nivel de dificultad de aprendizaje.

Este tipo de test resulta adecuado cuando se intenta realizar un *estudio absoluto* del desempeño de un alumno, como oposición al estudio comparativo descripto en el apartado anterior.

Finalmente, pueden señalarse dos propósitos básicos de esta clase de tests:

- ✓ La estimación del Score sobre el Dominio, que se define como la proporción de items del dominio que el individuo puede responder correctamente.

- ✓ La asignación de individuos en alguna de las categorías mutuamente excluyentes en las que se divide la escala del score de dominio, de acuerdo a la destreza alcanzada en el dominio definido representada por el score obtenido en el test. Estas categorías se separan por *puntos de corte* y en el caso más sencillo un solo punto de corte divide la escala en dos regiones que permiten decidir la suficiencia o no del nivel de destreza logrado.

CAPITULO 2

Teoría Clásica del Score Verdadero: Modelo y Supuestos

Una *Teoría de Test* o *Modelo de Test* es una representación simbólica de los factores que influyen en los scores observados de un test y se describe mediante un modelo y sus supuestos.

Muchos de los procedimientos estándares para construir y evaluar tests están basados en un conjunto de hipótesis, generalmente conocido como *Teoría Clásica (o Débil) del Score Verdadero.*

La Teoría Clásica del Score Verdadero es un modelo simple y útil que describe cómo los errores de medición pueden influenciar las puntuaciones (scores) observadas. Permite establecer principios para la construcción de un test, juzgar su validez y confiabilidad. Como cualquier modelo supone ciertas condiciones como ciertas; si estos supuestos son razonables, entonces las conclusiones derivadas del modelo serán razonables. Sin embargo si las hipótesis no son razonables, entonces el uso del modelo puede llevar a conclusiones erróneas.

La tarea de presentar y describir la Teoría Clásica es asumida aquí como una instancia previa, necesaria, en el proceso de aproximación hacia lo que constituye el núcleo de este trabajo: *La Teoría de la Generalizabilidad,* entendida como un esfuerzo que se orienta a incrementar la precisión en la interpretación de un test.

Este capítulo incluye, en primer lugar, un listado de los supuestos de la Teoría Clásica del Score Verdadero y luego explicaremos cada uno de ellos en detalle. Finalmente se presentan también algunas conclusiones que se derivan de estos supuestos y que resultan de gran utilidad en posteriores resultados.

2.1. LAS HIPÓTESIS DE LA TEORÍA CLÁSICA DEL SCORE VERDADERO

La Teoría Clásica del Score Verdadero se basa en cinco supuestos elementales que se presentan a continuación.

Hipótesis 1:

El Score Observado en un individuo "j", en una ocasión "k" de test, denotado como X_{jk}, es la suma del Score Verdadero T_j de ese individuo, más una componente aleatoria de error E_{jk}, denominado Score de Error o Error de Medición del individuo "j" en la ocasión "k".

En símbolos:

$$X_{jk} = T_j + E_{jk} \qquad k = 1,2,... \tag{2.1}$$

Para el j-ésimo individuo, en un test determinado, T_j se asume como un valor fijo, a pesar que X_{jk} y E_{jk} varían para ese individuo en diferentes ocasiones de evaluación (al variar el índice *k*).

Imaginemos, por ejemplo que se administra un test para medir el desarrollo de una competencia en Matemática a un grupo de alumnos. Supongamos que el score verdadero de José es 58, pero su score observado es 62, entonces X_{j1} =62, T_j =58 y E_{j1} = +4. Si José es evaluado nuevamente y su score observado es 52, entonces X_{j2} =52, T_j sigue siendo 58 y E_{j2} = -6.

Como se ve, el supuesto establece que los componentes del score se combinan en forma aditiva. La *hipótesis de aditividad* es común en trabajos estadísticos, porque es matemáticamente simple y aparece como razonable.

Debe notarse la naturaleza *aleatoria* de la variable *Score Observado X* que representa la calificación asignada a un individuo en un test. Supongamos por un momento que el test administrado consta de 100 items y que se asigna un punto por cada item correctamente contestado. Para un individuo cualquiera que tomara el test, la variable X podría entonces asumir cualquier valor entre 0 y 100 en esta escala. Obviamente no hay forma de anticipar el score que obtendrá el estudiante ya que no es posible predecir con exactitud la cantidad de veces que se distraerá, que no interpretará correctamente un enunciado o que *adivinará* con suerte algunas respuestas.

De esta forma, puede pensarse en X como una variable aleatoria a la que se asocia una determinada distribución de probabilidad y el score observado en un ensayo concreto del test como una *realización* de esta variable aleatoria. Una forma de estimar esta distribución de probabilidad consistiría en aplicar el test a ese estudiante un gran número de veces, adoptando las medidas

necesarias para que el alumno *no recuerde* cada ensayo previo. Luego la distribución de frecuencias de los scores observados construida a partir de una larga secuencia de ensayos constituye una estimación de esa distribución de probabilidad.

Hipótesis 2:

El valor esperado (media poblacional) del Score Observado X_{jk} del j-ésimo individuo, sobre todas las posibles ocasiones "k" del test, es el Score Verdadero T_j de ese individuo.

En símbolos:

$$\underset{k}{\mathrm{E}}\left(X_{jk}\right)=T_j \qquad (2.2)$$

Este supuesto constituye una definición para el Score Verdadero T_j, que se asume como la media de la distribución teórica de probabilidad de los scores X_{jk} que se encontraría en ensayos independientes *en la misma persona con el mismo test.* Por ejemplo si pudiéramos evaluar a José un número infinito de veces, la media de sus scores observados sería 58.

Debe advertirse que en la definición del score verdadero se da por supuesto que los ensayos son independientes, es decir que cada ocasión en que se aplica el test no influencia ninguna evaluación subsiguiente. En la práctica, sin embargo, resulta casi imposible conseguir esta situación de total independencia entre ensayos (dado que por ejemplo el individuo puede recordar sus respuestas del ensayo anterior o haber *aprendido* durante el test o bien no encontrarse en la misma condición física o anímica cada vez que es expuesto al test, etc.). Por otra parte, tampoco es posible disponer de infinitas pruebas repetidas. Todo esto conduce a concebir T como un valor teórico. Sin embargo, como veremos, este valor teórico permite lograr algunos resultados muy útiles.

Una cuestión que no puede obviarse surge del hecho que el verdadero score es un concepto probabilístico, esto es se define en términos del valor esperado para el score observado y no en términos de alguna característica *real* del individuo como ocurre cuando se trabaja en áreas como la biología o la física. Algunos ejemplos pueden auxiliarnos en la interpretación de este aspecto. Supongamos el caso de dos individuos a los que se administra el mismo test que intenta medir ciertas habilidades en el desarrollo de la escritura y que tiene un valor máximo en su escala de 50 puntos. Imaginemos que ambos individuos alcanzan el score máximo de 50 puntos debido a que el nivel de dificultad del test es demasiado bajo. Aunque por otros medios tuviéramos ocasión de conocer que uno de estos individuos es realmente más hábil que el

otro, los scores verdaderos de ambas personas pueden coincidir en la escala, dado que éstos se definen en relación a un proceso de medición específico (un dado test) más que en torno al nivel real de aptitud que cada individuo ha desarrollado. O bien puede ocurrir que un estudiante, que ha desarrollado una gran aptitud en dicha competencia, tuviera problemas en interpretar las consignas de cada item del test. Es probable que sus calificaciones en ensayos repetidos sean sistemáticamente bajas debido a esta causa y su score verdadero será también bajo a pesar que su nivel de desarrollo en esta área pueda ser elevado. El error de medición es una desviación asistemática o aleatoria del score observado de un individuo del score teórico. Los errores sistemáticos no son considerados *errores de medición* para la Teoría Clásica.

En la teoría clásica de test, el score verdadero es la media teórica de los resultados de ensayos repetidos en forma independiente. Si este score verdadero refleja en forma adecuada alguna habilidad teórica o característica es una cuestión de *validez* del test que se presentará más adelante.

Nota: La hipótesis 2 trata sobre la distribución teórica de scores observados en diferentes ensayos *para un individuo y un test*. Las hipótesis 3 a 5 tratan sobre el score verdadero y de error para uno o dos test para una población de individuos en un ensayo del test.

Hipótesis 3:

Los scores de error y verdaderos obtenidos en una población de individuos, en una ocasión de un test, no están correlacionados.

En símbolos:

$$\rho_{E_j T_j} = 0 \qquad j = 1,2,... \tag{2.3}$$

donde el índice *j* identifica a cada individuo de esa población.

Este supuesto implica que no hay asociación entre valores altos de scores verdaderos con errores de signo positivo o negativo, respecto a valores bajos en estos scores.

Alternativamente podríamos interpretar esta hipótesis estableciendo que el valor esperado de los scores de error, para un dado score verdadero, es nulo. Esto es si μ_{Ej} la esperanza de los errores de medición para el individuo j, debe cumplirse que $\mu_{Ej} = 0$.

Supongamos por un momento que de alguna forma conociéramos los verdaderos scores sobre un test de un grupo de individuos y lo aplicáramos un gran número de veces. Podríamos entonces representar gráficamente estas observaciones como puntos en un par de ejes coordenados: sobre el eje horizontal se localizarían los valores de los verdaderos scores mientras que sobre el eje vertical los correspondientes a los errores registrados para cada individuo. Bajo el supuesto de no correlación entre T_j y E_j, el gráfico no debería mostrar ningún patrón sistemático de relaciones entre estas cantidades.

La Figura 2.1 podría ser una representación adecuada de esa hipótesis, mientras que las Figuras 2.2. y 2.3. deberían tomarse como un serio indicio de la falta de cumplimiento de este supuesto.

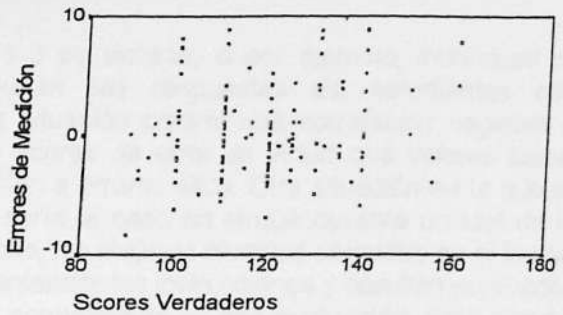


Fig. 2.1. No correlación entre los verdaderos scores y los errores.

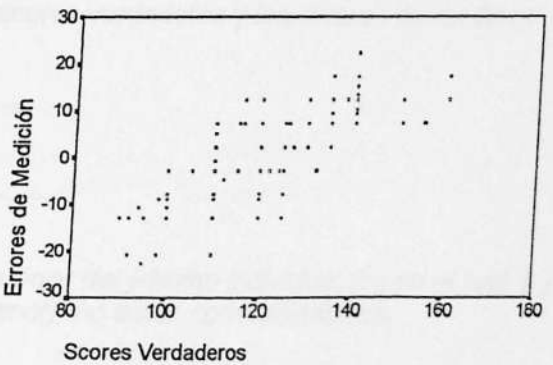


Fig. 2.2. Correlación positiva entre los verdaderos scores y los errores.

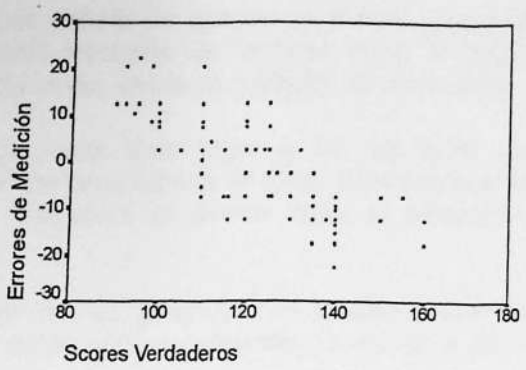


Fig. 2.3. Correlación negativa entre los verdaderos scores y los errores.

La hipótesis 3 se violaría, si por ejemplo, individuos con bajos scores verdaderos copiaran las respuestas de estudiantes con altos scores verdaderos. Esta situación crearía una correlación negativa entre los scores verdaderos y los scores de error en virtud que valores bajos de verdaderos scores se asociarían a errores altos. Otra situación en la que este supuesto se podría contrariar sería el caso en el que durante un test de logros, debido a problemas acústicos, los mejores alumnos ubicados en el fondo del aula tienen dificultades para entender las instrucciones y resultan puntuados por debajo de lo que obtendrían normalmente en esa evaluación. Esta circunstancia también generaría una correlación negativa entre los scores verdaderos y los scores de error debido a que valores altos del verdadero score se asociarían a valores negativos del error. Claramente, si la situación fuera la inversa y los mejores alumnos se ubicaran en el frente del aula, podría generarse una correlación positiva entre los scores verdaderos y los errores de medición.

Hipótesis 4:

Los scores de error del j-ésimo individuo, E_{1j} en el test 1 y E_{2j} en otro test 2 (diferente del anterior), no están correlacionados.

En símbolos:

$$\rho_{E_{1j}E_{2j}} = 0 \qquad j = 1,2,... \tag{2.4}$$

Es decir si una persona tiene un score de error positivo en un test, esto no significa que tendrá un score de error positivo o negativo en otro test.

Esta situación se violaría por ejemplo en el caso en que las puntuaciones de los test estuvieran afectadas de *factores como la fatiga, el humor del individuo, efectos del medio ambiente o efectos de aprendizaje.*

El factor *fatiga* puede tener lugar si los dos tests considerados son aplicados al final de una larga batería de tests. Esto provocaría que los errores de medición fueran negativos en ambos tests, al producirse observaciones inusualmente bajas.

El factor *humor* (o más generalmente el estado anímico) del individuo puede influenciar el desempeño provocando correlaciones entre los errores de medición.

Otros individuos, por ejemplo, pueden beneficiarse de la práctica provista por exámenes previos y en consecuencia luego de un *proceso de aprendizaje* pueden obtener puntuaciones inusualmente altas en los dos últimos tests, resultando errores de medición positivos en ambos tests. Una situación como esta puede provocar una correlación positiva entre los errores de medición de ambos tests.

En situaciones para las que se conoce que afectarán factores de este tipo, como efectos de aprendizaje, fatiga o condiciones ambientales y se pretende aplicar la teoría clásica de score se debería asegurar que las condiciones de evaluación sean lo más *homogéneas* posible para todos los individuos sobre todos los test en todos los ensayos. Este control puede reducir la medida del error de medición en cada test y las correlaciones entre los scores de error entre los tests.

Hipótesis 5:

Si E_{1j} es el score de error del j-ésimo individuo en un test 1 y T_{2j} es el score verdadero del mismo individuo en otro test 2, entonces E_{1j} y T_{2j} son no correlacionados.

En símbolos:

$$\rho_{E_{1j}T_{2j}} = 0 \qquad j = 1,2,... \tag{2.5}$$

Este supuesto se violaría en situaciones idénticas a las que contrarían el supuesto número 3.

2.2. TESTS *ESTRICTAMENTE PARALELOS* Y TESTS *ESENCIALMENTE τ (tau) EQUIVALENTES*

2.2.1. TESTS *ESTRICTAMENTE PARALELOS*

Si el j-ésimo individuo tiene scores observados X_{1j} y X_{2j} sobre dos tests diferentes que satisfacen los supuestos 1 a 5 y si se cumple que:

$$\begin{aligned} T_{1j} &= T_{2j} \\ \sigma^2_{E_{1j}} &= \sigma^2_{E_{2j}} \end{aligned} \qquad j = 1,2,.. \tag{2.6}$$

entonces, los test se dicen *estrictamente paralelos.*

En la definición anterior, X_{1j} es el score observado en el j-ésimo individuo en el test 1, T_{1j} es su score verdadero en ese test y σ^2_{E1j} es la varianza del error. La varianza del error es la varianza del score de error para ese test entre todos los individuos de una población particular. Por otra parte, X_{2j}, T_{2j} y σ^2_{E2j} son el score observado y el score verdadero del j-ésimo individuo y la varianza del error, en el test 2.

La definición dada establece que dos tests estrictamente paralelos, también llamados *formas alternativas* no sólo coinciden en sus scores verdaderos sino también en sus varianzas teóricas de los errores de medición para toda población de alumnos que tome ambos tests

Que las varianzas en ambos tests deban ser iguales, implica que las condiciones prácticas bajo las cuales se llevan a cabo los ensayos deben ser lo más homogéneas posibles, es decir condiciones ambientales (como el ruido) o personales (físicas o anímicas) deberían permanecer constantes para ambos tests.

Debe notarse que de la definición dada para tests estrictamente paralelos se deduce que las puntuaciones observadas, no sólo comparten el mismo valor esperado sino también la misma varianza. Es decir se cumple que:

$$\sigma^2_{X_j} = \sigma^2_{X_{j'}} \qquad j = 1,2,...$$

resultado que será probado más adelante en este capítulo.

Pero no debe suponerse que los scores observados en ambos tests deban coincidir exactamente de una prueba a otra para una persona cualquiera. Dos individuos pueden (y en general esto sucederá) tener diferentes scores observados, a menos que la varianza de error sea igual a cero.

2.2.2. TESTS *ESENCIALMENTE* τ *(tau) EQUIVALENTES*

Si el j-ésimo individuo tiene scores observados X_{1j} y X_{2j} sobre dos tests diferentes que satisfacen los supuestos 1 a 5 y si se cumple que:

$$T_{1j} = T_{2j} + c \qquad j = 1,2,... \tag{2.7}$$

para todo individuo de la población, donde c es una constante, entonces los test se dicen *esencialmente* τ *equivalentes*.

Esto significa que los tests *esencialmente* τ *equivalentes* tienen scores verdaderos que difieren sólo en una constante.

Por otra parte, las varianzas del error pueden ser distintas, a diferencia de las formas paralelas. Esto significa que el score verdadero se puede medir en forma más precisa con uno de estos test.

Debemos notar que dos tests paralelos son esencialmente τ equivalentes, pero lo contrario no es cierto.

2.3. CONCLUSIONES QUE SE DERIVAN DE LAS HIPÓTESIS DE LA TEORÍA CLÁSICA DEL SCORE VERDADERO

A continuación se presentan dieciocho resultados que se deducen de los supuestos del modelo propuesto por la Teoría Clásica. Estas conclusiones son de gran importancia en aspectos que se discutirán posteriormente.

Conclusión 1:

Para cualquier individuo particular, el valor esperado del score de error, sobre todas las ocasiones posibles del test, es cero.

En símbolos,

$$\underset{k}{\mathbf{E}}(E_{jk}) = 0 \qquad (2.8)$$

donde el índice *k* varía recorriendo todas las ocasiones posibles del test.

Demostración:

Como por (2.1) se cumple que:

$$X_{jk} = T_j + E_{jk} \qquad k = 1,2,...$$

Aplicando el operador Esperanza,

$$\underset{k}{\mathbf{E}}(X_{jk}) = \underset{k}{\mathbf{E}}(T_j + E_{jk}) \qquad k = 1,2,...$$

$$= \underset{k}{\mathbf{E}}(T_j) + \underset{k}{\mathbf{E}}(E_{jk})$$

donde la última igualdad se logra por la linealidad del Valor Esperado. Pero teniendo presente el supuesto 2 y que T_j es una constante,

$$T_j = T_j + \underset{k}{\mathbf{E}}(E_{jk}) \qquad k = 1,2,...$$

$$\therefore \underset{k}{\mathbf{E}}(E_{jk}) = 0. \; \Diamond$$

Conclusión 2:

Para una población de individuos que toman un test, el valor esperado del producto del score de error por el score verdadero es cero .

En símbolos,

$$\mathbf{E}(E_j \cdot T_j) = 0. \qquad (2.9)$$

Demostración:

Por el supuesto 3,

$$\rho_{E_j T_j} = \frac{\sigma_{E_j T_j}}{\sigma_{E_j} \cdot \sigma_{T_j}} = 0,$$

lo que ocurre si y sólo si

$$\sigma_{E_j T_j} = 0,$$

puesto que

$$\sigma_{E_j} \sigma_{T_j} \neq 0.$$

Pero por propiedades de covarianza,

$$\sigma_{E_j T_j} = \mathbf{E}(E_j T_j) - \mathbf{E}(E_j).\mathbf{E}(T_j),$$

de manera que

$$0 = \mathbf{E}(E_j T_j) - 0,$$

$$\therefore \mathbf{E}(E_j T_j) = 0. \lozenge$$

Como se ve, también la covarianza entre el score de error y el score verdadero es cero.

Conclusión 3:

Para una población de alumnos que toman un test, la varianza del score observado es la suma de la varianza del score verdadero más la varianza del error.

En símbolos,

$$\sigma^2_{X_j} = \sigma^2_{T_j} + \sigma^2_{E_j}. \qquad \textbf{(2.10)}$$

Demostración:

Tomando varianza en ambos miembros de la expresión (2.1),

$$\sigma^2_{X_j} = \sigma^2_{(T_j+E_j)},$$

y desarrollando la varianza de la suma del segundo miembro,

$$\sigma^2_{X_j} = \sigma^2_{T_j} + \sigma^2_{E_j} + 2 \cdot \sigma_{E_j T_j},$$

pero teniendo en cuenta la Conclusión 2 resulta que

$$\sigma^2_{X_j} = \sigma^2_{T_j} + \sigma^2_{E_j} \; . \; \Diamond$$

Hay que notar que si las observaciones se midieran sin error, entonces $\sigma^2_{Ej}=0$ y en este caso toda la varianza del score observado correspondería a la varianza del score verdadero es decir reflejaría las diferencias entre los scores verdaderos de los individuos.

Si las mediciones se registran con errores, entonces parte de la varianza de los scores observados podrán atribuirse a las diferencias entre los scores verdaderos, pero otra parte deberá atribuirse al error.

Conclusión 4:

Para una población de alumnos que toman un test, el cuadrado de la correlación entre los scores observados y verdaderos es el cociente entre la varianza del score verdadero y la varianza del score observado.

En símbolos,

$$\rho^2_{X_j T_j} = \frac{\sigma^2_{T_j}}{\sigma^2_{X_j}} . \qquad \textbf{(2.11)}$$

Demostración:

Por definición de coeficiente de correlación y por propiedades de la covarianza,

$$\rho^2_{X_jT_j} = \left[\frac{\sigma_{X_jT_j}}{\sigma_{X_j}\cdot\sigma_{T_j}}\right]^2,$$

$$= \left[\frac{\mathbf{E}(X_jT_j) - \mathbf{E}(X_j)\cdot\mathbf{E}(T_j)}{\sigma_{X_j}\cdot\sigma_{T_j}}\right]^2,$$

$$= \left[\frac{\mathbf{E}[(T_j + E_j)\cdot T_j] - \mathbf{E}(X_j)\cdot\mathbf{E}(T_j)}{\sigma_{X_j}\cdot\sigma_{T_j}}\right]^2,$$

por lo que si expandimos el numerador y tenemos en cuenta la Conclusión 2,

$$\rho^2_{X_jT_j} = \left[\frac{\mathbf{E}(T_j^2) + \mathbf{E}(E_jT_j) - \mathbf{E}(T_j)\cdot\mathbf{E}(T_j)}{\sigma_{X_j}\cdot\sigma_{T_j}}\right]^2,$$

$$= \left[\frac{\mathbf{E}(T_j^2) - \mathbf{E}^2(T_j)}{\sigma_{X_j}\cdot\sigma_{T_j}}\right]^2,$$

Advirtiendo que el numerador del segundo miembro es precisamente la varianza de los scores verdaderos y simplificando

$$\rho^2_{X_jT_j} = \left[\frac{\sigma^2_{T_j}}{\sigma_{X_j}\cdot\sigma_{T_j}}\right]^2,$$

$$= \left[\frac{\sigma_{T_j}}{\sigma_{X_j}}\right]^2,$$

$$= \frac{\sigma^2_{T_j}}{\sigma^2_{X_j}}. \Diamond$$

Conclusión 5:

Para una población de alumnos que toman un test, el cuadrado de la correlación entre los scores observados y verdaderos es igual a uno menos el cociente entre la varianza del score de error y la varianza del score observado.

En símbolos,

$$\rho^2_{X_jT_j} = 1 - \frac{\sigma^2_{E_j}}{\sigma^2_{X_j}}. \qquad (2.12)$$

Demostración:

Haciendo uso de la Conclusión 4,

$$\rho^2_{X_jT_j} = \frac{\sigma^2_{T_j}}{\sigma^2_{X_j}},$$

$$= \frac{\sigma^2_{X_j} - \sigma^2_{E_j}}{\sigma^2_{X_j}},$$

$$= 1 - \frac{\sigma^2_{E_j}}{\sigma^2_{X_j}}. \Diamond$$

Conclusión 6:

Si X_{1j} y X_{2j} son scores observados de tests estrictamente paralelos, entonces sus varianzas son iguales.

En símbolos,

$$\sigma^2_{X_{1j}} = \sigma^2_{X_{2j}}. \qquad (2.13)$$

Demostración

Para el test 1 vale que:

$$\sigma^2_{X_{1j}} = \sigma^2_{T_{1j}} + \sigma^2_{E_{1j}} ,$$

mientras que para el test 2:

$$\sigma^2_{X_{2j}} = \sigma^2_{T_{2j}} + \sigma^2_{E_{2j}} ,$$

pero por ser tests aleatoriamente paralelos, las varianzas de sus errores son iguales y también lo son las varianzas de sus scores verdaderos, por lo cual,

$$\sigma^2_{X_{1j}} = \sigma^2_{X_{2j}} \cdot \Diamond$$

Observar que esta es una condición necesaria pero no suficiente para paralelismo.

Conclusión 7:

Si X_{1j} y X_{2j} son scores de tests aleatoriamente paralelos, se correlacionan en forma idéntica con otro score Y.

En símbolos,

$$\rho_{X_{1j}Y} = \rho_{X_{2j}Y} \cdot \qquad \textbf{(2.14)}$$

Demostración:

De la definición de coeficiente de correlación se sigue que:

$$\rho_{X_{1j}Y_j} = \frac{\sigma_{X_{1j}Y_j}}{\sigma_{X_{1j}} \cdot \sigma_{Y_j}} ,$$

$$= \frac{\sigma_{(T_{1j}+E_{1j})Y_j}}{\sigma_{X_{1j}} \cdot \sigma_{Y_j}} ,$$

Ahora se puede expandir el numerador y teniendo presente que la covarianza σ_{Ey} es cero,

$$\rho_{XY} = \frac{\sigma_{T_{1j}Y_j} + \sigma_{E_{1j}Y_j}}{\sigma_{X_{1j}} \cdot \sigma_{Y_j}},$$

$$= \frac{\sigma_{T_{1j}Y_j}}{\sigma_{X_{1j}} \cdot \sigma_{Y_j}},$$

Pero dado que los tests 1 y 2 son aleatoriamente paralelos, podemos intercambiar sus scores verdaderos como así también las varianzas de sus scores observados:

$$\rho_{X_{1j}Y_j} = \frac{\sigma_{T_{2j}Y_j}}{\sigma_{X_{2j}} \cdot \sigma_{Y_j}},$$

$$\rho_{X_{1j}Y_j} = \rho_{X_{2j}Y_j} \cdot \Diamond$$

Este resultado sugiere una técnica alternativa para probar el paralelismo entre dos pruebas, comprobando la igualdad de las correlaciones de los scores observados de ambos tests con una tercera variable.

Conclusión 8:

La correlación entre los scores observados de dos formas paralelas es igual al cociente entre la varianza de los scores verdaderos sobre la varianza de los scores observados sobre la base de cualquiera de los tests.

En símbolos,

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma^2_{T_{1j}}}{\sigma^2_{X_{1j}}} = \frac{\sigma^2_{T_{2j}}}{\sigma^2_{X_{2j}}}. \qquad \textbf{(2.15)}$$

Demostración:

Por definición de coeficiente de correlación

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma_{X_{1j}X_{2j}}}{\sigma_{X_{1j}} \cdot \sigma_{X_{2j}}},$$

$$= \frac{\sigma_{(T_{1j}+E_{1j})(T_{2j}+E_{2j})}}{\sigma^2_{X_{1j}}},$$

donde la última igualdad procede de expresar en el numerador, los scores observados de cada tests en función de sus scores verdaderos y errores, mientras que los factores del denominador son iguales por ser formas paralelas. Desarrollando el numerador y teniendo presente que:

$$\sigma_{E_{1j}T_{2j}} = \sigma_{E_{2j}T_{1j}} = \sigma_{E_{1j}E_{2j}} = 0$$

se llega a que:

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma_{T_{1j}T_{2j}} + \sigma_{T_{1j}E_{2j}} + \sigma_{T_{2j}E_{1j}} + \sigma_{E_{1j}E_{2j}}}{\sigma^2_{X_{1j}}},$$

$$= \frac{\sigma_{T_{1j}T_{2j}}}{\sigma^2_{X_{1j}}},$$

$$= \frac{\sigma^2_{T_{1j}}}{\sigma^2_{X_{1j}}},$$

$$= \frac{\sigma^2_{T_{2j}}}{\sigma^2_{X_{2j}}} \cdot \Diamond$$

Esta correlación toma su máximo valor 1 cuando las mediciones se hacen sin error, esto es la varianza del error es nula y por lo tanto las varianzas de los scores observados y verdaderos coinciden exactamente.

Por otra parte, es claro que este coeficiente asumirá su mínimo, igual a 0, cuando toda la varianza de los scores observados corresponda a varianza del término de error y en consecuencia sea nula la varianza de los verdaderos scores.

Conclusión 9:

La correlación entre los scores observados de dos formas paralelas es igual a uno menos el cociente entre la varianza de los scores de error sobre la varianza de los scores observados.

En símbolos,

$$\rho_{X_{1j}X_{2j}} = 1 - \frac{\sigma^2_{E_{1j}}}{\sigma^2_{X_{1j}}}. \qquad \textbf{(2.16)}$$

Demostración:

Por la conclusión anterior se sabe que

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma^2_{T_{1j}}}{\sigma^2_{X_{1j}}},$$

por lo que haciendo uso de la Conclusión 3,

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma^2_{X_{1j}} - \sigma^2_{E_{1j}}}{\sigma^2_{X_{1j}}},$$

$$= 1 - \frac{\sigma^2_{E_{1j}}}{\sigma^2_{X_{1j}}}. \Diamond$$

La ecuación (2.16) es una forma alternativa del resultado expresado en la (2.15). De nuevo se pone de manifiesto que si la varianza del error es cero, entonces la correlación de las formas "estrictamente paralelas" es máxima. Será mínima en el supuesto que la varianza de las observaciones correspondan enteramente a la variabilidad del *ruido* del modelo.

Conclusión 10:

La correlación entre dos scores observados de tests paralelos es uno menos el cuadrado de la correlación (coeficiente de determinación) entre los scores observados y los errores.

En símbolos,

$$\rho_{X_{1j}X_{2j}} = 1 - \rho^2_{X_{1j}E_{1j}} \,. \tag{2.17}$$

Demostración:

Por definición de coeficiente de correlación,

$$\rho^2_{X_{1j}E_{1j}} = \left[\frac{\sigma_{X_{1j}E_{1j}}}{\sigma_{X_{1j}} \cdot \sigma_{E_{1j}}}\right]^2,$$

$$= \left[\frac{\sigma_{(T_{1j}+E_{1j})E_{1j}}}{\sigma_{X_{1j}} \cdot \sigma_{E_{1j}}}\right]^2,$$

donde la última igualdad se logra utilizando la expresión (2.1). Expandiendo el numerador y recordando el supuesto de que la covarianza entre los errores y los scores verdaderos en un test es nula,

$$\rho^2_{X_{1j}E_{1j}} = \left[\frac{\sigma_{T_{1j}E_{1j}} + \sigma^2_{E_{1j}}}{\sigma_{X_{1j}} \cdot \sigma_{E_{1j}}}\right]^2,$$

$$= \left[\frac{\sigma^2_{E_{1j}}}{\sigma_{X_{1j}} \cdot \sigma_{E_{1j}}}\right]^2,$$

$$= \frac{\sigma^2_{E_{1j}}}{\sigma^2_{X_{1j}}} \,.$$

Por lo tanto, como por la Conclusión 9 se tenía que

$$\rho_{X_{1j}X_{2j}} = 1 - \frac{\sigma^2_{E_{1j}}}{\sigma^2_{X_{1j}}},$$

resulta finalmente que

$$\rho_{X_{1j}X_{2j}} = 1 - \rho^2_{X_{1j}E_{1j}} \cdot \Diamond$$

Se ve que la correlación entre los scores observados de tests paralelos será perfecta cuando no haya correlación entre los scores observados y los errores.

Conclusión 11:

La correlación entre los scores observados de dos tests paralelos es igual al cuadrado de la correlación entre los scores observados y los verdaderos de uno de ellos.

En símbolos,

$$\rho_{X_{1j}X_{2j}} = \rho^2_{X_{1j}T_{1j}} \cdot \qquad \textbf{(2.18)}$$

Demostración:

La igualdad sigue de las Conclusiones 4) y 8). ◊

Conclusión 12:

La varianza del score verdadero es igual a la covarianza de los scores observados de dos tests estrictamente paralelos.

En símbolos,

$$\sigma^2_{T_{1j}} = \sigma_{X_{1j}X_{2j}} \cdot \qquad \textbf{(2.19)}$$

Demostración

Teniendo en cuenta la definición del coeficiente de correlación,

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma_{X_{1j}X_{2j}}}{\sigma_{X_{1j}} \cdot \sigma_{X_{2j}}},$$

pero por tratarse de test estrictamente paralelos, las varianzas de sus scores observados son iguales, de manera que

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma_{X_{1j}X_{2j}}}{\sigma^2_{X_{1j}}},$$

pero dado que la Conclusión 8 asegura que

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma^2_{T_{1j}}}{\sigma^2_{X_{1j}}},$$

comparando los numeradores de los segundos miembros se llega a que

$$\sigma^2_{T_{1j}} = \sigma_{X_{1j}X_{2j}}. \quad \Diamond$$

Conclusión 13:

La varianza del error es el producto de la varianza del score observado por uno menos la correlación entre los scores observados de dos tests estrictamente paralelos.

En símbolos,

$$\sigma^2_{E_{1j}} = \sigma^2_{X_{1j}} \cdot \left(1 - \rho_{X_{1j}X_{2j}}\right). \qquad \textbf{(2.20)}$$

Demostración:

De la expresión (2.10),

$$\sigma^2_{E_{1j}} = \sigma^2_{X_{1j}} - \sigma^2_{T_{1j}},$$

pero como por (2.15)

$$\sigma^2_{T_{1j}} = \sigma^2_{X_{1j}} \cdot \rho_{X_{1j}X_{2j}}$$

reemplazando en la expresión anterior y extrayendo factor común,

$$\sigma^2_{E_{1j}} = \sigma^2_{X_{1j}} .\left(1 - \rho_{X_{1j}X_{2j}}\right). \ \Diamond$$

Conclusión 14:

La correlación entre los scores verdaderos de dos test es igual al cociente entre la correlación entre los scores observados de ambos tests dividido en la raíz cuadrada del producto de las correlaciones entre los scores observados de dos formas paralelas de ambos tests.

En símbolos,

$$\rho_{T_{X_{1j}}T_{Z_{1j}}} = \frac{\rho_{X_{1j}Z_{1j}}}{\sqrt{\rho_{X_{1j}X_{2j}} \cdot \rho_{Z_{1j}Z_{2j}}}}. \qquad (2.21)$$

Demostración

Teniendo en cuenta que:

$$\begin{aligned} X_{1j} &= T_{X_{1j}} + E_{X_{2j}} \\ X_{2j} &= T_{X_{2j}} + E_{X_{2j}} \\ Z_{1j} &= T_{Z_{1j}} + E_{Z_{2j}} \\ Z_{2j} &= T_{Z_{2j}} + E_{Z_{2j}} \end{aligned}$$

donde X_{1j} y X_{2j} son scores de tests estrictamente paralelos, lo mismo que Z_{1j} y Z_{2j} .

Por definición de coeficiente de correlación,

$$\rho_{X_{1j}Z_{1j}} = \frac{\sigma_{X_{1j}Z_{1j}}}{\sigma_{X_{1j}} \cdot \sigma_{Z_{1j}}},$$

$$= \frac{\sigma_{\left(T_{X_{1j}} + E_{X_{1j}}\right)\left(T_{Z_{1j}} + E_{Z_{1j}}\right)}}{\sigma_{X_{1j}} \cdot \sigma_{Z_{1j}}},$$

donde la última igualdad se obtiene al aplicar la expresión (2.1). Pero por propiedad lineal de la covarianza:

$$\rho_{X_{1j}Z_{1j}} = \frac{\sigma_{T_{X_{1j}}T_{Z_{1j}}} + \sigma_{T_{X_{1j}}E_{Z_{1j}}} + \sigma_{T_{Z_{1j}}E_{X_{1j}}} + \sigma_{E_{X_{1j}}E_{Z_{1j}}}}{\sigma_{X_{1j}} \cdot \sigma_{Z_{1j}}},$$

pero dados los supuestos que:

$$\sigma_{T_{X_{1j}}E_{Z_{1j}}} = \sigma_{T_{Z_{1j}}E_{X_{1j}}} = \sigma_{E_{X_{1j}}E_{Z_{1j}}} = 0$$

la expresión anterior se reduce a

$$\rho_{X_{1j}Z_{1j}} = \frac{\sigma_{T_{X_{1j}}T_{Z_{1j}}}}{\sigma_{X_{1j}} \cdot \sigma_{Z_{1j}}},$$

de manera que,

$$\sigma_{T_{X_{1j}}T_{Z_{1j}}} = \rho_{X_{1j}Z_{1j}} \cdot \sigma_{X_{1j}} \cdot \sigma_{Z_{1j}}.$$

Por otra parte, nuevamente por definición de coeficiente de correlación,

$$\rho_{T_{X_{1j}}T_{Z_{1j}}} = \frac{\sigma_{T_{X_{1j}}T_{Z_{1j}}}}{\sigma_{T_{X_{1j}}} \cdot \sigma_{T_{Z_{1j}}}},$$

por lo cual si se reemplaza la expresión anterior en el numerador del segundo miembro se tiene que:

$$\rho_{T_{X_{1j}}T_{Z_{1j}}} = \frac{\rho_{X_{1j}Z_{1j}} \cdot \sigma_{X_{1j}} \cdot \sigma_{Z_{1j}}}{\sigma_{T_{X_{1j}}} \cdot \sigma_{T_{Z_{1j}}}},$$

que puede reescribirse como:

$$\rho_{T_{X_{1j}}T_{Z_{1j}}} = \frac{\rho_{X_{1j}Z_{1j}}}{\frac{\sigma_{T_{X_{1j}}}}{\sigma_{X_{1j}}} \cdot \frac{\sigma_{T_{Z_{1j}}}}{\sigma_{Z_{1j}}}},$$

en la que las fracciones del denominador se pueden expresar como:

$$\rho_{T_{X_{1j}} T_{Z_{1j}}} = \frac{\rho_{X_{1j} Z_{1j}}}{\sqrt{\rho_{X_{1j} X_{1j}}} \cdot \sqrt{\rho_{Z_{1j} Z_{1j}}}},$$

$$= \frac{\rho_{X_{1j} Z_{1j}}}{\sqrt{\rho_{X_{1j} X_{1j}} \cdot \rho_{Z_{1j} Z_{1j}}}}. \quad \Diamond$$

Dado que el denominador de la fracción es un número menor o igual que 1, se infiere que la correlación entre los scores verdaderos de dos tests será siempre mayor que la correlación entre los scores observados de dichos tests.

Decimos entonces que esta última correlación está *atenuada* en relación a la correlación de los scores verdaderos. Esta expresión se conoce como *corrección por atenuación.*

Conclusión 15:

Si un test se construye combinando N versiones paralelas de él, la varianza del score verdadero del test completo será N^2 veces la varianza del test original.

En símbolos,

$$\sigma^2_{T_X} = N^2 \cdot \sigma^2_{T_Y} . \qquad \textbf{(2.22)}$$

donde X es la suma de N scores de tests paralelos:

$$X = \sum_{i=1}^{N} Y_i$$

donde Y_i es uno de estos tests paralelos.

Demostración:

El score verdadero para el test combinado se obtiene al aplicar el operador esperanza sobre los scores observados del mismo,

$$T_X = \mathbf{E}(X),$$

pero reescribiendo X como una combinación lineal de tests aleatoriamente paralelos y por propiedades del operador esperanza,

$$T_X = \mathbf{E}\left(\sum_{i=1}^{N} Y_i\right),$$

$$= \sum_{i=1}^{N} \mathbf{E}(Y_i) = N \cdot T_Y ,$$

pero si restamos en ambos miembros la cantidad $\mathbf{E}(T_X)$ se tiene que:

$$T_X - \mathbf{E}(T_X) = N \cdot T_Y - \mathbf{E}(N \cdot T_Y),$$

$$= N \cdot T_Y - N \cdot \mathbf{E}(T_Y),$$

$$= N[T_Y - \mathbf{E}(T_Y)].$$

Elevando al cuadrado ambos miembros y tomando esperanza,

$$\mathbf{E}[T_X - \mathbf{E}(T_X)]^2 = N^2 \cdot \mathbf{E}[T_Y - \mathbf{E}(T_Y)]^2,$$

$$\sigma^2_{T_X} = N^2 \cdot \sigma^2_{T_Y} . \quad \Diamond$$

Conclusión 16:

La varianza del error del test combinado es N veces la varianza del error del test original.

En símbolos,

$$\sigma^2_{E_X} = N \cdot \sigma^2_{E_Y} . \qquad \textbf{(2.23)}$$

donde X e Y se definen igual que en la deducción anterior.

Demostración:

Por la expresión (2.1) se sabe que

$$E_X = X - T_X ,$$

y utilizando los resultados de la Conclusión anterior,

$$E_X = \sum_{i=1}^{N} Y_i - N \cdot T_Y ,$$

expresión en la que es posible aplicar nuevamente la igualdad (2.1) para Y_i:

$$E_X = \sum_{i=1}^{N} \left(T_Y + E_{Y_i} \right) - N \cdot T_Y .$$

Desarrollando la suma,

$$E_X = N \cdot T_Y + \sum_{i=1}^{N} E_{Y_i} - N \cdot T_Y$$

$$E_X = \sum_{i=1}^{N} E_{Y_i} .$$

Este resultado muestra que el score de error en X, es la suma de los scores de error de los Y´s. Tomando varianza en ambos miembros,

$$\sigma^2_{E_X} = \sum_{i=1}^{N} \sigma^2_{E_Y} + \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{E_{Y_i} E_{Y_j}} ,$$

$$= \sum_{i=1}^{N} \sigma^2_{E_Y} . \quad \Diamond$$

donde la última igualdad de logra debido a que todas las covarianzas entre los errores de los tests componentes de la combinación lineal son nulas.

Es importante destacar de las consecuencias 15) y 16) que la varianza del score verdadero crece más que la varianza del score de error para el test compuesto de N veces un test original, por lo cual puede pensarse que adicionando items a un test se aumenta la precisión de la medición.

Conclusión 17:

La correlación entre los scores observados entre dos tests paralelos, en los que cada uno de ellos resultan de la suma de N tests paralelos, se puede computar a partir de la correlación entre los scores de estos últimos de la siguiente forma:

En símbolos,

$$\rho_{X_{1j}X_{2j}} = \frac{N \cdot \rho_{Y_{1j}Y_{2j}}}{1+(N-1)\cdot \rho_{Y_{1j}Y_{2j}}}. \qquad (2.24)$$

donde X e Y se definen igual que en la conclusión anterior.

Demostración

Utilizando la expresión (2.15):

$$\rho_{X_{1j}X_{2j}} = \frac{\sigma^2_{T_{X_{1j}}}}{\sigma^2_{X_{1j}}},$$

en la que podemos utilizar la Conclusión 15 en el numerador y la varianza de una suma en el denominador:

$$\rho_{X_{1j}X_{2j}} = \frac{N^2 \cdot \sigma^2_{T_{Y_{1j}}}}{\sum_{i=1}^{N}\sigma^2_{Y_i} + \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{N}\sigma_{Y_iY_j}}_{i\neq k}},$$

$$= \frac{N^2 \cdot \sigma^2_{T_{Y_{1j}}}}{N\cdot\sigma^2_{Y_{1j}} + \left(N^2-N\right)\cdot\sigma^2_{T_{Y_{1j}}}},$$

$$= \frac{N^2 \cdot \rho_{Y_{1j}Y_{2j}} \cdot \sigma^2_{Y_{1j}}}{N\cdot\sigma^2_{Y_{1j}} + N(N-1)\cdot\rho_{Y_{1j}Y_{2j}}\cdot\sigma^2_{Y_{1j}}},$$

donde la última igualdad se obtiene al aplicar la Conclusión 8 a la varianza de los scores verdaderos del test *Y*. Finalmente, simplificando, resulta que:

$$\rho_{X_{1j}X_{2j}} = \frac{N \cdot \rho_{Y_{1j}Y_{1j}}}{1+(N-1)\cdot \rho_{Y_{1j}Y_{2j}}} \cdot \Diamond$$

Esta fórmula se conoce como la Fórmula de Spearman – Brown, donde *X, Y, N* se definen como en la deducción 15). Además la correlación ρ_{Y1jY2j} es la que existe entre cualquier par de tests paralelos que forman la combinación X y la correlación ρ_{X1jX2j} es la que existe entre los scores observados del test completo con una forma paralela.

Conclusión 18:

Conforme la longitud del test crece cada vez más, adicionando componentes paralelas, la correlación entre los scores observados de dos formas paralelas del test completo tiende a la unidad.

En símbolos,

$$\textbf{Si } \rho_{Y_{1j}Y_{2j}} \neq 0, \lim_{N\to\infty} \rho_{X_{1j}X_{2j}} = 1. \qquad \textbf{(2.25)}$$

donde X e Y se definen igual que en la conclusión anterior.

Demostración:

La expresión (2.24) puede escribirse como:

$$\rho_{XX'} = \frac{\rho_{YY'}}{\frac{1}{N} + \frac{(N-1)}{N}\cdot \rho_{YY'}} = \frac{\rho_{YY'}}{\frac{1}{N} + \frac{(N-1)}{N}\cdot \rho_{YY'}}$$

y tomando límite cuando N tiende a infinito en esta última expresión,

$$\underbrace{\lim}_{N\to\infty} \frac{\rho_{YY'}}{\frac{1}{N} + \frac{(N-1)}{N}\cdot \rho_{YY'}} = \frac{\rho_{YY'}}{\rho_{YY'}} = 1. \Diamond$$

Debemos notar que como ya habíamos establecido que esta correlación también era igual a uno menos el cociente entre la varianza del error sobre la varianza de los scores observados, deducimos que conforme la longitud del test tiende a infinito, la varianza del error tiende a 0.

CAPITULO 3

Análisis de Confiabilidad

La consistencia de los resultados de un test puede pensarse como la posibilidad de replicar los scores observados si los mismos individuos fueran expuestos al mismo test en idénticas condiciones en diferentes ocasiones.

Como quedará de manifiesto más adelante, existen varias acepciones para el concepto de *Confiabilidad de un test*, pero podríamos caracterizarlo genéricamente como el grado en el cual los scores individuales se mantienen relativamente consistentes en distintas administraciones del mismo test a lo largo del tiempo o bien de formas alternas de ese test.

A partir del sencillo modelo propuesto por la Teoría Clásica del Score Verdadero, que adjudica cada observación a la suma de dos componentes, el score verdadero (inobservable) y una componente aleatoria que representa el conjunto de todas las fuentes de errores no sistemáticos, es posible aproximarse a la interpretación del concepto de confiabilidad del test. La cuestión central consiste en determinar el grado en el cual los errores de medición, expresados en la componente aleatoria, afectan o influencian los scores observados.

En el capítulo anterior hemos introducido una distinción entre errores de medición sistemáticos y no sistemáticos, determinando que sólo estos últimos integran el término aleatorio. Los errores de medición sistemáticos, que afectan las observaciones siempre en la misma dirección si bien producen una disminución de la precisión (y por lo tanto de la utilidad) de las observaciones, no provocan su inconsistencia, a diferencia de los errores no sistemáticos, cuyos efectos no muestran una dirección privilegiada en el sentido que pueden alterar las observaciones positiva o negativamente, que no sólo disminuyen la precisión de los datos recolectados sino también le restan consistencia, cuestión que pone en juego la utilidad de éstos.

Este capítulo se dedica por completo al *Análisis de Confiabilidad* de un test, en función de la importancia decisiva que debe asignarse a la responsabilidad de contar con datos confiables como condición absolutamente necesaria para construir inferencias verdaderas. Se inicia con la definición del Indice de Confiabilidad y del Coeficiente de Confiabilidad, insistiendo en la interpretación de sus posibles valores. Luego se da lugar a los diferentes procedimientos para estimar la confiabilidad de un test distinguiendo entre métodos que requieren de una o dos aplicaciones y se finaliza con comentarios acerca de los factores prácticos que afectan la confiabilidad de una prueba y que deben tenerse en cuenta en cualquier trabajo de campo.

3.1. EL INDICE DE CONFIABILIDAD

El *Indice de Confiabilidad* de un test se define como el coeficiente de correlación ρ_{XT} entre los scores verdaderos y observados.

En la Conclusión 4 del capítulo anterior, hemos visto que la correlación entre el score X y el componente T cumple:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2},$$

y por lo tanto (teniendo presente que los desvíos estándares son cantidades positivas):

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X}. \qquad \textbf{(3.1)}$$

De esta forma, el *Indice de Confiabilidad* se puede definir también como el cociente entre los desvíos estándares de los scores verdadero y observado.

En la práctica, esta definición tiene poco valor por cuanto los scores verdaderos son inobservables, lo que imposibilita el cómputo directo de su varianza.

Sin embargo es posible establecer una relación entre el coeficiente de correlación ρ_{XT} y $\rho_{XX'}$ el coeficiente de correlación entre los scores observados de dos formas paralelas del test. Esto conduce a una definición del *Coeficiente de Confiabilidad.*

3.2. EL COEFICIENTE DE CONFIABILIDAD

3.2.1. DEFINICION

El *Coeficiente de Confiabilidad* de un test se define como la correlación $\rho_{XX'}$ entre los scores observados X y X´ de dos tests paralelos.

Pero la Conclusión 11 del capítulo anterior muestra que este coeficiente de correlación se puede definir como la varianza del score verdadero sobre la varianza del score observado.

De esta forma podemos establecer ahora una relación matemática entre el Indice de Confiabilidad y el Coeficiente de Confiabilidad. Como por la expresión (3.1) se tiene que:

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X},$$

y por la Conclusión 11:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}, \qquad \textbf{(3.2)}$$

por lo que resulta:

$$\therefore \rho_{XX'} = \rho_{XT}^2 \,. \qquad \textbf{(3.3)}$$

De esta forma el coeficiente de confiabilidad es el cuadrado del índice de confiabilidad.

Supongamos que se establece que el coeficiente de confiabilidad de un test es 0.90, entonces podríamos interpretar este valor indicando que:

- ✓ el 90% de la variabilidad de los scores observados puede atribuirse a la variabilidad de los verdaderos scores en ese grupo de alumnos, es decir:

$$\sigma_T^2 = 0{,}90 \cdot \sigma_X^2$$

- ✓ Un 81% $[(0{,}90)^2]$ de la varianza de los scores del segundo test puede explicarse a partir de la varianza de las puntuaciones del primer test.

✓ la correlación entre los scores observados y los verdaderos es $\sqrt{0,90}$ = 0,9487 para este grupo de alumnos

3.2.2. INTERPRETACION

Es interesante notar que de acuerdo a las definiciones dadas, el coeficiente de confiabilidad será un número comprendido en el intervalo [0 ; 1]. Esto se puede ver a partir de que:

$$X = T + E \quad \Rightarrow \quad \sigma_X^2 \geq \sigma_T^2$$

por lo cual resulta que:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} \leq 1,$$

y que:

✓ Cuando $\rho_{XX'} = 1$:

- Las mediciones se hacen sin error ($E = 0$).
- Para cada alumno el score observado coincide con el score verdadero, es decir, en términos probabilísticos $P(X=T)=1$.
- Toda la varianza del score observado refleja la varianza del score verdadero ($\sigma^2_X = \sigma^2_T$).
- Todas las diferencias entre los scores observados representan las diferencias entre los verdaderos puntajes.
- La correlación entre los scores observados y verdaderos es 1 ($\rho_{XT} = 1$).
- La correlación entre los scores observados y los errores es 0 ($\rho_{XE} = 0$).

✓ Cuando $\rho_{XX'} = 0$:

- Las mediciones sólo incluyen errores de medición.
- Para cada alumno ocurre que $X = E$.

- Toda la varianza del score observado refleja la varianza del error ($\sigma^2_X = \sigma^2_E$).
- Todas las diferencias entre los scores observados representan sólo errores aleatorios.
- La correlación entre los scores observados y verdaderos es 0 ($\rho_{XT} = 0$).
- La correlación entre los scores observados y los errores es 1 ($\rho_{XE} = 1$).

✓ Cuando $0 < \rho_{XX'} < 1$:

- Las mediciones contienen una componente aleatoria de error.
- $X=T+E$ para cada alumno.
- Toda la varianza del score observado refleja, en parte, la varianza del score verdadero y en parte la varianza del error:

$$\sigma^2_X = \sigma^2_T + \sigma^2_E$$

- Todas las diferencias entre los scores observados representan las diferencias entre los verdaderos puntajes y también errores aleatorios.
- La correlación entre los scores observados y verdaderos es $\rho_{XT} = \sqrt{\rho_{XX'}}$
- La correlación entre los scores observados y los errores es $\rho_{XE} = \sqrt{1 - \rho_{XX'}}$
- El coeficiente de confiabilidad es la proporción de la varianza del score observado que corresponde a la varianza del score verdadero:

$$\rho_{XX'} = \sigma^2_T / \sigma^2_X$$

- Cuanto más cercano a 1 sea el coeficiente $\rho_{XX'}$, más confiablemente podemos estimar T a partir de X, puesto que la varianza del error será más pequeña.
- Las relaciones indicadas entre el coeficiente de confiabilidad $\rho_{XX'}$ y las cantidades ρ_{XT} y ρ_{XE} se ilustran en los gráficos siguientes:

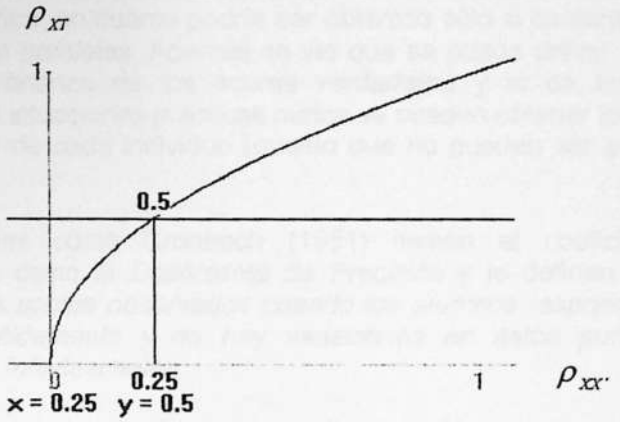


Figura 3.1. Relación entre ρ_{XT} y $\rho_{XX'}$

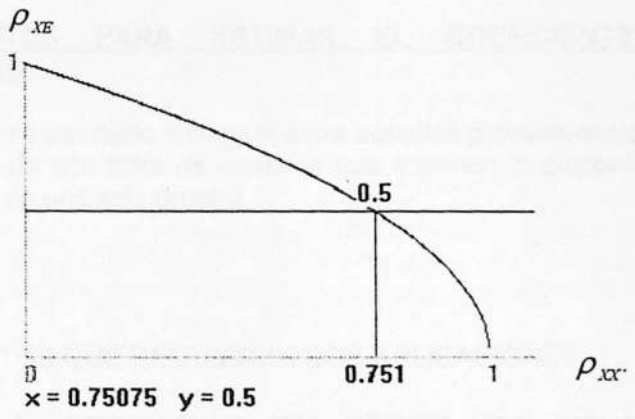


Figura 3.2. Relación entre ρ_{XE} y $\rho_{XX'}$

3.2.3. EL COEFICIENTE DE PRECISION

Es importante notar en este momento que el coeficiente de confiabilidad es una cantidad teórica, en cuanto podría ser obtenido sólo si contáramos con formas *estrictamente paralelas.* Además se vio que se puede definir como un cociente entre la varianza de los scores verdaderos y la de los scores observados, pero en situaciones prácticas nunca se pueden obtener los valores del score verdadero de cada individuo (puesto que no pueden ser evaluados infinitas veces).

Algunos autores como Cronbach (1951) llaman al coeficiente de confiabilidad teórico como el *Coeficiente de Precisión* y lo definen como *la correlación entre los scores observados cuando los alumnos responden a los mismos items repetidamente y no hay variaciones en estos puntajes en intervalos de tiempo infinitesimales.*

Es claro entonces que en situaciones prácticas, sólo podemos *estimar* el coeficiente de precisión mediante una muestra de individuos (extraída de una población de individuos) que responden a una muestra de items (que representan un universo de items) de un test.

3.3. PROCEDIMIENTOS PARA ESTIMAR EL COEFICIENTE DE CONFIABILIDAD

En este apartado es oportuno distinguir entre aquellos procedimientos que suponen la aplicación de dos tests de aquellos que suponen la disponibilidad de scores observados en una sola prueba.

3.3.1. PROCEDIMIENTOS QUE REQUIEREN DOS APLICACIONES

A continuación se describen los tres métodos más comúnmente utilizados:

3.3.1.1. ***Método de las Formas Alternativas (o Alternas) – El Coeficiente de Equivalencia***

La imposibilidad de disponer en la práctica de formas *estrictamente paralelas* para un test conduce a la necesidad de definir las *formas alternativas.*

Dos *formas alternativas* de un test son dos formas que se construyen en un esfuerzo por obtener formas paralelas y deben presentar valores similares de las medias, varianzas y correlaciones con otras variables, de sus scores observados.

En otros términos las formas alternativas son dos muestras aleatorias de items extraídas del mismo *dominio de contenidos* que se establezca.

Una forma de estimar el Coeficiente de Confiabilidad requiere la construcción de dos formas alternativas del test y la administración de ambas formas al mismo grupo de alumnos. La estimación del coeficiente de confiabilidad para este caso se obtiene computando el coeficiente de correlación entre los scores obtenidos por cada alumno en ambas formas alternativas y este número de conoce como *Coeficiente de Equivalencia*.

El Coeficiente de Equivalencia será entonces una medida, tanto de la confiabilidad del test como del "paralelismo" de las formas. Cuando más alto sea el valor del coeficiente de Equivalencia, más confiados podemos estar que los puntajes de ambas formas son *intercambiables*.

Los tests construídos para evaluar logros o aptitudes generalmente tienen múltiples formas, dado que por lo general se prevé el hecho que el individuo tenga la oportunidad de "re-tomar" la prueba.

Es claro que problemas en la aplicación de las pruebas, cambios temporarios en el estado del alumno, diferencias en las condiciones ambientales, etc. influirán en los scores observados. Sin embargo, una cuestión de suma importancia es advertir que la principal fuente de error de medición proviene de *posibles diferencias en los contenidos* de ambas formas del test.

Es interesante advertir que las formas alternativas pueden no ser paralelas, pero si son *funciones lineales de formas paralelas*, la correlación entre ellas es igual a la correlación entre formas paralelas. Supongamos que:

$$X = T_x + E_x ,$$

$$X' = T_{x'} + E_{x'}$$

donde X y X´ son scores de tests paralelos. Si tomamos:

$$Z = a X' + b ,$$

vemos que claramente Z y X *no son puntajes de tests paralelos,* puesto que sus scores verdaderos no coinciden:

$$Z = a \cdot (T_{X'} + E_{X'}) + b,$$

$$= \underbrace{(aT_{X'} + b)}_{T_Z} + \underbrace{aE_{X'}}_{E_Z}$$

$$= T_Z + E_Z,$$

de donde se ve que:

$$T_Z = a\,T_{X'} + b$$

pero dado que $T_X = T_{X'}$ por ser tests estrictamente paralelos,

$$T_Z = a\,T_X + b.$$

$$T_Z \neq T_X.$$

Además, por ser X y X' scores de tests estrictamente paralelos,

$$\sigma^2_{E_X} = \sigma^2_{E_{X'}},$$

por lo que resulta que:

$$\sigma^2_{E_Z} = a^2\,\sigma^2_{E_{X'}},$$

$$\sigma^2_{E_Z} \neq \sigma^2_{E_X}.$$

es decir que si $a \neq \pm 1$, las varianzas de sus errores tampoco son iguales. Sin embargo aún tenemos que $\rho_{XZ} = \rho_{XX'}$ puesto que el coeficiente de correlación es *invariante* ante transformaciones lineales para valores positivos de *a*:

$$\rho_{XZ} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z},$$

$$= \frac{\mathbf{E}(X - T_X) \cdot (Z - T_Z)}{\sigma_X \sigma_Z},$$

$$= \frac{\mathbf{E}(X - T_X) \cdot [(aX' + b) - (aT_{X'} + b)]}{\sigma_X \left(\sqrt{a^2 \cdot \sigma^2_{X'}}\right)},$$

$$\rho_{XZ} = \frac{a \cdot \mathbf{E}(X - T_X) \cdot (X' - T_{X'})}{|a| \cdot \sigma_X \cdot \sigma_{X'}},$$

$$= \rho_{XX'} \ .$$

Un problema típico asociado a este método es la decisión que debe tomarse en relación al lapso de tiempo que debe transcurrir entre las aplicaciones. Pequeños intervalos de tiempo entre ellas, permitirían la presencia de efectos indeseables como la memorización de los items mientras que períodos demasiado largos podrían permitir cambios en las competencias que se pretenden medir. Las dos formas deberían administrarse dentro de un intervalo de tiempo relativamente corto, permitiendo a los individuos que toman el test disponer sólo del tiempo necesario para no encontrarse fatigados.

Una sugerencia razonable que se encuentra frecuentemente en la bibliografía es la de balancear el orden de aplicación de las dos formas, por lo que usualmente una mitad del grupo, seleccionada aleatoriamente, es asignada a la forma 1 seguida de la forma 2, mientras que la otra mitad es asignada a la forma 2 en primer lugar seguido de la forma 1. Esta estrategia contribuye al control de factores indeseados que pueden afectar a las observaciones.

Finalmente, aunque no hay reglas para decidir cuál es un valor mínimo aceptable para el coeficiente de confiabilidad (equivalencia) en este caso, generalmente se reportan coeficientes estimados entre 0,80 y 0,90, acompañados de los valores de las medias, desvíos estándares y errores estándares de medición para cada forma, valores que deben ser próximos entre sí.

3.3.1.2. ***Método Test – Retest – El Coeficiente de Estabilidad***

En muchas ocasiones, el interés del investigador reside en analizar la consistencia de las respuestas de los individuos a *una forma única* de un test cuando ésta se administra en diferentes instantes de tiempo.

A diferencia del método de las formas alternas, aquí la principal fuente para el error de medición no está en potenciales diferencias en los contenidos de las pruebas (puesto que se trata de un mismo test) sino en las fluctuaciones de los scores observados del individuo alrededor de su verdadera puntuación debido a cambios temporarios en el estado del individuo.

En este caso, se aplica una prueba a un grupo de individuos, se espera y se re-administra la misma prueba al mismo grupo.

El coeficiente de confiabilidad se estima calculando el coeficiente de correlación entre los scores observados de cada alumno en ambos ensayos. Este número se conoce como el *Coeficiente de Estabilidad.*

Hay pocos valores que se toman como *estándares* para el coeficiente de estabilidad. Cuando se aplica el método con tests de inteligencia se han reportado valores entre 0,70 y 0,90, pero con tests referidos a personalidad o actitud, los valores suelen ser más bajos.

Es importante destacar que cuando se evalúa la magnitud del coeficiente de estabilidad deben tenerse en cuenta factores como el tiempo transcurrido entre pruebas así como la edad de los individuos y la naturaleza de la competencia teórica que se mide:

- ✓ Como en el caso anterior no es sencillo decidir cuánto tiempo debe transcurrir entre los ensayos. Este tiempo debe ser suficientemente largo como para eliminar efectos como la memoria o aprendizajes, pero suficientemente corto como para evitar cambios en la competencia que se intenta medir.
- ✓ El uso ulterior de los scores del test debe ser un aspecto a considerar: si, por ejemplo, se trata de pruebas sobre el desarrollo motriz de un niño pequeño en orden a decidir una intervención médica o psicológica, es claro que no se presentarán problemas de aprendizaje en pruebas repetidas en intervalos de tiempos cortos, pero las condiciones cambiarían por la maduración del niño si se tratara de intervalos temporales prolongados.

3.3.1.3. ***Método Test – Retest con Formas Alternativas – El Coeficiente de Estabilidad y Equivalencia***

Simplemente se trata de una combinación de los dos métodos anteriores. En este caso se procede a administrar la forma 1, se espera y se administra la forma 2, teniendo en cuenta de invertir el orden de aplicación de las formas para la mitad del grupo. El coeficiente de correlación en este caso se conoce como el *Coeficiente de Estabilidad y Equivalencia.*

En esta situación, en relación a las fuentes principales para el error se deben considerar tanto las diferencias procedentes de los contenidos de las pruebas como las fluctuaciones en el comportamiento del individuo a lo largo del tiempo, además de otras circunstancias como las descriptas precedentemente.

3.3.2. PROCEDIMIENTOS QUE REQUIEREN UNA SOLA APLICACION

Las típicas evaluaciones en un curso de una escuela son ejemplos de situaciones en las sólo se aplica una vez un test a un grupo de individuos.

Aunque la cuestión de fondo continúa siendo el análisis de la medida en que los scores observados de los alumnos reflejan sus verdaderos scores, el interés del investigador ya no radica en el desempeño de los alumnos en los items que constituyen el test, sino más bien en la medida en que se pueden *generalizar* los resultados, a partir de los items incluidos en la prueba, sobre un dominio de contenidos mucho más general que podrían haber sido incluidos en el test.

En consecuencia, el análisis de la consistencia de los resultados ya no se construye juzgando la equivalencia de los contenidos muestreados o la estabilidad de los scores a lo largo del tiempo sino que ahora se trata de investigar la *consistencia interna* de los items de la prueba, que puede tomarse como una medida de cuán consistentemente se pueden generalizar los resultados obtenidos, a un dominio de contenidos más extenso. En otras palabras no interesa ya la consistencia del test completo sino la de los items que lo integran.

Los procedimientos para estimar la confiabilidad en estos casos se conocen como *Métodos de consistencia interna*.

Debería notarse que la principal fuente de error deriva de posibles diferencias en la *muestra* de contenidos contemplados por los items del test en relación a aquellos contenidos que no han sido incluidos en la evaluación, además de otras fuentes de errores que también pueden afectar las puntuaciones observadas como fallas en la administración de la prueba, en su corrección, fluctuaciones en el estado del individuo, etc.

Cuando los individuos se comportan consistentemente a través de los items del test, se dice que éste tiene *homogeneidad en los items*, en el sentido que se acepta que los items representan el mismo dominio de contenidos.

A continuación se describen los métodos más populares que suponen la división del test en N componentes, siendo N un número mayor o igual que 2 y también un método propuesto por Hoyt (1941) que hace uso del Análisis de la Varianza.

3.3.2.1. ***Método de la División del Test en Componentes***

En este caso el test se divide en *N* partes o componentes, que se seleccionan de modo que tales partes sean *paralelas* o *esencialmente* τ *equivalentes*.

En este contexto:

- ✓ Si los componentes del test son formas paralelas se puede aplicar la *Fórmula de Spearman – Brown.*
- ✓ Si los componentes son *esencialmente* τ *equivalentes*, se puede calcular una cantidad conocida como el *Coeficiente* α *de Cronbach.*
- ✓ Si los componentes *no son esencialmente* τ *equivalentes*, se puede obtener una *cota inferior* utilizando el Coeficiente α de Cronbach.

3.3.2.1.1. *LA FÓRMULA DE SPEARMAN - BROWN*

La Fórmula de Spearman-Brown se introdujo ya en el capítulo anterior, en la Conclusión 17 del capítulo previo. Si un test se asume como una *combinación de N componentes paralelos*, su confiabilidad se puede expresar en función de la confiabilidad de sus componentes $\rho_{YY'}$ como:

$$\rho_{XX'} = \frac{N \cdot \rho_{YY'}}{\left[1 + (N-1) \cdot \rho_{YY'}\right]}. \qquad \textbf{(3.4)}$$

Debemos notar que el coeficiente de confiabilidad del test completo siempre será mayor o igual que el coeficiente de confiabilidad de cualquiera de sus componentes, motivo por el cual se conoce como el *coeficiente de confiabilidad escalado:*

$$\rho_{XX'} = \frac{N \cdot \rho_{YY'}}{1 + (N-1) \cdot \rho_{YY'}},$$

$$\rho_{XX'} \cdot \left[1 + (N-1) \cdot \rho_{YY'}\right] = N \cdot \rho_{YY'},$$

$$\rho_{XX'} + N \cdot \rho_{XX'} \cdot \rho_{YY'} - \rho_{XX'} \cdot \rho_{YY'} = N \cdot \rho_{YY'},$$

de lo cual se deduce que:

$$\rho_{XX'} = N \cdot \rho_{YY'} - N \cdot \rho_{XX'} \cdot \rho_{YY'} + \rho_{XX'} \cdot \rho_{YY'},$$
$$= \rho_{YY'} \cdot (N - N \cdot \rho_{XX'} + \rho_{XX'}).$$

Pero se puede demostrar que la cantidad entre paréntesis es un número mayor o igual que 1. Dado que *N* es un número mayor o igual que 1,

$$N \geq N - 1 \geq \rho_{XX'} \cdot (N - 1),$$

puesto que $\rho_{XX'}$ es un número mayor o igual que 0 y menor o igual que 1. Entonces la expresión anterior se puede escribir como:

$$1 \geq -N + \rho_{XX'} \cdot (N - 1),$$
$$1 \leq N - N \cdot \rho_{XX'} + \rho_{XX'},$$

razón por la cual se deduce que:

$$\rho_{XX'} \geq \rho_{YY'}$$

como se quería demostrar.

- ***Caso Particular*: Dos Componentes**

Cuando el test se considera compuesto de sólo dos partes, se conoce como *División por mitades* (*Split Half Method*), la expresión anterior se reduce a:

$$\rho_{XX'} = \frac{2 \cdot \rho_{YY'}}{1 + \rho_{YY'}}, \qquad \textbf{(3.5)}$$

donde $\rho_{YY'}$ es el coeficiente de confiabilidad de cada mitad.

La tabla siguiente muestra la relación entre la confiabilidad de una mitad con la confiabilidad de test completo:

$\rho_{YY'}$	$\rho_{xx'}$
0,00	0,00
0,20	0,33
0,40	0,57
0,60	0,75
0,80	0,89
1,00	1,00

Tabla 3.1.: Confiabilidad del test combinado $\rho_{xx'}$ a partir de la confiabilidad de una mitad $\rho_{YY'}$

La relación admite la representación gráfica de la Figura 3.3.

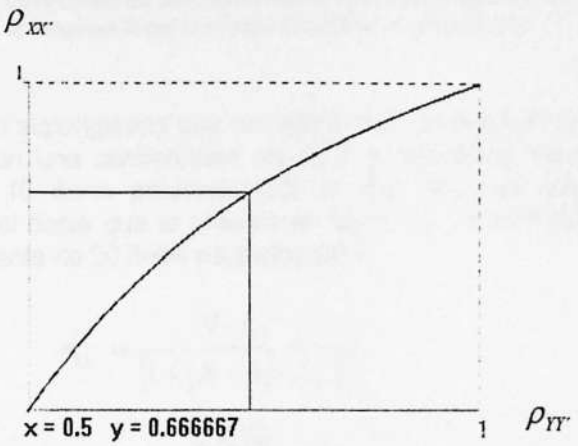


Figura nº 3.3: El coeficiente de confiabilidad del test combinado en función del coeficiente de confiabilidad de una mitad.

- ***Otros usos de la fórmula de Spearman-Brown:***

1. *Efecto sobre la confiabilidad de un cambio en la cantidad de componentes de un test*

A partir de la fórmula general de Spearman-Brown (para k componentes) podemos inferir que a medida que se adicionan subtest paralelos, la combinación resultante será cada vez más confiable, aunque, los mayores aumentos en el coeficiente de confiabilidad del test completo se obtienen para valores bajos de k. Dicho efecto aparece con nitidez en el siguiente gráfico:

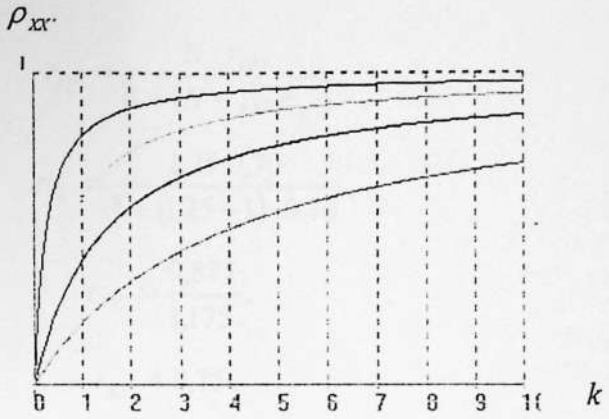


Figura nº 3.4: El coeficiente de confiabilidad del test combinado $\rho_{XX'}$ en función de la longitud N del test para distintos valores de $\rho_{YY'}$

Como ejemplo supongamos que contamos con un test compuesto de 10 items paralelos, con una confiabilidad de 0,70 y decidimos incrementar su longitud en otros 10 items paralelos (por lo que N=2, ya que la nueva combinación será el doble que la primera en tamaño). La confiabilidad de la combinación resultante de 20 items paralelos será:

$$r_{XX'} = \frac{N \cdot r_{YY'}}{[1 + (N-1) \cdot r_{YY'}]},$$

$$= \frac{2.0,70}{1 + (2-1) \cdot 0,70},$$

$$= \frac{1,40}{1,70},$$

$$= 0,82.$$

donde $r_{XX'}$ es una estimación de $\rho_{XX'}$. Al pasar de una confiabilidad de 0,70 a 0,82 hemos logrado un aumento de 17,65%.

Por otra parte supongamos que tenemos una combinación de 40 items con una confiabilidad de 0,70 y decidimos también añadir 10 items paralelos. El nuevo test compuesto por 50 items (N=1,25) tendrá una confiabilidad de:

$$r_{XX'} = \frac{N \cdot r_{YY'}}{[1 + (N-1) \cdot r_{YY'}]},$$

$$r_{XX'} = \frac{1{,}25.0{,}70}{1 + (1{,}25 - 1) \cdot 0{,}70},$$

$$r_{XX'} = \frac{0{,}875}{1{,}175},$$

$$r_{XX'} = 0{,}75.$$

En esta situación, al pasar de una confiabilidad de 0,70 a 0,75, el aumento logrado es sólo de un 6,4%, bastante menor al 17,65% anterior.

Como se recordará, ya se probó que a medida que el número de componentes paralelos tiende a infinito, la confiabilidad del test tiende a 1.

2. Cálculo de la confiabilidad de un subtest de longitud 1/N a partir del test completo

Si el interés reside en lograr una versión abreviada del test, puede resultar útil contar con el coeficiente de confiabilidad del subtest. Esto puede hacerse a partir de la fórmula de Spearman-Brown:

$$\rho_{YY'} = \frac{\frac{1}{N} \cdot \rho_{XX'}}{1 + \left(\frac{1}{N} - 1\right) \cdot \rho_{XX'}}. \qquad \textbf{(3.6)}$$

A manera de ejemplo de aplicación supongamos que se tiene un test compuesto de 20 items paralelos con una confiabilidad de 0,90 y deseamos eliminar cinco de ellos para tener una versión abreviada del test ¿Cuál sería la confiabilidad del nuevo test integrado solo por los quince items restantes?

$$r_{YY'} = \frac{\frac{1}{N} \cdot r_{XX'}}{1 + \left(\frac{1}{N} - 1\right) \cdot r_{XX'}},$$

$$= \frac{\frac{3}{4} \cdot 0{,}90}{1 + \left(\frac{3}{4} - 1\right) \cdot 0{,}90},$$

$$= 0{,}87.$$

En consecuencia, el test compuesto de los quince items restantes tendrá una confiabilidad de 0,87. Como era de esperar el test abreviado presenta una confiabilidad menor que la del test completo.

3. Cálculo de la longitud necesaria de una combinación para lograr una confiabilidad predeterminada

Nuevamente, a partir de la fórmula de Spearman-Brown se tiene que:

$$N = \frac{\rho_{XX'} \cdot (1 - \rho_{YY'})}{\rho_{YY'} \cdot (1 - \rho_{XX'})}. \qquad \textbf{(3.7)}$$

Como ejemplo supongamos que tenemos un test compuesto de cinco items con una confiabilidad de 0,6. ¿Cuántos items deben agregarse para obtener una confiabilidad de 0,85?

$$N = \frac{r_{XX'} \cdot (1 - r_{YY'})}{r_{YY'} \cdot (1 - r_{XX'})},$$

$$= \frac{0{,}85 \cdot (1 - 0{,}6)}{0{,}6 \cdot (1 - 0{,}85)},$$

$$= 3{,}78.$$

Es decir debería tener una longitud 3,78 veces la longitud de test original, esto es alrededor de 19 items (3,78 . 5 = 18,88).

3.3.2.1.2. *EL COEFICIENTE α DE CRONBACH (1951)*

- ***Derivación del Coeficiente α de Cronbach***

Supongamos que los componentes del test no tienen todos la misma varianza o existe alguna indicación que los subtests no pueden considerarse paralelos. En tal circunstancia no es lícito aplicar la fórmula de Spearman-Brown.

Se presenta ahora una medida de confiabilidad de un test que sólo requiere el conocimiento de la varianza del test completo y de las varianzas de los componentes.

Comencemos suponiendo que los scores observados del test X se consideran como la suma de N subtests paralelos. La varianza del score verdadero del test completo puede expresarse como:

$$\begin{aligned}\sigma^2_{T_X} &= \sigma^2_{T_A} + \sigma^2_{T_B} + \ldots + \sigma^2_{T_N} + \sum_{i \neq j} \rho_{T_i T_j} \sigma_{T_i} \sigma_{T_j}, \\ &= \sigma^2_{T_A} + \sigma^2_{T_B} + \ldots + \sigma^2_{T_N} + \sum_{i \neq j} \sigma_{T_i T_j}, \\ &= N \cdot \sigma^2_{T_i} + N \cdot (N-1) \cdot \sigma_{T_i T_j}. \end{aligned} \qquad \textbf{(3.8)}$$

En la expresión anterior, σ_{TiTj} es la covarianza entre los verdaderos scores de dos subtests paralelos *i* y *j* cualesquiera, que es idéntica para todos los pares que se consideren por ser paralelos. Debe notarse que por esta misma razón también coinciden las varianzas de los scores verdaderos de cada subtest.

Pero podemos mostrar que para dos tests *paralelos i* y *j*, la varianza del score verdadero de uno cualquiera de ellos es igual a la covarianza entre los scores verdaderos de ambos tests.

$$\begin{aligned}\sigma_{T_i T_j} &= \mathrm{E}\left[\left(T_i - \mu_{T_i}\right) \cdot \left(T_j - \mu_{T_j}\right)\right], \\ &= \mathrm{E}\left[\left(T_i - \mu_{T_i}\right)^2\right], \\ &= \sigma^2_{T_i}. \end{aligned} \qquad \textbf{(3.9)}$$

Por otro lado sabemos también que, en general, para dos test *cualesquiera* i, j, la covarianza de los scores verdaderos de ambos test es igual a la covarianza entre sus scores observados:

$$\sigma_{T_i T_j} = \sigma_{x_i x_j} . \tag{3.10}$$

Una notación abreviada para $\sigma_{X_i X_j}$ es σ_{ij}. Este resultado se muestra tomando:

$$\begin{aligned} \sigma_{ij} &= \mathbf{E}\left[\left(X_i - \mu_{X_i}\right)\cdot\left(X_j - \mu_{X_j}\right)\right], \\ &= \mathbf{E}\left[\left(T_i + E_i - \mu_{X_i}\right)\cdot\left(T_j + E_j - \mu_{X_j}\right)\right], \end{aligned}$$

Si de distribuye el producto y se tiene en cuenta que los scores verdaderos no están correlacionados con los errores y además:

$$\mu_{X_i} = \mu_{T_i} \quad ; \quad \mu_{X_j} = \mu_{T_j}$$

resulta que:

$$\begin{aligned} \sigma_{ij} &= \mathbf{E}\left[\left(T_i - \mu_{T_i}\right)\cdot\left(T_j - \mu_{T_j}\right)\right], \\ &= \sigma_{T_i T_j} . \end{aligned}$$

Reemplazando (3.9) en (3.8) y utilizando (3.10), tenemos que:

$$\begin{aligned} \sigma^2_{T_X} &= N\cdot\sigma^2_{T_i} + N\cdot(N-1)\cdot\sigma_{T_i T_j} , \\ &= N\cdot\sigma_{T_i T_j} + N\cdot(N-1)\cdot\sigma_{T_i T_j} , \\ &= N^2\cdot\sigma_{T_i T_j} , \\ &= N^2\cdot\sigma_{ij} . \end{aligned} \tag{3.11}$$

Utilizando (3.11) el coeficiente de confiabilidad de la combinación lineal se puede expresar como:

$$\rho_{XX'} = \frac{\sigma^2_{T_X}}{\sigma^2_X}$$

$$\rho_{XX'} = \frac{N^2 \cdot \sigma_{ij}}{\sigma^2_X} \qquad \textbf{(3.12)}$$

Pero esta expresión sólo es válida *bajo el supuesto que todos los componentes son estrictamente paralelos.*

En situaciones reales, nunca se puede asegurar que los tests sean paralelos. Sin embargo es posible encontrar una expresión que permita estimar una cota inferior para el coeficiente de confiabilidad, conocida como α *de Cronbach.*

Un resultado previo que puede demostrarse es que la suma de las varianzas de los scores verdaderos de N subtests no necesariamente paralelos, será mayor o igual que la suma de sus N (N-1) covarianzas dividida en (N-1), esto es:

$$\sum \sigma^2_{T_i} \geq \frac{\sum_{i \neq j} \sigma_{ij}}{N-1}$$

Si sumamos a ambos miembros la suma de las covarianzas, tenemos que:

$$\sum \sigma^2_{T_i} + \sum_{i \neq j} \sigma_{ij} \geq \frac{\sum_{i \neq j} \sigma_{ij}}{N-1} + \sum_{i \neq j} \sigma_{ij} \,,$$

$$\sum \sigma^2_{T_i} + \sum_{i \neq j} \sigma_{ij} \geq \frac{\sum_{i \neq j} \sigma_{ij} + (N-1) \cdot \sum_{i \neq j} \sigma_{ij}}{N-1} \,,$$

$$\sum \sigma^2_{T_i} + \sum_{i \neq j} \sigma_{ij} \geq \frac{N \cdot \sum_{i \neq j} \sigma_{ij}}{N-1} \,, \qquad \textbf{(3.13)}$$

donde la suma del segundo miembro de la desigualdad recorre N · (N-1) covarianzas de subtests no necesariamente paralelos. Además reconocemos que la suma del primer miembro es simplemente la varianza de los scores verdaderos del test completo:

$$\sum \sigma_{T_i}^2 + \sum_{i \neq j} \sigma_{ij} = \sigma_{T_X}^2 .$$

A partir de estos resultados, dividiendo ambos miembros de (3.13) por la varianza de los scores observados de la combinación lineal se tiene que:

$$\sigma_{T_X}^2 \geq \frac{N}{N-1} \cdot \sum_{i \neq j} \sigma_{ij} ,$$

$$\frac{\sigma_{T_X}^2}{\sigma_X^2} \geq \frac{N}{N-1} \cdot \left(\frac{\sum_{i \neq j} \sigma_{ij}}{\sigma_X^2} \right),$$

$$\frac{\sigma_{T_X}^2}{\sigma_X^2} \geq \frac{N}{N-1} \cdot \left(\frac{\sigma_X^2 - \sum \sigma_i^2}{\sigma_X^2} \right),$$

$$\frac{\sigma_{T_X}^2}{\sigma_X^2} \geq \frac{N}{N-1} \cdot \left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2} \right).$$

El primer miembro de esta desigualdad es el coeficiente de confiabilidad del test completo mientras que el segundo miembro se conoce usualmente como *"α de Cronbach"*:

$$\alpha = \frac{N}{N-1} \cdot \left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2} \right), \qquad \textbf{(3.14)}$$

$$\rho_{XX'} \geq \alpha .$$

En situaciones prácticas se reemplazan los parámetros por sus estimadores.

La utilidad de esta desigualdad reside en que permite estimar una cota inferior para el coeficiente de confiabilidad de una combinación, conociendo la cantidad de subtests que la componen, la varianza del test completo y las varianzas de cada subtests. Dado cualquier test, se puede considerar que se trata de una combinación en la que los items del test constituyen los *subtests* de la combinación.

Si el valor estimado de α es alto entonces la confiabilidad debe ser alta, pero si es pequeño no podemos saber si realmente la magnitud de la confiabilidad del test es baja dado que se trata de una estimación de una cota inferior para el verdadero coeficiente si los componentes no son "esencialmente τ equivalentes".

Es importante destacar que:

- ✓ Si los componentes son *esencialmente τ equivalentes*, el valor del coeficiente α de Cronbach puede tomarse como el coeficiente de confiabilidad del test completo, puesto que las desigualdades se transforman en igualdades.

- ✓ Si los componentes son *esencialmente τ equivalentes* y además sus varianzas (de los scores observados) son iguales, entonces el valor de α coincide con el obtenido por la fórmula de Spearman-Brown y ambos iguales al coeficiente de confiabilidad, dado que estaríamos en presencia de tests *estrictamente paralelos*.

- ***Caso Particular*: Dos Componentes**

Cuando el test se considera compuesto de sólo dos subtests, la expresión (3.14) se reduce a:

$$\alpha = 2 \cdot \left(1 - \frac{\sum_{i=1}^{2} \sigma_i^2}{\sigma_X^2} \right), \qquad \textbf{(3.15)}$$

Para ilustrar el uso práctico de (3.15) supongamos que un test se compone de dos partes, las que tienen una correlación estimada (confiabilidad) de 0,5. Las varianzas estimadas de cada mitad son 7 y 5 y se sabe que la varianza estimada del score total es 18. Usando la fórmula de Spearman-Brown, la confiabilidad del test completo resulta igual a:

$$r_{XX'} = \frac{2 \cdot r_{YY'}}{1 + r_{YY'}} = \frac{2.0{,}5}{1 + 0{,}5} = 0{,}67$$

Si decidimos estimar el coeficiente α de Cronbach, tenemos que:

$$\alpha = 2 \cdot \left(1 - \frac{7+5}{18} \right) = 0{,}67$$

- ***Coeficiente α de Cronbach para items dicotómicos: Fórmula KR_{20} de Kuder- Richardson***

Si los items del test son dicotómicos, el coeficiente α de Cronbach toma una forma particular, conocida como la Fórmula de Kuder-Richardson (KR_{20}).

La derivación de esta fórmula no es complicada. En el caso de items dicotómicos (sólo pueden tomar dos valores 0 ó 1), la varianza del score de cada item vendrá dada por:

$$\sigma_i^2 = \pi_i \cdot (1 - \pi_i),$$

donde π_i es la proporción de individuos que responden correctamente al item i, cantidad que se conoce como *dificultad del item i.*

Por lo tanto, el coeficiente de confiabilidad asume la forma:

$$\rho_{XX'} \geq \frac{N}{N-1} \cdot \left(1 - \frac{\sum \pi_i \cdot (1 - \pi_i)}{\sigma_X^2} \right).$$

El segundo miembro de esta desigualdad se denota generalmente como KR_{20}:

$$\mathbf{KR}_{20} = \frac{N}{N-1} \cdot \left(1 - \frac{\sum \pi_i \cdot (1 - \pi_i)}{\sigma_X^2} \right). \qquad \textbf{(3.16)}$$

Supongamos que se aplica un test compuesto de cuatro items dicotómicos en un curso de una escuela primaria y que la varianza de los scores observados resultante es $s^2_x = 2$, mientras que las proporciones de alumnos que contestaron correctamente cada item son, respectivamente: 0,2 ; 0,6 ; 0,4 y 0,5.

Entonces la suma de las varianzas estimadas de los cuatro items será:

$$\sum_{i=1}^{5} p_i q_i = 0{,}2 \cdot 0{,}8 + 0{,}6 \cdot 0{,}4 + 0{,}4 \cdot 0{,}6 + 0{,}5 \cdot 0{,}5 = 0{,}89.$$

Podemos ahora utilizar (3.16):

$$r_{XX'} \geq \mathbf{KR}_{20},$$

$$\geq \frac{4}{4-1} \cdot \left(1 - \frac{0{,}89}{2}\right),$$

$$\geq 0{,}74.$$

3.3.2.2. ***Método de Hoyt*** *(1941)*

Este método utiliza el Análisis de la Varianza (ANOVA), considerando al conjunto de individuos y al conjunto de items como fuentes de variación. Es decir se plantea un ANOVA de dos factores con una observación por celda, para estimar los cuadrados medios correspondiente a cada fuente de variación. Luego Hoyt identifica la varianza de los scores de error con la varianza residual y la varianza de los scores observados con la varianza del factor "individuos". Como consecuencia de ello propone utilizar los cuadrados medios de estos efectos como estimadores en la expresión del coeficiente de confiabilidad.

La tabla del ANOVA de dos factores presenta el aspecto siguiente, donde "n" es la cantidad de inidividuos que toman el test y "k" la cantidad de items que contiene.

Fuente de Variación	S.S. (Suma de Cuadrados)	df (Grados de libertad)	MS (Cuad. Medios)
Personas	SS(p)	n-1	MS(p)=SS(p)/(n-1)
Items	SS(i)	k-1	MS(i)=SS(i)/(k-1)
residual	SS(r)	(n-1) (k-1)	MS(r)=SS(r)/(n-1)(k-1)
Total	SST	nk-1	

Tabla 3.2. Tabla de ANOVA en dos direcciones para el Método de Hoyt

Las expresiones para la Suma de Cuadrados de los factores son las siguientes:

$$SS(p) = n_i \sum_p \left(X_p - \overline{X}\right)^2 = n_i \sum_p \overline{X_p}^2 - n_p n_i \overline{X}^2,$$

$$SS(i) = n_p \sum_i \left(X_i - \overline{X}\right)^2 = n_p \sum_i \overline{X_i}^2 - n_p n_i \overline{X}^2,$$

$$SS(pi) = \sum_p \sum_i \left(X_{pi} - \overline{X_p} - \overline{X_i} + \overline{X}\right)^2,$$

$$SS(pi) = \sum \sum X_{pi}^2 - n_i \sum_p \overline{X_p}^2 - n_p \sum_i \overline{X_i}^2 + n_p n_i \overline{X}^2.$$

Finalmente, Hoyt propone que el coeficiente de confiabilidad:

$$\rho_{xx'} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

se estime mediante:

$$\hat{\rho}_{xx'} = 1 - \frac{MS(r)}{MS(p)} \qquad \textbf{(3.18)}$$

Aunque este es un método sencillo para estimar el coeficiente de confiabilidad, sobre todo cuando no se dispone de un software que lo ofrezca, los resultados de un Análisis de Varianza no son suficientes para estimar otras características de los items de un test que serán presentadas en el Capítulo 5.

Para ilustrar el método, supongamos que se disponen de los siguientes scores observados de 10 individuos en un test compuesto por 8 items:

Alumno	Item								Total
	1	2	3	4	5	6	7	8	
1	1	1	1	1	1	1	0	0	6
2	1	1	1	0	0	1	0	0	4
3	1	1	0	1	0	0	0	0	3
4	0	0	0	0	1	0	1	1	3
5	1	1	1	1	1	1	1	1	8
6	1	1	1	1	1	0	0	0	5
7	0	1	1	1	1	0	0	1	5
8	1	1	1	1	1	1	1	0	7
9	1	1	0	0	0	0	0	0	2
10	1	1	1	1	1	0	1	1	7

Tabla 3.3. Puntajes asignados en ocho items a diez alumnos.

Los resultados del ANOVA se presentan en la Tabla 3.4.

Efecto	df	SS	MS
Persona	9	4,500	0,500
Item	7	2,750	0,393
Residual	63	11,500	0,183

Tabla 3.4. ANOVA en dos direcciones para los datos de la Tabla 3.3.

Por lo tanto:

$$r_{xx'} = 1 - \frac{MS(r)}{MS(p)},$$

$$= 1 - \frac{0{,}189}{0{,}500},$$

$$= 0{,}622.$$

3.4. CONFIABILIDAD DE TESTS REFERENCIADOS EN CRITERIOS

En el capítulo 1 se introdujo el concepto de *test referenciado en un criterio*. Para estas pruebas la idea central gira en torno a medir el desempeño de un individuo sin hacer referencias o comparaciones con otros individuos que toman el test. En general se trata de pruebas construidas mediante muestreos a partir de un dominio más amplio de contenidos que se conocen como *tests aleatoriamente paralelos*. Esta expresión intenta dejar de manifiesto que las pruebas se consideran paralelas en contenidos, pero a diferencia de tests estrictamente paralelos, no requieren igualdad de medias, varianzas o correlaciones.

Más aún se expuso que los dos objetivos básicos que se persigue con este tipo de test son, por un lado, la estimación del *score sobre el dominio* (concepto equivalente al de score universal según se verá en Teoría de la Generalizabilidad y que no es otra cosa que el score medio del individuo sobre todos los posibles tests paralelos que conforman el Universo de Generalización) y por otro lado la clasificación de los individuos según el grado de destreza que hayan alcanzado en el constructo que se investiga.

Dado que las medidas de confiabilidad del score sobre el dominio requieren de la Teoría de la Generalizabilidad, su tratamiento será pospuesto hasta el Capítulo 8. Por el momento interesa la descripción de la confiabilidad para clasificaciones construidas en base a la destreza lograda por el individuo.

Para ilustrar la situación supongamos que se construye un test orientado a medir los conocimientos alcanzados por un grupo de 100 alumnos en Cálculo Elemental. Un único punto de corte (*cut score*) separa dos regiones: aquellos estudiantes cuyo score observado se localice por encima del punto de corte serán promocidos a un curso de Cálculo Superior mientras que los demás deberán revisar el material el próximo semestre. Es fácil advertir que se trata de un test referenciado en un criterio, donde el dominio de contenidos está

constituido por todos los temas del Cálculo Elemental, de los cuales se extraerá, aleatoriamente, una muestra de contenidos que integrarán el test. Debe observarse que las decisiones que se tomen en relación a un individuo no se verán afectadas en absoluto por el desempeño de los restantes miembros del grupo.

En esta circunstancia, la precisión de las estimaciones que se obtengan para los scores sobre el dominio no son tan importantes como la *consistencia de las decisiones* que se hagan en base a tales estimaciones.

3.4.1. ESTIMACION DE LA CONSISTENCIA DE LAS DECISIONES

Con la expresión *consistencia de las decisiones* se alude al grado con el cual se toman idénticas decisiones a partir de dos conjuntos diferentes de mediciones.

Continuando con el ejemplo anterior consideremos la siguiente tabla que muestra las decisiones que se tomarían en base a dos formas paralelas:

		Decisiones basadas en Test 1		
		No Promovido (0)	Promovido (1)	Total
Decisiones basadas en Test 2	No Promovido (0)	45	10	55
	Promovido (1)	25	20	45
	Total	70	30	100

Tabla 3.5. Clasificaciones de 100 alumnos en dos categorías (promovido y no promovido) en base a dos test (Test 1 y Test 2)

A continuación se presentan un par de medidas de esta consistencia de las decisiones.

3.4.1.1. ***Probabilidad estimada de una clasificación consistente***

Si asumimos las frecuencias observadas como estimaciones de las probabilidades de clasificación en cada celda, la probabilidad estimada de una decisión consistente es la suma de la probabilidad estimada de clasificación consistente como *promovido* más de la de la categoría *no promovido*:

$$\hat{p} = \hat{p}_{00} + \hat{p}_{11}, \tag{3.19}$$

$$\hat{p} = \frac{45}{100} + \frac{20}{100},$$

$$\hat{p} = 0{,}65.$$

Este resultado puede interpretarse diciendo que el 65% de los alumnos fueron clasificados consistentemente. Es claro que la probabilidad estimada de una clasificación inconsistente es 1-0,65 = 0,35.

3.4.1.2. ***La medida k (kappa) de Cohen***

Swaminathan *et al* (1974) sugirieron el uso de la Kappa de Cohen como medida de consistencia de las decisiones:

$$\kappa = \frac{p - p_0}{1 - p_0} \tag{3.20}$$

donde p_0 es la probabilidad de tomar decisiones *estadísticamente independientes* y se estima como:

$$\hat{p}_0 = \hat{p}_{1.} \cdot \hat{p}_{.1} + \hat{p}_{0.} \cdot \hat{p}_{.0}, \tag{3.21}$$

es decir la suma de los productos de las probabilidades estimadas marginales de las celdas correspondientes a clasificaciones consistentes.

El denominador de la expresión de la kappa de Cohen es el máximo incremento posible en la consistencia de las decisiones sobre la situación de independencia (denominada *consistencia casual*), mientras que el numerador de dicho cociente representa el aumento real en la consistencia respecto del caso de independencia.

De esta forma la kappa de Cohen puede interpretarse como el incremento en la consistencia de las decisiones que logran las formas paralelas por encima de la consistencia casual.

Para el ejemplo anterior:

$$\hat{p}_0 = 0{,}45 \cdot 0{,}30 + 0{,}55 \cdot 0{,}70,$$

$$\hat{p}_0 = 0{,}52.$$

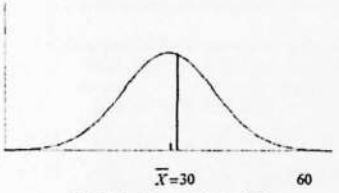
La kappa de Cohen estimada para el ejemplo es:

$$\hat{\kappa} = \frac{0{,}65 - 0{,}52}{1 - 0{,}52},$$

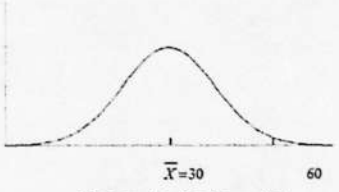
$$= 0{,}27.$$

En este caso concluimos que los tests paralelos proveen un incremento del 27% sobre el posible aumento total de la consistencia sobre la hipótesis de independencia.

Uno de los factores que afecta seriamente la consistencia de las decisiones es la localización del punto de corte en relación al desempeño medio del grupo. Podemos preguntarnos en que situación se lograría un mayor porcentaje de casos consistentemente clasificados, cuando el punto de corte es próximo al score medio del grupo o cuando está alejado de este? La situación se ilustra en los gráficos siguientes:



Score de corte = 31



Score de Corte = 56

Es fácil advertir que cuando el punto de corte se localiza en las colas de la distribución, habrá una frecuencia menor de individuos próximos a dicho punto que podrían *saltar* el score de corte en la segunda prueba. Si por otra parte el punto de corte es próximo a la media de la distribución habrá una mayor cantidad de personas que podrían superar ese punto en la segunda instancia, lo que redundaría en una menor consistencia en las decisiones.

3.5. FACTORES QUE AFECTAN EL COEFICIENTE DE CONFIABILIDAD

Entre los factores más relevantes que tienen una gran incidencia en el Coeficiente de Confiabilidad se destacan cuatro:

- ✓ *La homogeneidad del grupo de individuos*
- ✓ *Los límites de tiempo en algunos tests*
- ✓ *La longitud de la prueba*
- ✓ *La dificultad de los items*

Comentaremos brevemente cada uno de ellos.

• *Homogeneidad del grupo de individuos*

Es razonable imaginar que la homogeneidad del grupo de individuos tendrá influencia sobre la variabilidad tanto de los scores verdaderos como de los errores.

Imaginemos que se dispone de dos grupos, en el primero de los cuales la varianza de los verdaderos scores es inferior que la del otro. Si se conjetura además que la varianza de los errores en dos grupos fuese aproximadamente igual, entonces el mismo test tendrá una confiabilidad más elevada para el segundo grupo. La lógica de este razonamiento reside en que en el segundo caso la varianza de los verdaderos scores captan una mayor proporción de la varianza de las observaciones. Un ejemplo numérico ilustra este punto.

	Grupo 1	Grupo 2
Varianza de los Scores Verdaderos	10	30
Varianza de los Scores de Error	5	5
Varianza de los Scores Observados	15	35
Coeficiente de Confiabilidad	**0,67**	**0,86**

Tabla 3.5. Coeficiente de Confiabilidad para dos grupos con igual varianza de error y diferentes varianzas de los verdaderos scores.

Como consecuencia es razonable pensar que la confiabilidad de una prueba es una propiedad del test *en relación a un determinado grupo de individuos*. Esto motiva un cuidado especial al momento de aplicar tests estándares que han sido diseñados y analizados con una particular población de individuos.

- *Límites de tiempo*

En oportunidades, el tiempo disponible para tomar un test está acotado y los individuos logran responder diferentes porcentajes del total de la prueba. Estas diferencias en la cantidad de respuestas inciden en la varianza de los scores verdaderos incrementándola al agregar una variabilidad extra.

El efecto de este incremento en la varianza de los verdaderos scores se traducirá en un aumento artificial del coeficiente de confiabilidad de la prueba.

Esta circunstancia debería evitarse (asegurando una cantidad de tiempo suficiente para completar el test) siempre que la proporción de trabajo no sea de interés.

Tests de Velocidad

En otros casos, interesa aplicar un *test de velocidad* que es sencillamente un test compuesto por items que pueden ser respondidos correctamente por todos los individuos si se diera suficiente tiempo. Pero el test se administra con una cantidad de tiempo reducida para determinar cuán rápidamente pueden trabajar los individuos.

Los métodos de *consistencia interna* descriptos no son adecuados para estimar el coeficiente de confiabilidad en tests de velocidad. Veamos dos situaciones en las que aparece un efecto distorsivo en la confiabilidad:

- Supongamos que se decide dividir el test en dos mitades asignando a una de ellas los items pares y a la otra los impares. Dada la sencillez de los items muy probablemente el individuo haya podido contestar correctamente todos los items que respondió, por lo que los items pares e impares estarán casi perfectamente correlacionados lo que significa que la confiabilidad estimada se situará cercana a uno.

- Supongamos el caso en que las mitades del test se construyan separando los primeros de los últimos. Es muy probable que los primeros items presenten altos scores mientras los últimos items tengan puntuaciones bajas. En tal circunstancia, la correlación entre ambas mitades será casi nula y en consecuencia también lo será la confiabilidad.

En consecuencia, cuando se trata de test de velocidad, se sugiere aplicar métodos de test-retest o formas equivalentes para medir la confiabilidad de estas pruebas.

- *Longitud de la prueba*

Pruebas más extensas serán más confiables, en virtud que al adicionar items al test, el número de términos de covarianza que se adicionan (en el cálculo de la varianza total) se incrementa más rápidamente que el número de términos de varianzas de items. Por ejemplo si a un test que inicialmente consta de cinco items (y por lo tanto el cálculo de la varianza total implica cinco términos de varianzas y n(n-1) = 5 . 4 = 20 términos de covarianzas, de los cuales la mitad serán distintos numéricamente por una cuestión de simetría) se adicionan otros cinco items, entonces la combinación de 10 items implicará que en el cálculo de la varianza total se computen 10 términos de varianzas (una por cada item de la prueba) y 10*9=90 términos de covarianza. Esto produce que la varianza del score verdadero de la prueba crezca más rápidamente que la varianza del error.

- *Dificultad de los items*

Un nivel de dificultad similar en los items de una prueba tiene el efecto de incrementar la confiabilidad del test en contraste con otro cuyos items presenten mayor variabilidad en este aspecto.

Es importante advertir que cuando un test es demasiado *fácil* (o *difícil*) para un grupo de individuos y en consecuencia los scores observados son mayoritariamente altos (o bajos), situación que se conoce como *efecto techo* (o *efecto piso*), la restricción en el rango de los scores que se observen hará que la varianza de los scores verdaderos se vea afectada (reducida) y con ello la confiabilidad de la prueba.

Desde otro punto de vista, debemos tener presente que cuanto más altas sean las covarianzas entre pares de items (y por lo tanto las correlaciones entre ellos sean más elevadas) mayor será el coeficiente de confiabilidad de la prueba. Esto tiene lugar cuando una misma persona que contesta correctamente el item i también responde acertadamente al item j y aquella que responde incorrectamente al item i también lo hace con el item j, es decir cuando los items tengan iguales dificultades.

No solo importa que los items presenten dificultades similares, sino que también aporta el hecho que se trate de niveles medios de dificultad. Esto puede apreciarse teniendo en cuenta que la varianza de un item viene dada (en el caso dicotómico) por el producto π_i .$(1-\pi_i)$. Dado que se trata de números comprendidos entre 0 y 1 (por ser proporciones), sus valores más altos se logran en las cercanías de 0,5.

3.6. EL ERROR ESTANDAR DE MEDICION

3.6.1. DEFINICION

Todos los métodos descriptos hasta este momento se dirigían a estimar el coeficiente de confiabilidad a través de un estudio de la consistencia del conjunto de scores observados en una prueba. En este punto nos preguntamos por la *consistencia de los scores individuales*. El concepto de *Error Estándar de Medición* está vinculado a la precisión en la medida del desempeño individual de una persona que toma la prueba.

En la conclusión 13 del capítulo anterior se mostró que el error estándar de medición viene dado por:

$$\sigma_E = \sigma_X \cdot \sqrt{1 - \rho_{XX'}}$$

Para un dado individuo, la distribución de los scores observados, según los supuestos de la Teoría Clásica de Score Verdadero, está centrada en T (el verdadero score), y presenta un error estándar igual a σ_E (igual al error estándar de medición).

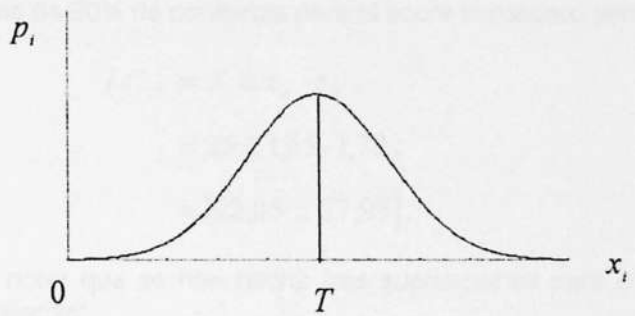


Figura nº 3.5: Distribución de probabilidad de los scores observados en ensayos repetidos de un mismo individuo

Como se ve sería necesario disponer de un gran número de pruebas sobre un mismo individuo para computar su error estándar de medición. Esto no es precisamente lo que ocurre en situaciones prácticas. Sin embargo si *asumimos que todos los alumnos tendrán un error estándar similar*, podemos utilizar la expresión:

$$s_E = s_X \cdot \sqrt{1 - r_{XX'}} \qquad \textbf{(3.22)}$$

para estimar el error estándar de estimación.

3.6.2. INTERVALOS DE CONFIANZA PARA EL SCORE VERDADERO

Suponiendo que la distribución de las observaciones es aproximadamente normal, es posible construir intervalos de confianza para estimar el verdadero score de un individuo:

$$P(X - z_c \cdot s_E < T < X + z_c \cdot s_E) = 1 - \alpha \qquad \textbf{(3.23)}$$

donde z_c es el coeficiente de confianza de la distribución normal para un determinado nivel de confianza (1-α). Vemos que cuanto mayor es el error estándar de medición, mayor será la longitud del intervalo de confianza y por lo tanto menor será la precisión de la estimación. Esto ciertamente ocurrirá cuando la confiabilidad sea cada vez más baja.

Supongamos, como ejemplo, que un individuo presenta un score observado de 25. Además se sabe que $s_X = 4$ y $r_{XX'} = 0{,}8$. Entonces el error estándar de estimación vendría dado por:

$$s_E = 4 \cdot \sqrt{1 - 0{,}8} = 1{,}79.$$

Un intervalo de 90% de confianza para el score verdadero será:

$$\begin{aligned} I.C._T &= X \pm z_c \cdot s_E \,, \\ &= 25 \pm 1{,}65 \cdot 1{,}79 \,, \\ &= [22{,}05 \;;\; 27{,}95]. \end{aligned}$$

Debemos notar que se han hecho tres suposiciones para construir este intervalo de confianza:

- ✓ Valen los supuestos de la Teoría Clásica del Score Verdadero.
- ✓ La distribución de los scores observados es aproximadamente normal.
- ✓ Hay *homocedasticidad*, es decir todos los alumnos tienen errores estándares de medición iguales.

En la medida que estas hipótesis no sean sustentables en una situación práctica, los intervalos de confianza que se construyan pueden conducir a serios equívocos. Si los supuestos de normalidad y homocedasticidad no pueden sostenerse, es posible trabajar con otros modelos como el Modelo Binomial que integra la Teoría Fuerte de los Verdaderos Scores y que no se incluye en este trabajo.

CAPITULO 4

Análisis de Validez

Un test es *válido* si mide lo que se propone medir. Un test de logros en Estadística es válido si discrimina entre estudiantes que han alcanzado diferentes niveles de aprendizaje; un test de personalidad es válido si logra clasificar correctamente a individuos según sus características personales, etc.

Para Cronbach, la validación de un test es un proceso en el cual el investigador recolecta evidencia que sustenta los tipos de inferencias que se hacen a partir de los scores de la prueba.

Un *Análisis de Validez* puede referirse a uno o más de los siguientes aspectos:

- ✓ *Validez de Contenido*: si las inferencias que se desean construir se refieren a un dominio de items más amplio que aquellos incluidos en el test.

- ✓ *Validez referenciada en un criterio*: si las inferencias que se construyan se refieren al desempeño de los individuos en alguna variable que represente algún comportamiento de importancia.

- ✓ *Validez de un constructo*: cuando las inferencias que se elaboren se relacionen con desempeños propios de algún constructo psicológico.

Debería ser claro que se trata de diferentes clases de análisis de validez adecuados a distintos tipos de inferencias y que por lo tanto son *no intercambiables*.

En ocasiones se requiere más un tipo de análisis de la validez de la prueba como condición necesaria para su uso.

Este capítulo se dedica completamente al estudio de la validez de un test, estableciendo en cada caso las instancias que deben contemplarse, los problemas que usualmente se presentan y las técnicas apropiadas. Se presenta con especial detalle dos herramientas del Análisis Estadístico Multivariado: el Análisis Discriminante y el Análisis Factorial. Finalmente se hace una breve referencia a la corrección por atenuación para verdaderos scores y al efecto que una selección de individuos puede tener en el coeficiente de validez.

4.1. VALIDEZ DE CONTENIDO

Con un análisis de este tipo se trata de establecer si los items incluidos en la prueba son una representación adecuada del dominio de contenidos del que han sido extraídos.

4.1.1. ETAPAS DE UN ANALISIS PARA ESTABLECER LA VALIDEZ DE CONTENIDO

Conducir un estudio tendiente a determinar la validez de contenido de una prueba, por lo general implica las siguientes instancias:

- Definición del *Dominio de Contenidos* que se estudia.
- Selección de un panel de expertos en ese dominio de contenidos.
- Establecer un proceso sistemático que vincule items con el dominio de contenidos.
- Recolectar y analizar datos generados por el proceso anterior.

4.1.2. PROBLEMAS ASOCIADOS A ANALISIS DE LA VALIDEZ DE CONTENIDO

Dado que un estudio de la validez de contenido se basa en juicios subjetivos está más sujeta a errores que otros tipos de análisis. Por lo general este análisis constituye un primer paso, pero aunque un test exhiba validez de contenido, ésta no es condición suficiente para su aplicación.

4.2. VALIDEZ REFERENCIADA EN UN CRITERIO

En ocasiones resulta de interés utilizar los scores de un test para obtener inferencias relativas al desempeño de los individuos en cuestiones que no son medidas directamente por la prueba (criterio). Podría ser importante pronosticar adecuadamente el desempeño laboral a partir de los resultados alcanzados por los candidatos a un empleo o predecir el rendimiento académico en base a los scores de un examen de admisión en una carrera universitaria.

Pero en cualquier caso, antes de proceder a utilizar los puntajes de una prueba con tales finalidades, es absolutamente necesario reunir evidencia suficiente sobre la existencia de una relación entre los scores del test (predictor) y la variable que representa el comportamiento que se predice (criterio) mediante un análisis de validez referenciada en ese criterio.

4.2.1. ETAPAS DE UN ANALISIS PARA ESTABLECER LA VALIDEZ REFERENCIADA EN UN CRITERIO

Los pasos que se siguen en un estudio de esta naturaleza son:

- Identificar un criterio (variable) adecuado que se relaciona con el test.
- Seleccionar una muestra de individuos representativa de la población sobre la cual se pretende inferir.
- De cada individuo registrar el score del test y el correspondiente al criterio.
- Establecer la dirección e intensidad de la relación entre estos puntajes.

4.2.2. LA VALIDEZ PREDICTIVA Y LA VALIDEZ CONCURRENTE

Algunos autores distinguen dos clases de estudios de validación con referencia a un criterio:

✓ *La Validez Predictiva* : Cuando los scores del test se usan para predecir el comportamiento de los individuos *en el futuro*, es decir que las mediciones del criterio tienen lugar en algún instante del tiempo posterior al test. Por ejemplo, para otorgar validez predictiva a los scores de la prueba de admisión a un programa, se debería administrar el test, admitir a todos los candidatos y registrar los puntajes promedios que cada individuo logra en el examen final, al término del programa. Luego, debería establecerse la magnitud y dirección de la correlación entre estos puntajes y si esta es

adecuada, recién entonces se estaría en condiciones de justificar el uso de los puntajes del test en la toma de decisiones de selección de candidatos, en admisiones subsiguientes.

✓ *La Validez Concurrente*: El registro de los scores del test y los del criterio asociado se hace simultáneamente. Este análisis resulta adecuado cuando los scores del test se utilizarán para estimar un criterio (más que para predecir sus valores en el futuro). Supongamos, por ejemplo, que se dispone de un largo y costoso test de inteligencia. Una forma de ahorrar recursos podría consistir en administrar una prueba relativamente breve y seleccionar aquellos individuos que hayan mostrado un resultado particularmente elevado para participar del test más extenso. Para ello resultaría necesario demostrar la validez concurrente entre los scores de ambos tests. El coeficiente de validez que se obtiene en este caso tiende a subestimar la verdadera correlación entre las variables (scores del test y del criterio) debido a la fuerte restricción del rango de puntajes que puede tener lugar si la selección de individuos elimina una proporción importante de candidatos.

4.2.3. PROBLEMAS ASOCIADOS A UN ANALISIS DE LA VALIDEZ REFERENCIADA EN UN CRITERIO.

Entre los problemas que surgen en un intento por establecer la validez relacionada a un criterio, deben mencionarse al menos:

✓ *La determinación de un criterio adecuado*: en algunas situaciones puede ocurrir que el criterio adecuado para establecer la validez del test no resulte sencillo en cuanto a su definición operacional a través de variables observables, como la *efectividad del desempeño* o la *independencia de juicios de valor.*

✓ *Tamaño de la Muestra*: si la muestra con la que se intenta establecer la correlación entre las variables de interés es pequeña, puede ocurrir que los errores de muestreo sean importantes y se reduzca el poder de las pruebas estadísticas que se lleven a cabo.

✓ *Restricción del rango:* Como se mencionó anteriormente, una reducción en el rango de valores de los scores del test y/o del criterio puede conducir a una subestimación del coeficiente de validez observado. Además se producirse en situaciones de selección de individuos, puede también tener lugar cuando se produce un *efecto techo* (el nivel de dificultad del test es bajo y la gran mayoría de los individuos alcanzan puntajes altos) o un *efecto piso* (el nivel de dificultad del test es muy alto y la mayoría de las personas que lo toman obtienen scores bajos).

✓ *Confiabilidad del Predictor y del Criterio:* Los errores de medición tanto del predictor (test) como del criterio pueden afectar seriamente el valor de la correlación que exista entre ellos. Una relación importante, que luego se probará establece que:

$$\rho_{xy} \leq \sqrt{\rho_{xx'}} \cdot \sqrt{\rho_{yy'}} \qquad (4.1)$$

De esta forma, conociendo los valores de los coeficientes de confiabilidad para el predictor y el criterio es posible establecer una cota máxima para el coeficiente de correlación. Cuanto más bajas sean estas confiabilidades, más lo será la correlación entre ambas cantidades.

4.2.4. COEFICIENTES DE VALIDEZ A PARTIR DE UN UNICO PREDICTOR

Los estadísticos que se utilizan para medir la fuerza de la asociación entre el predictor y el criterio se conoce como *Coeficiente de Validez.*

Según el nivel de medición de la variable que representa el criterio asociado al test se selecciona una medida de correlación adecuada:

✓ Si el criterio se representa por una variable continua: Por lo general se calcula el Coeficiente de Correlación Lineal de Pearson.

✓ Si el criterio se representa por una variable categórica: Se puede llevar a cabo una prueba estadística para determinar si existen diferencias significativas entre los scores medios de cada categoría como el *test t* para dos grupos o ANOVA para más de dos grupos. Otros estadísticos adecuados se presentarán en el capítulo siguiente.

✓ Si el predictor y el criterio se representan ambos por variables categóricas: es posible utilizar alguna medida de asociación adecuada a los niveles de medición como por ejemplo el coeficiente Φ si ambas son dicotómicas y nominales, o τ_b de Kendall si ambas son ordinales, etc.

4.2.5. EL COEFICIENTE DE DETERMINACION

Generalmente suele reportarse, además del coeficiente de validez, el cuadrado de su valor, conocido como el *Coeficiente de Determinación* que al multiplicarse por 100 se interpreta como el porcentaje de la varianza de los scores del criterio que puede explicarse a partir de la variabilidad de los scores del predictor (test).

Por ejemplo, un coeficiente de determinación de 0,75 entre los scores del rendimiento académico durante el primer año de un grupo de ingresantes universitarios y los puntajes del examen de ingreso, puede interpretarse diciendo que el 75% de la varianza de la variable que mide el rendimiento académico durante el primer año puede justificarse por la varianza que presentan los scores del examen de admisión.

4.2.6. ESTIMACION DEL CRITERIO A PARTIR DE UN UNICO PREDICTOR

Este apartado se dedica a mostrar como se puede utilizar los scores de un test (predictor) en la estimación o predicción de un criterio asociado.

Luego de seleccionar una muestra aleatoria de individuos que representan la población sobre la cual interesa construir las inferencias, se obtienen los scores del test y del criterio para cada persona. A partir de estos datos se construye la ecuación de regresión lineal simple (que es el mejor predictor lineal por mínimos cuadrados) que puede utilizarse en individuos para los cuales se conoce su score del test pero no su score sobre el criterio:

$$Y_i' = r_{XY} \cdot \left(\frac{s_Y}{s_X}\right) \cdot \left(X_i - \overline{X}\right) + \overline{Y}. \qquad \textbf{(4.2)}$$

Esta ecuación construida con la muestra seleccionada para el análisis de validez estima los valores del criterio Y_i a partir de los estadísticos r_{XY}, s_X, s_Y y las medias muestrales de X e Y.

Si se sostienen tres supuestos es posible derivar intervalos de confianza para el verdadero valor (desconocido) de Y_i:

- Existe una relación lineal entre X e Y.
- Hay *homocedasticidad*, es decir $s_{Y.X}$ es el mismo para cualquier valor de X.
- La distribución condicional de Y dado X es normal

El intervalo de confianza tendrá la forma

$$I.C._{Y_i} = Y_i' \pm z_c \cdot s_{Y.X}, \qquad \textbf{(4.3)}$$

donde z_c es el coeficiente de confianza que surge de la distribución normal estándar y $s_{Y.X}$ es el *error estándar de estimación muestral*:

$$s_{Y.X} = s_Y \cdot \sqrt{1 - r_{XY}^2} \qquad (4.4)$$

Como ejemplo de aplicación de estas expresiones, supongamos que X representa el score en un examen de ingreso, Y el rendimiento medio alcanzado durante el primer año de estudio y que $r_{XY} = 0,7$; $s_X = 20$; $s_Y = 15$ y las medias muestrales de X e Y iguales a 65 y 50 respectivamente. Entonces para la i-ésima persona cuya nota de ingreso es 80 puntos, el intervalo de confianza será:

$$\begin{aligned} Y_i' &= 0,7 \cdot \left(\frac{15}{20}\right) \cdot (80 - 65) + 50, \\ &= 57,875, \\ s_{Y.X} &= 15 \cdot \sqrt{1 - 0,7^2} = 10,712, \\ I.C._{Y_i} &= 57,875 \pm 1,96 \cdot 10,712, \\ &= [36,879 \ ; \ 78,870] \end{aligned}$$

Debería observarse que la falta de confiabilidad en las mediciones del predictor y/o del criterio implicaría una atenuación del valor del coeficiente de validez r_{XY} lo que a su vez se traduciría en un intervalo de confianza más ancho, esto es una estimación menos precisa.

4.2.7. ESTIMACION DEL CRITERIO A PARTIR DE MULTIPLES PREDICTORES

En muchas situaciones en las que se requiere la estimación del desempeño individual de una persona en un criterio, se dispone de dos o más variables predictoras, como por ejemplo podría registrarse no sólo el score en el test de admisión a una carrera universitaria sino también la nota promedio que el alumno alcanzó durante sus estudios secundarios, como predictores del rendimiento que podría lograr en el primer año de sus estudios en la Universidad.

En este apartado consideraremos en primer lugar el caso de tener múltiples predictores para estimar un criterio que se asume continuo. Para tal circunstancia describiremos dos técnicas alternativas: El coeficiente de correlación parcial y El Análisis de Regresión Múltiple. Luego se presenta el caso en que la variable que representa el criterio es categórica y entonces haremos uso del Análisis Discriminante.

4.2.7.1. ***El coeficiente de correlación parcial***

En ocasiones el interés está centrado en establecer la intensidad y dirección de la relación entre el criterio de interés y alguno de los múltiples predictores de los que se dispone, una vez que se ha descontado el efecto de estos últimos. Una medida adecuada de esa relación la ofrece el *Coeficiente de Correlación Parcial.*

Alternativamente, podemos interpretar dicho coeficiente como una medida de la relación que vincula el criterio con uno de los predictores para un subconjunto de individuos que son homogéneos con respecto a los restantes predictores. Veamos el caso sencillo de dos predictores, aunque los resultados admiten una generalización a más variables.

Para continuar con el ejemplo donde los predictores eran la nota promedio en la escolaridad media (X_1) y los scores en el examen de admisión (X_2), para el criterio definido como el rendimiento que alcanza el individuo durante el primer año de estudios superiores (Y), podemos interesarnos en la medida en que se interrelacionan los scores del test de ingreso con el criterio que se analiza, para subgrupos de estudiantes que comparten la misma nota promedio del nivel secundario. Una medida de esta correlación la dará:

$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1} \cdot r_{X_1X_2}}{\sqrt{1 - r_{YX_1}^2} \cdot \sqrt{1 - r_{X_1X_2}^2}} \cdot \qquad \textbf{(4.5)}$$

Esta expresión muestra que la correlación que buscamos no sólo depende de las variables involucradas sino también de la correlación entre el criterio y la variable controlada y de la correlación entre los predictores. En la medida que crezca la correlación entre los predictores, decrecerá la correlación parcial entre el criterio y uno de ellos, como se muestra a través de un ejemplo numérico sencillo en la tabla siguiente:

$r_{x_1x_2}$	r_{yx_2}	r_{yx_1}	$r_{yx_2.x_1}$
0,4	0,35	0,38	0,212
0,8	0,35	0,38	0,083

Tabla 4.1. Ejemplo numérico de correlación parcial.

Un valor positivo para el coeficiente de correlación parcial debe interpretarse en el sentido que, cuando se descuenta el efecto de la variable controlada, un aumento en el criterio está asociado a un aumento en el predictor que se considera.

Ya hemos indicado que los errores de medición en el criterio y/o el predictor disminuyen el valor absoluto de las correlaciones simples (de orden 0), según se desprende de la ecuación (4.1). Sin embargo esta afirmación no es válida cuando se consideran las correlaciones parciales, puesto que es posible que ésta sea mayor que la correpondiente correlación de orden 0 entre tales variables si la asociación entre los predictores es negativa. Una expresión para corregir por confiabilidad el coeficiente de correlación parcial es la siguiente:

$$r_{yx_2.x_1} = \frac{r_{yx_2} \cdot r_{11} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{r_{11} \cdot r_{yy'} - r_{yx_1}^2} \cdot \sqrt{r_{11} \cdot r_{22} - r_{x_1x_2}^2}}. \qquad \textbf{(4.6)}$$

donde r_{11} , r_{22} y $r_{YY'}$ son los coeficientes de confiabilidad de X_1 , X_2 e Y.

4.2.7.2. ***El Análisis de Regresión Múltiple***

- ***Los coeficientes de regresión y el término independiente***

Cuando se trata de predecir el desempeño en un criterio en base a los valores registrados en múltiples predictores, resulta necesario construir una ecuación que generalice el caso simple dado por:

$$Y_i' = b_{Y.X} \cdot X_i + c$$

que no es otra cosa que la ecuación (4.2) reescribiendo:

$$b_{Y.X} = r_{XY} \cdot \left(\frac{s_Y}{s_X} \right),$$
$$c = \overline{Y} - b_{Y.X} \cdot \overline{X}. \qquad (4.7)$$

En lo que sigue se mostrará el caso de dos predictores, por sencillez, aunque debe tenerse presente en todo momento que las expresiones pueden generalizarse a más predictores.

Para dos variables explicativas, la ecuación de regresión lineal es:

$$Y_i' = b_{YX_1.X_2} \cdot X_1 + b_{YX_2.X_1} \cdot X_2 + c. \qquad (4.8)$$

En la ecuación (4.8), las cantidades que multiplican a cada predictor se conocen como *coeficientes de regresión* y deben interpretarse como el efecto en el criterio provocado por un cambio unitario en ese predictor cuando la otra variable independiente se mantiene constante.

Estos coeficientes de regresión y el intersecto, se obtienen con el criterio de *mínimos cuadrados*, es decir se eligen de tal forma que se minimice la suma de cuadrados de los desvíos entre los valores observados y predichos por la ecuación para el criterio:

$$b_{YX_1.X_2} = \frac{s_Y \cdot \left(r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2} \right)}{s_{X_1} \left(1 - r_{X_1X_2}^2 \right)},$$
$$b_{YX_2.X_1} = \frac{s_Y \cdot \left(r_{YX_2} - r_{YX_1} \cdot r_{X_1X_2} \right)}{s_{X_2} \left(1 - r_{X_1X_2}^2 \right)}, \qquad (4.9)$$
$$c = \overline{Y} - b_{YX_1.X_2} \cdot \overline{X_1} - b_{YX_2.X_1} \cdot \overline{X_2}.$$

Para dos predictores, las ecuaciones análogas a la ecuación (4.7) para calcular los coeficientes de regresión involucran las correlaciones parciales:

$$b_{YX_2.X_1} = r_{YX_2.X_1} \cdot \left(\frac{s_{Y.X_1}}{s_{X_1.X_2}} \right),$$

$$b_{YX_1.X_2} = r_{YX_1.X_2} \cdot \left(\frac{s_{Y.X_2}}{s_{X_1.X_2}} \right). \qquad (4.10)$$

Se presentan ahora dos medidas que pueden utilizarse para evaluar la precisión de las predicciones que se hagan a partir del modelo presentado.

- ***El cuadrado del coeficiente de correlación de validación cruzada***

No debe perderse de vista que la ecuación de regresión muestral dada por la ecuación (4.8) es una estimación de la ecuación de regresión poblacional:

$$Y_i'' = B_{YX_1.X_2} \cdot X_1 + B_{YX_2.X_1} \cdot X_2 + C. \qquad (4.11)$$

la que se obtendría si sobre todos y cada uno de los individuos de la población en estudio se registraran los valores del criterio y de ambos predictores.

Sobre cada individuo, debe distinguirse claramente entre las siguientes cantidades:

- El verdadero score del criterio, denotado por Y.
- El score del criterio predicho por la ecuación de regresión muestral, dado por la ecuación (4.8) y denotado por Y'.
- El score del criterio predicho por la ecuación de regresión poblacional, dado por la ecuación (4.11), representado por Y''.

Ahora supongamos que pudiéramos aplicar la ecuación (4.8) de regresión muestral sobre todos y cada uno de los individuos de la población, calculando el valor de Y' para cada uno de ellos.

El cuadrado del coeficiente de correlación entre los valores de Y (verdaderos scores del criterio) y los de Y' (predichos por la ecuación de regresión muestral) se conoce como *El cuadrado del coeficiente de correlación de validación cruzada*.

Esta cantidad es una medida de la precisión con que se predicen los verdaderos scores a partir de la ecuación de regresión muestral y se denota como ρ^2_{cv}. La expresión *validación cruzada* hace referencia al hecho que la

precisión de las predicciones se evalúan a partir de las puntuaciones de individuos que no integran la muestra seleccionada.

Llamamos ρ , el coeficiente de correlación múltiple, a la correlación entre los scores del criterio *Y* y los scores predichos por la ecuación de regresión poblacional *Y´´*. En otras palabras, ρ representa una medida de la magnitud y dirección de la asociación entre la variable dependiente y el conjunto de predictores.

Denotemos por $\rho^2_{YY'}$ al cuadrado del coeficiente de correlación entre los valores de *Y´* (scores de criterio predichos por la ecuación de regresión muestral) y los de *Y´´* (scores de criterio predichos por la ecuación de regresión poblacional). Es fácil ver que se trata de una medida de la similitud entre ambas ecuaciones.

Rozeboom (1981) demostró que el cuadrado del coeficiente de correlación de validación cruzada se puede expresar como el producto del coeficiente de correlación múltiple y el coeficiente de correlación de los scores predichos por las ecuaciones de regresión muestral y poblacional:

$$\rho^2_{cv} = \rho^2 \cdot \rho^2_{YY'} \qquad \textbf{(4.12)}$$

El cuadrado de la correlación entre los scores de ambas ecuaciones de regresión, $\rho^2_{YY'}$ depende del cociente entre N, el tamaño muestral y el número de predictores.

Por lo tanto, deducimos que el cuadrado del coeficiente de correlación de validación cruzada, como medida de la precisión de las predicciones de los scores del criterio en base a la ecuación de regresión muestral, puede ser pequeño debido a una relación débil entre el criterio y el conjunto de variables explicativas o bien por un tamaño muestral insuficiente para estimar la ecuación de predicción.

Estimación de ρ

El coeficiente de correlación lineal múltiple, en el caso de dos variables predictoras, se estima mediante la conocida relación:

$$R^2 = \frac{s_{X_1} \cdot b_{YX_1 \cdot X_2} \cdot r_{YX_1} + s_{X_2} \cdot b_{YX_2 \cdot X_1} \cdot r_{YX_2}}{s_Y}. \qquad \textbf{(4.13)}$$

Estimación de ρ^2_{cv}

Browne (1975) propuso un procedimiento para estimar ρ^2_{cv} como:

$$R^2_{cv} = \frac{(N-k-3)\cdot R^4_c + R^2_c}{(N-2k-2)\cdot R^2_c + k}, \quad (4.14)$$

donde k es el número de regresores y R^2_c se define de la siguiente forma:

$$R^2_c = R^2 - \frac{k\cdot(1-R^2)}{N-k-1}, \quad (4.15)$$

y comúnmente se conoce como R^2 *ajustado*, mientras que R^4_c es:

$$R^4_c = (R^2_c)^2 - \frac{2k\cdot(1-R^2_c)^2}{(N-1)\cdot(N-k+1)}, \quad (4.16)$$

Nota: dado que tanto R^2_c como R^4_c pueden ser negativos, se sugiere que en tales casos, se asuma que su valor es cero.

Un método alternativo para estimar el valor de ρ^2_{cv} es dividir la muestra en dos partes: utilizar una parte para estimar la ecuación de predicción y luego aplicar esta ecuación en la otra parte de la muestra para obtener las predicciones del criterio Y' para finalmene calcular el coeficiente de correlación entre estos últimos valores y los correspondientes valores observados Y del criterio.

- ***El error estándar de estimación de Y'***

El *error estándar de estimación* para múltiples regresores se calcula a través de la relación conocida:

$$s_{Y.X} = \sqrt{\frac{N-1}{N-k}\cdot s^2_Y\cdot(1-R^2)}. \quad (4.17)$$

Como se sabe, el error estándar de estimación surge de considerar los desvíos entre los valores observados del criterio Y y los predichos por la línea de regresión construida a partir de la muestra seleccionada Y'.

El error estándar de estimación de Y´ puede usarse para construir intervalos de confianza para el verdadero valor del criterio, de la forma:

$$I.C._{Y_i} = Y_i' \pm t_c \cdot s_{YX}, \qquad \textbf{(4.20)}$$

Finalmente, digamos que la selección del mejor subconjunto de variables predictoras se lleva acabo mediante alguna de las conocidas técnicas del Análisis de Regresión Lineal Múltiple, como el procedimiento Stepwise, Forward o Backward, etc.

4.2.7.3. ***El Análisis Discriminante***

- ***La lógica del Análisis Discriminante***

Dada una variable dependiente cualitativa (criterio) que clasifica los individuos en alguna de las categorías que la componen y un conjunto de una o más variables (en general p variables) cuantitativas independientes (predictores), el análisis discriminante consiste en obtener unas funciones lineales de las variables independientes (las *funciones discriminantes*) que permitan clasificar a los individuos en alguno de los grupos establecidos por la variable dependiente.

La expresión de la función discriminante tiene la forma:

$$D_s = B_{s1} \cdot X_1 + ... + B_{sp} \cdot X_p, \qquad \textbf{(4.21)}$$

Para el i-ésimo individuo, con i = 1, 2, ..., n, las puntuaciones de las funciones discriminantes serán:

$$d_{is} = B_{s1} \cdot x_{i1} + ... + B_{sp} \cdot x_{ip}, \qquad \textbf{(4.22)}$$

A partir de las puntuaciones discriminantes, un individuo i, para el cual se conoce a qué grupo pertenece, será clasificado en uno de ellos. El porcentaje de casos correctamente clasificado será un índice de la efectividad de las funciones discriminantes. Si estas funciones son efectivas en una muestra observada, esperamos que también lo sean al momento de clasificar un individuo para el cual no se conoce su grupo de pertenencia.

Se trata entonces de buscar las funciones dadas por la ecuación (4.22) de tal forma que a partir de las puntuaciones de estas funciones podamos establecer la probabilidad que un individuo pertenezca a alguno de los grupos establecidos por alguna variable dependiente de clasificación.

- ***Selección de Variables***

La técnica de selección de las variables que se utiliza es muy parecida a la que comúnmente se practica con el Análisis de Regresión Múltiple, salvo por el hecho que en aquel caso los valores de la variable dependiente no están agrupados como ahora. Es decir que con la Regresión Múltiple, podemos estimar directamente el valor de la variable dependiente mientras que con el Análisis Discriminante, estimamos la probabilidad de pertenencia a alguno de los grupos y en función de esas probabilidades luego estimaremos a cuál de estos grupos pertenece cada individuo.

Es usual poner en práctica un procedimiento por pasos (Stepwise) para la selección de las variables que intervendrán. En el Análisis de Regresión la idea básica es incluir aquel subconjunto de variables independientes que mayor información aporte sobre los valores de la variable dependiente. En el Análisis Discriminante la cuestión será seleccionar aquel subconjunto de variables independientes que mejor discrimine los grupos establecidos por la variable de clasificación. En la regresión lineal múltiple el criterio consiste en seleccionar la primera variable con la mayor correlación lineal simple y en los sucesivos pasos utilizar el criterio de la mayor correlación parcial. En el caso del Análisis Discriminante, el criterio a utilizar para seleccionar una variable será el de la Lambda de Wilks (aunque es posible disponer de otros).

Supongamos que para predecir el valor de la variable dependiente sólo dispusiéramos de una variable independiente. Si ocurriera que las medias de esta variable en cada grupo establecido por la variable de clasificación fueran muy distintas entre sí y por otra parte el comportamiento dentro de cada grupo fuera muy homogéneo, con pocos valores dispersos y próximos a la media, los grupos estarían separados y dicha variable predictora sería una buena variable discriminante. En consecuencia puede pensarse que el criterio consiste en buscar aquellas variables que mejor separan los grupos en el sentido descripto. Generalizando este criterio podemos pensar en seleccionar aquel subconjunto de variables independientes de tal modo que al proyectar todo el conjunto de observaciones en el subespacio generado por los valores de esas variables, los centros de los grupos estuvieran muy separados entre sí y dentro de cada grupo el comportamiento fuera homogéneo.

La Lambda de Wilks para un conjunto de p variables independientes mide las desviaciones dentro de cada grupo respecto a las desviaciones totales sin distinguir grupos en el espacio p dimensional generado por los valores de las p variables. Si su valor es pequeño, la variabilidad total será debida a las diferencias entre grupos y por lo tanto el conjunto de variables correspondiente discriminará bien a los grupos. Por otra parte si su valor es próximo a 1, los grupos estarán mezclados y ese conjunto de variables no será adecuado para construir las funciones discriminantes.

- ***Extracción de las Funciones Discriminantes***

Ya hemos dicho que la cuestión radica en extraer, a partir de los datos de las nxp observaciones, un nuevo espacio de pequeña dimensión, de tal forma que al proyectar la nube de puntos sobre dicho espacio, los puntos correspondientes a individuos de un mismo grupo estén próximos y los correspondientes a individuos de grupos diferentes estén alejados. Los ejes de este nuevo espacio serán las funciones discriminantes. El espacio se extraerá de la siguiente forma:

- el primer eje o función discriminante del nuevo espacio será el que más discrimine a los grupos.
- el segundo, de entre todos los posibles perpendiculares al primero, será aquel que junto con el primero mejor discrimine los grupos.
- En general el s-ésimo eje será de entre todos los perpendiculares a todos los anteriores, aquel que mejor discrimine los grupos conjuntamente con todos los precedentes.

Sabemos que si el número de grupos es k, entonces el máximo número de funciones discriminantes será c = mín (k-1 ; p) donde p es el número de variables independientes.

Pero debe advertirse que el hecho que dicho máximo sea igual a "c" no significa que al representar los individuos en el nuevo espacio c-dimensional los grupos estén completamente separados, ya que la mayor o menor separación de los grupos dependerá de la capacidad discriminante del conjunto de variables independientes seleccionado. Podemos calcular el valor de Lambda de Wilks para ese conjunto de variables y si ese valor es pequeño, al representar los individuos en el espacio de las variables, los grupos estarán separados y por lo tanto también estarán separados en el espacio de las funciones discriminantes. Es posible además que con la representación de las s primera funciones discriminantes baste para discriminar los grupos, pudiéndose eliminar las (c-s) funciones discriminantes restantes.

Para aclarar las ideas digamos que si se dispone de las puntuaciones discriminantes para cualquier individuo i-ésimo. Ahora bien, si las medias de las puntuaciones discriminantes correspondiente a la función discriminante D_s en los grupos establecidos por la variable de clasificación son muy distintas entre si y si dentro de cada grupo el comportamiento es homogéneo, entonces esa s-ésima función discriminará bien los grupos. En este caso también es posible utilizar el estadístico Lambda de Wilks para comparar las desviaciones entre grupos con las desviaciones dentro de los grupos.

En el caso del conjunto de las funciones discriminantes, la Lambda de Wilks mide las desviaciones de las puntuaciones discriminantes dentro de los grupos respecto de las desviaciones totales sin distinguir grupos. Si su valor es grande, próximo a 1, la dispersión será debida a las diferencias dentro de los grupos y en consecuencia al representarlos en el espacio de las funciones discriminantes, los grupos estarán poco separados. El valor de la Lambda de Wilks para el conjunto de las funciones discriminantes coincide con el valor de este estadístico que corresponde al conjunto de variables independientes seleccionadas.

Generalmente, los programas de computación estadísticos, ofrecen la posibilidad de contrastar la hipótesis nula que el conjunto de funciones discriminantes no aporta información suficiente para clasificar a los individuos, calculando en cada caso un valor de Chi – cuadrado y su significación. De esta forma es posible decidir con qué conjunto de funciones discriminantes debemos quedarnos.

- ***Autovalores asociados a las Funciones Discriminantes***

Hasta donde se explicó precedentemente, sabemos cuál es el conjunto de las funciones discriminantes significativo, pero no sabemos que proporción de la información se puede atribuir a cada función discriminante del conjunto. Para esto disponemos de dos medidas: la correlación canónica y los autovalores asociados a cada función discriminante.

La correlación canónica mide las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones totales sin distinguir grupos. Un autovalor mide las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones dentro de los grupos.

En ambos casos si el valor obtenido es grande (en el caso de la correlación canónica próximo a 1) implica que la dispersión se debe a diferencias entre grupos y por lo tanto la función discriminará mucho los grupos.

Es evidente que los valores de la correlación canónica y del autovalor decrecen desde la primera función hasta la última. Un autovalor asociado a una función puede interpretarse como la parte de la variabilidad total de la nube de puntos proyectada sobre el conjunto de funciones discriminantes que es atribuible a esa función discriminante.

- ***Clasificación de los individuos***

Veamos ahora como obtener una regla que permita clasificar un individuo en uno de los k grupos definidos por la variable dependiente, a partir de sus puntuaciones discriminantes, es decir, los valores que se obtienen al reemplazar los valores de las variables independientes especializadas para el i-ésimo individuo en las funciones discriminantes. Una técnica muy utilizada es la Regla de Bayes:

La probabilidad estimada que un individuo i, con puntuaciones discriminantes: d_{i1} ; d_{i2} ; ... ; d_{ic} pertenezca al grupo j se denota como $P(G_j|D)$ y se calcula como:

$$P(G_j \mid D) = \frac{P(D \mid G_j) \cdot P(G_j)}{\sum_{j=1}^{k} P(D \mid G_j) \cdot P(G_j)} \quad j = 1,...,k, \qquad \textbf{(4.23)}$$

donde $D = (d_{i1} ; d_{i2} ; ... ; d_{ic})$; $P(G_j)$ es la probabilidad de pertenecer al grupo "j" y $P(D|G_j)$ es la probabilidad de que, supuesto que el individuo pertenezca al grupo "j" sus puntuaciones discriminantes sean d_{i1} ; d_{i2} ; ... ; d_{ic} .

Un individuo será clasificado en el grupo para el que la probabilidad a posteriori sea máxima, es decir será clasificado en el grupo G_j si:

$$P(G_j \mid D) = \text{máx}\{P(G_1 \mid D),...,P(G_k \mid D)\}, \qquad \textbf{(4.24)}$$

El porcentaje de casos correctamente clasificados será un índice de efectividad de la función discriminante. En ocasiones las probabilidades a priori son desconocidas pero bajo el supuesto que la muestra es representativa de la población objeto de estudio, se puede tomar el tamaño relativo de cada grupo como el valor de esta probabilidad a priori.

- ***Predicción***

Otra utilidad que se puede dar al Análisis Discriminante es la de predecir el grupo al que pertenece un individuo para el que no se conoce su pertenencia. Esto se lleva a cabo con el criterio de máxima probabilidad a posteriori.

- ***Utilidad del Análisis Discriminante en el Análisis de Validez***

Supongamos por un momento que se conoce la pertenencia a un grupo determinado (categoría del criterio) de un conjunto de individuos y que se dispone de un conjunto de tests con cuyos scores podría clasificarse a tales individuos. El análisis de la validez de los scores de estas pruebas como herramientas de clasificación puede realizarse conduciendo un Análisis Discriminante. De los resultados de este proceso se puede disponer de una tabla de doble entrada conocida como la matriz de confusión que permite apreciar la efectividad de las funciones de discriminación construidas a partir de los scores de la batería de tests.

Nota: En el Capítulo 6 se aplicará esta técnica a un ejemplo real.

4.3. VALIDEZ DE UN CONSTRUCTO

Durante la introducción de esta primera parte, hemos definido un *constructo* como un atributo psicológico desarrollado en un intento por construir teorías capaces de explicar el comportamiento humano.

La *inteligencia*, la *integración social*, etc. son conceptos teóricos que no pueden mensurarse directamente.

Lord y Novick (1968) establecieron que un constructo debería ser definido en dos niveles:

- ✓ A un nivel semántico: mediante una definición operacional que permitiera elaborar un procedimiento capaz de obtener una medición del constructo.
- ✓ A un nivel sintáctico: a través de la postulación de ciertas relaciones con otros constructos psicológicos del mismo cuerpo teórico y sus vínculos con otras variables del mundo real.

4.3.1. ETAPAS DE UN ANALISIS PARA ESTABLECER LA VALIDEZ DE UN CONSTRUCTO

Un proceso tendiente a establecer la validez de un constructo debería contener, al menos, las siguientes etapas:

- En base a una teoría establecida, formular hipótesis acerca de las características demográficas, de desempeño o sobre otros constructos, de aquellos individuos que difieren en la magnitud del desarrollo del constructo en estudio.
- Construir un instrumento de medición para registrar medidas sobre comportamientos o desempeños concretos, que sean manifestaciones observables del constructo.
- Recolectar datos reales que permitan testar las hipótesis del primer paso.
- Determinar si la evidencia muestral es consistente con las hipótesis formuladas y explicitar hasta qué punto los resultados alcanzados son explicables por otras teorías alternativas.

4.3.2. PROCEDIMIENTOS PARA LA VALIDACION DE UN CONSTRUCTO

4.3.2.1. ***La Matriz de Métodos y Constructos Múltiples***

Esta técnica fue propuesta por Campbell y Fiske (1959). Resulta útil cuando se intenta medir dos o más constructos mediante dos o más métodos.

Supongamos que dos constructos A y B se miden por dos métodos distintos 1 y 2. Los cuatro tests se aplican a un conjunto de individuos y se construye una matriz de correlaciones entre los scores de las cuatro pruebas, conocida como *Matriz de Validez para Métodos y Constructos Múltiples.*

Un ejemplo hipotético se muestra a continuación:

	A1	B1	A2	B2
A1	0,93	0,65	0,34	0,46
B1	0,65	0,90	0,35	0,78
A2	0,34	0,35	0,95	0,47
B2	0,46	0,78	0,47	0,87

Tabla 4.1. Matriz de Validez para dos métodos y dos constructos (hipotética)

Una diferencia importante entre este tipo de matrices y una matriz de correlación usual residen en que en estas últimas la diagonal principal está integrada por 1´s puesto que representan las correlaciones de una variable cualquiera consigo misma, pero aquí la diagonal principal está ocupada por los coeficientes de confiabilidad estimados del test que mide un constructo particular.

¿Qué propiedades serían esperables en una matriz de esta naturaleza? Para comenzar digamos que los valores de la diagonal principal deberían ser altos, puesto que se trata de las confiabilidades de las pruebas utilizadas. También deberían ser altas las correlaciones entre los scores de dos métodos diferentes que midan el mismo constructo, esto es las celdas intersección de A1 y A2 ; B1 y B2. Por otra parte las correlaciones de los puntajes correspondientes a diferentes constructos deberían ser bajas: A1 y B1 ; A1 y B2 ; A2 y B1 ; A2 y B2.

Una matriz de validez para dos métodos y dos constructos, ideal, debería entonces exhibir un aspecto como el siguiente:

	A1	**B1**	**A2**	**B2**
A1	alto	bajo	alto	bajo
B1	bajo	alto	bajo	alto
A2	alto	bajo	alto	bajo
B2	bajo	alto	bajo	alto

Tabla 4.2. Matriz de Validez para dos métodos y dos constructos (ideal)

Algunos autores distinguen dos tipos de validación, con el auxilio de estas matrices:

- ✓ *Validez Convergente*: queda demostrada por altas correlaciones entre los scores de diferentes métodos que miden el mismo concepto, es decir los coeficientes r_{A1A2} y r_{B1B2}. La expresión *convergente* hace referencia a que los tests *convergen* en el constructo que miden.

- ✓ *Validez Discriminante*: queda de manifiesto por las bajas correlaciones entre los scores correspondientes a diferentes conceptos: r_{A1B1} , r_{A1B2} , r_{A2B1} , r_{A2B2} y en particular cuando se miden con el mismo método: r_{A1B1} y r_{A2B2} . Estas correlaciones bajas muestran que el método discrimina entre diferentes conceptos.

La matriz de la Tabla 4.1. muestra coeficientes de confiabilidad bastante altos en la diagonal principal. Hay validez convergente cuando se mide el concepto B con ambos métodos, pero al parecer esto no ocurre con el constructo A. Ambos métodos aparentan no discriminar adecuadamente entre los conceptos A y B, pero el método 2 parece tener un poder de discriminación algo mayor.

4.3.2.2. ***El Análisis Factorial***

- ***La lógica del Análisis Factorial***

Cuando se administran varias pruebas a un conjunto de individuos, el análisis de la validez de la batería de tests debería tomar en consideración la posibilidad que algunos de éstos conformaran subconjuntos en los que los participantes se desempeñan de manera similar, sugiriendo que los tests pertenecientes a cada subconjunto miden, en realidad, un mismo constructo.

Una herramienta fundamental para este propósito es el Análisis Factorial. Esta es una técnica estadística que analiza las interrelaciones entre las variables de un conjunto e intenta explicar estos vínculos en base a un reducido número de nuevas variables inobservables subyacentes, llamadas *factores*. Una simple inspección visual a la matriz de correlaciones de un conjunto de variables puede sugerir la presencia de uno o más factores. Tomemos en cuenta las matrices de correlación que se muestran en las tablas siguientes:

		Test			
		1	**2**	**3**	**4**
	1	1,00	0,96	0,95	0,89
Test	**2**	0,96	1,00	0,93	0,90
	3	0,95	0,93	1,00	0,90
	4	0,89	0,90	0,90	1,00

Tabla 4.3. Matriz de Correlaciones para cuatro tests exhibiendo un factor (hipotética)

		Test			
		1	**2**	**3**	**4**
	1	1,00	0,94	0,01	0,02
Test	**2**	0,94	1,00	0,03	0,02
	3	0,01	0,03	1,00	0,90
	4	0,02	0,02	0,90	1,00

Tabla 4.4. Matriz de Correlaciones para cuatro tests exhibiendo dos factores no relacionados (hipotética)

		Test			
		1	2	3	4
Test	1	1,00	0,90	0,45	0,38
	2	0,90	1,00	0,50	0,42
	3	0,45	0,50	1,00	0,95
	4	0,38	0,42	0,95	1,00

Tabla 4.5. Matriz de Correlaciones para cuatro tests exhibiendo dos factores relacionados (hipotética)

En la tabla 4.3, las elevadas correlaciones entre los scores de los cuatro tests utilizados hace pensar que en realidad parecen estar mensurando un mismo factor. A partir de esta evidencia, parece innecesario aplicar las cuatro pruebas, con el consiguiente gasto de tiempo y costos, sino sólo una de ellas.

Del análisis de la matriz contenida en la Tabla 4.4, podemos suponer la existencia de dos factores, el primero medido por los tests 1 y 2 y un segundo factor detectado por los scores de las pruebas 3 y 4. Debe notarse los valores muy bajos de las correlaciones entre los tests de los dos diferentes grupos, lo que sugiere que ambos factores no se asocian entre sí.

Finalmente, a partir de la matriz que exhibe la Tabla 4.5 podríamos sostener la hipótesis de dos factores, como antes medidos por los tests 1 y 2 por un lado (primer factor) y 3 y 4 por otro (segundo factor). Sin embargo, a diferencia del caso anterior, parece existir cierta interrelación entre estos factores, puesta de manifiesto por las correlaciones entre los tests de los distintos subconjuntos.

Por lo general, no resulta sencillo determinar la cantidad de factores desde la simple inspección de la matriz de correlaciones, en ocasiones porque hay involucradas una gran cantidad de variables y en otras porque no aparece claramente un esquema que pueda interpretarse sencillamente (como en los ejemplos hipotéticos). El Análisis Factorial se convierte entonces en un instrumento invalorable, no sólo para determinar la cantidad de factores subyacentes sino también sus conexiones.

El análisis factorial arranca en el estudio de la matriz de correlaciones de las variables que se estudian y trata de determinar si las variaciones que exhiben estas asociaciones pueden ser justificadas con un menor número de categorías básicas respecto del conjunto original de variables. Una solución satisfactoria generará factores que contengan toda la información contenida en el conjunto de partida.

Supongamos a manera de ejemplo que mediante varios tests se mide: el recuerdo de las ideas, el recuerdo de los símbolos, el recuerdo de números, el recuerdo de palabras, la aplicación de algoritmos, la resolución de problemas matemáticos y la percepción de figuras abstractas. Probablemente las asociaciones entre los scores de estos tests puedan ser explicadas a partir de tres factores: un factor de razonamiento verbal, un factor de razonamiento numérico y un factor mnemotécnico.

- ***Los Factores y sus "Saturaciones"***

El Análisis Factorial pretende expresar una variable (por ejemplo los scores de un test) en su forma estándar (es decir una vez que se le ha restado la media y dividido por su desvío estándar) en términos de factores subyacentes, que como se explicó son construcciones hipotéticas. El modelo lineal es la forma más sencilla para describir una variable en función de estos factores.

Los factores pueden clasificarse como:

✓ *Factores Comunes*: son aquellos que pertenecen a más de una variable.

✓ *Factores Unicos*: son aquellos que pertenecen a una sola variable. Indican la medida en que los factores comunes no pueden explicar la varianza total de una variable y a su vez se pueden clasificar en:

- *Factores Específicos*
- *Factores de Error*

Denotemos con F a los factores comunes, con S a los factores específicos y con E a los factores que representan el error. Entonces podemos expresar una variable cualquiera *j* mediante la siguiente *ecuación factorial*:

$$z_j = a_{j1} \cdot F_1 + a_{j2} \cdot F_2 + \ldots + a_{jr} \cdot F_r + b_j \cdot S_j + c_j \cdot E_j \,. \qquad \textbf{(4.25)}$$

Como se puede observar, los coeficientes de los factores comunes (a_{j1} , a_{j2} , ...a_{jr}) tienen dos subíndices: el primero que hace referencia a la variable (*j* en este caso) y el segundo que se asocia al factor común. Los coeficientes de los factores específicos y del error, al ser únicos, sólo requieren de un solo subíndice que los vincula a la j-ésima variable.

Estos coeficientes se conocen como ponderaciones o más comúnmente como saturaciones del factor (*factor loadings*).

Si se tienen *n* variables (esto es los scores de *n* tests), puede construirse un sistema de ecuaciones lineales del tipo:

$$\begin{aligned} z_1 &= a_{11} \cdot F_1 + a_{12} \cdot F_2 + ... + a_{1r} \cdot F_r + b_1 \cdot S_1 + c_1 \cdot E_1 \\ z_2 &= a_{21} \cdot F_1 + a_{22} \cdot F_2 + ... + a_{2r} \cdot F_r + b_2 \cdot S_2 + c_2 \cdot E_2 \\ &... \\ z_n &= a_{n1} \cdot F_1 + a_{n2} \cdot F_2 + ... + a_{nr} \cdot F_r + b_n \cdot S_n + c_n \cdot E_n \end{aligned} \qquad \textbf{(4.26)}$$

A su vez, la matriz de coeficientes de este sistema, la *Matriz Factorial* es:

$$\overbrace{\left\langle \begin{matrix} a_{11} & a_{12} & ...a_{1r} \\ a_{21} & a_{22} & ...a_{2r} \\ ... & & \\ a_{n1} & a_{n2} & ...a_{nr} \end{matrix} \right.}^{1 \quad 2 \quad r} \overbrace{\left\| \begin{matrix} b_{11} & 0 & ...0 \\ 0 & b_{22} & ..0 \\ ... & & \\ 0 & 0 & ..b_{nn} \end{matrix} \right.}^{1 \quad 2 \quad n} \overbrace{\left\| \begin{matrix} c_{11} & 0 & ...0 \\ 0 & c_{22} & ..0 \\ ... & & \\ 0 & 0 & ..c_{nn} \end{matrix} \right\rangle}^{1 \quad 2 \quad n} \qquad \textbf{(4.27)}$$

El Análisis Factorial se propone encontrar la *Matriz Factorial Incompleta* , esto es la submatriz que contiene sólo las saturaciones de los factores comunes. Debe notarse que esta submatriz tiene tantas filas como variables y tantas columnas como factores comunes. Como la idea es encontrar un número reducido de factores, siempre habrá menos columnas que filas en esta submatriz..

Los datos de los que se parte en un Análisis Factorial es la Matriz de Correlaciones *R* y la cuestión es llegar a sustituir ésta por la Matriz Factorial Incompleta *F*, a través de la descomposición matricial:

$$R = F\,F'$$

Se dispone de varios métodos para lograr esta descomposición, entre ellos los más populares son el *Método diagonal* y el *Método del Centroide*.

Por el momento interesa remarcar el hecho que la Matriz de Correlaciones *R* exhibe las correlaciones entre los scores de las pruebas, mientras que la Matriz Factorial Incompleta *F* tiene como elementos las saturaciones de los factores comunes, que representan las correlaciones de cada test con cada factor (si los factores son *ortogonales* como se explicará más adelante).

A manera de ejemplo supongamos que se ha encontrado ya la matriz factorial incompleta, en un caso de seis variables y dos factores siguiente:

	Factor	
Test	I	II
1	0,90	0,10
2	0,90	0,10
3	0,90	0,10
4	0,12	0,85
5	0,12	0,85
6	0,12	0,85

Tabla 4.6. Saturaciones de dos factores comunes (ortogonales) para seis variables.

El test 5, por ejemplo, tiene una correlación baja con el primer factor, de 0,12 y una alta correlación con el segundo factor, de 0,85.

La correlación entre un par de tests cualesquiera guarda una estrecha relación con las saturaciones de ambos tests en cada factor. Para el caso de dos factores comunes, la relación es:

$$r_{ij} = a_{i1} \cdot a_{j1} + a_{i2} \cdot a_{j2} \,. \qquad \textbf{(4.28)}$$

A manera de ejemplo, la correlación entre los tests 1 y 5 se puede computar como:

$$r_{15} = 0{,}90 \cdot 0{,}12 + 0{,}10 \cdot 0{,}85 = 0{,}193 \,.$$

La ecuación (4.28) admite generalización:

$$r_{ij} = \sum_{k=1}^{r} a_{ik} \cdot a_{jk} \,. \qquad \textbf{(4.29)}$$

donde r es el número de factores comunes. En realidad no es difícil demostrar que esta expresión que representa la correlación entre dos tests no es otra cosa que el producto escalar de dos vectores, es decir cuando las saturaciones se consideran las coordenadas del *vector* z_j .

- ***La Rotación de Factores***

En el caso del ejemplo de seis tests y dos factores comunes, las saturaciones dadas por la Tabla 4.6 no son únicas. En realidad existe un número infinito de conjuntos de saturaciones que satisfacen la ecuación (4.28). Un ejemplo podría ser el siguiente:

	Factor	
Test	I	II
1	0,707	0,5656
2	0,707	0,5656
3	0,707	0,5656
4	0,68579	-0,51611
5	0,68579	-0,51611
6	0,68579	-0,51611

Tabla 4.7. Nuevo conjunto de Saturaciones alternativo al de la Tabla 4.6.

Una vez más, la correlación entre los tests 1 y 5 es:

$$r_{15} = 0{,}707 \cdot 0{,}68579 + 0{,}5656 \cdot (-0{,}51611) = 0{,}193 .$$

es decir lo mismo que antes.

Cualquiera de los infinitos conjuntos de saturaciones se puede construir por un proceso conocido como *rotación de los factores* que no es otra cosa que una transformación lineal adecuada. En el ejemplo anterior, las nuevas saturaciones para el test *j* se encontraron haciendo:

$$a^{*}_{j1} = \frac{1}{\sqrt{2}}\left(a_{j1} + a_{j2}\right),$$

$$a^{*}_{j2} = \frac{1}{\sqrt{2}}\left(a_{j1} - a_{j2}\right).$$

donde el coeficiente $1/\sqrt{2}$ se utiliza para lograr varianza unitaria.

Como se observa, con cada nuevo conjunto de saturaciones, no cambia ni el número de factores comunes ni la correlación entre cualquier par de tests.

Sin embargo, los factores ya no son los mismos en cualquier caso y por ende, la interpretación de la solución que se obtiene tampoco lo es.

Una pregunta inmediata es entonces ¿cómo se pude elegir el conjunto de saturaciones que pueda permitir una mejor interpretación?. Para Thurstone el mejor conjunto es aquel que ayuda a identificar la estructura más simple entre las variables. Esto implica que cada test debería mostrar altas correlaciones con muy pocos factores y muy bajas o nulas en los restantes.

Usualmente la primera solución que se encuentra en el Análisis Factorial se conoce como *Solución inicial no rotada* y generalmente no es sencilla de interpretar por lo que se procede a rotarla hasta encontrar la solución adecuada.

Existen dos tipos de rotaciones:

- ✓ *Rotaciones Ortogonales*: produce factores no correlacionados.
- ✓ *Rotaciones Oblicuas*: genera factores correlacionados.

- ***Factores Correlacionados***

Cuando se produce una rotación oblicua y los factores resultantes se hallan correlacionados, las saturaciones no pueden interpretarse como correlaciones (como en el caso de factores ortogonales), sino más bien como coeficientes de una ecuación de regresión múltiple estandarizada, es decir ponderaciones que muestran el efecto sobre la variable dependiente (el test en este caso) de un cambio unitario en un determinado factor, cuando se controla el otro factor.

Cuando los factores son ortogonales, la correlación entre dos tests cualesquiera viene dada por la ecuación (4.28) y su generalización, la ecuación (4.29). Para el caso de dos factores correlacionados, ésta viene dada por:

$$r_{ij} = a_{i1} \cdot a_{j1} + a_{i2} \cdot a_{j2} + a_{i1} \cdot a_{j2} \cdot \phi + a_{i2} \cdot a_{j1} \cdot \phi, \qquad \textbf{(4.30)}$$

donde ϕ es la correlación entre ambos factores.

Es claro que si la rotación es oblicua, entonces la ecuación (4.30) se reduce a la ecuación (4.28).

- ***Comunalidad, Especificidad, Singularidad, Confiabilidad, Varianza del Error y Varianza Total***

Regresemos a la ecuación (4.25) y tomemos varianza en ambos miembros de la igualdad. En el primer miembro, z_j es una variable estandarizada (con varianza unitaria), y si los factores del segundo miembro son independientes, resulta:

$$1 = a_{j1}^2 + a_{j2}^2 + ... + a_{jr}^2 + b_j^2 + c_j^2 . \qquad \textbf{(4.31)}$$

que se expresa diciendo que la suma de los cuadrados de las satuaraciones factoriales de una variable en todos sus factores independientes es igual a uno.

En Análisis Factorial es usual denominar:

✓ ***Comunalidad***: (ó Varianza Común) denotada como h^2_j , es la parte de la varianza total del test asociada a los factores comunes que explica la correlación con otras variables.

Para factores no correlacionados:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + ... + a_{jr}^2 , \qquad \textbf{(4.32)}$$

$$h_j^2 = \sum_{k=1}^{r} a_{jk}^2 .$$

Para dos factores correlacionados:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + a_{j1} \cdot a_{j2} \cdot \phi , \qquad \textbf{(4.33)}$$

mientras que para más de dos factores correlacionados se dispone de una expresión similar que generaliza (4.33).

✓ ***Singularidad:*** (ó Varianza Unica) simbolizada como u_j^2, es la porción de la varianza total asociada a los factores únicos (el factor específico S_j y el factor de error E_j) del test:

$$u_j^2 = 1 - h_j^2 . \qquad \textbf{(4.34)}$$

✓ ***Especificidad:*** (ó Varianza Específica), que se denota como b^2_j y que representa la porción de la variabilidad total debida al factor específico S_j de la prueba. De la ecuación (4.31) es claro que:

$$b_j^2 = 1 - \left(h_j^2 + c_j^2\right). \qquad \textbf{(4.35)}$$

✓ ***Confiabilidad:*** (ó Varianza Confiable), que se denota como r_{jj} y que representa la porción de la variabilidad total que puede considerarse no afectada por el error y por ende, confiable, es decir representa la confiabilidad del test:

$$r_{jj} = h_j^2 + b_j^2. \qquad \textbf{(4.36)}$$

✓ ***Varianza del Error:*** denotada como c^2_j y que representa la porción de la variabilidad total que puede considerarse no afectada por el error

$$c_j^2 = 1 - r_{jj}. \qquad \textbf{(4.37)}$$

Cuando un Análisis Factorial se lleva a cabo, es importante comparar la comunalidad del test con su confiabilidad en orden a establecer la proporción de la varianza total del test que resulta confiable pero asociada a un factor específico de la prueba. Esta cantidad, b^2_j, es claramente la diferencia entre la confiabilidad del test y su comunalidad.

Como ejemplo, consideremos las saturaciones del Test 1 en la Tabla 4.6. De esto se deriva que la comunalidad para esta prueba es:

$$h_1^2 = 0{,}9^2 + 0{,}1^2,$$

$$h_1^2 = 0{,}82.$$

Esto significa que el 82% de la variación total de esta prueba es compartida por los dos factores detectadosSupongamos que por algún método adecuado se ha estimado que la confiabilidad de este test es 0,85. Por lo tanto:

$$b_1^2 = 0{,}85 - 0{,}82,$$

$$b_1^2 = 0{,}03.$$

Sólo un 3% de su variabilidad total se asocia a un factor específico de esta prueba. Por lo tanto es razonable aceptar que el Test 1 es básicamente una medida del Factor I.

Pero supongamos que las saturaciones de este test en ambos factores hubiesen sido 0,55 en el factor I y 0,15 en el factor II. En principio, a juzgar por estos valores, seguiríamos pensando que esta prueba es una medida del Factor I, pero al calcular su comunalidad:

$$h_1^2 = 0{,}55^2 + 0{,}15^2,$$

$$h_1^2 = 0{,}325.$$

lo que significa que ahora el test comparte el 32,5% de su varianza total con ambos factores. Dado que su confiabilidad es 0,85, la parte de su variabilidad total asociada a un factor específico medido por el Test I es:

$$b_1^2 = 0{,}85 - 0{,}325,$$

$$b_1^2 = 0{,}525.$$

Este resultado indica que el Test I ahora no puede ser asumido básicamente como una medida del Factor I, sino que está influenciado fuertemente por otro factor no correlacionado con los dos factores comunes detectados por el Análisis Factorial.

4.4. COEFICIENTES DE VALIDEZ PARA LOS VERDADEROS SCORES

En el Capítulo 2, se había establecido como una de las conclusiones que se derivan del modelo y los supuestos de la Teoría Clásica de Tests, que la correlación entre los scores verdaderos de dos test es igual al cociente entre la correlación entre los scores observados de ambos tests dividido en la raíz cuadrada del producto de las correlaciones entre los scores observados de dos formas paralelas de ambos tests.

En símbolos,

$$\rho_{T_X T_Z} = \frac{\rho_{XZ}}{\sqrt{\rho_{XX'} \cdot \rho_{ZZ'}}}, \qquad \textbf{(4.38)}$$

Esta expresión que se conoce como *Corrección por Atenuación*, expresa que el coeficiente de validez entre los verdaderos scores de un test X y un criterio Z será mayor que el correspondiente a los scores observados de ambas variables, debido a la influencia de los errores de medición, en el test, en el criterio o en ambos.

En la práctica debe considerarse cuidadosamente la aplicación de esta corrección puesto que representa un incremento en la validez del test respecto de la que realmente puede exhibirse con los scores observados.

Sin embargo, la corrección por atenuación resulta de gran utilidad en una situación en la que debe decidirse cuál de dos criterios se asocia con mayor fuerza a los scores de un test y no es posible mensurar ambos criterios con la misma confiabilidad. En este escenario, es adecuado comparar los coeficientes de validez corregidos.

Supongamos, por ejemplo que se dispone de un test X y dos criterios Y y Z y se han calculado los coeficientes de validez (sin correcciones) y las confiabilidades siguientes: $r_{XY} = 0{,}50$; $r_{XZ} = 0{,}40$; $r_{XX'} = 0{,}62$; $r_{YY'} = 0{,}83$ y $r_{ZZ'} = 0{,}35$.

Entonces,

$$r_{T_X T_Y} = \frac{r_{XY}}{\sqrt{r_{XX'} \cdot r_{YY'}}},$$

$$r_{T_X T_Y} = \frac{0{,}50}{\sqrt{0{,}62 \cdot 0{,}83}},$$

$$r_{T_X T_Y} = 0{,}697.$$

$$r_{T_X T_Z} = \frac{r_{XZ}}{\sqrt{r_{XX'} \cdot r_{ZZ'}}},$$

$$r_{T_X T_Y} = \frac{0{,}40}{\sqrt{0{,}62 \cdot 0{,}35}},$$

$$r_{T_X T_Y} = 0{,}859.$$

De esto se deduce que el criterio Z está más fuertemente asociado con X que el criterio Y.

4.5. EFECTOS DE UNA SELECCIÓN DE INDIVIDUOS EN EL COEFICIENTE DE VALIDEZ

En muchas oportunidades se utilizan los scores de una prueba para seleccionar un subconjunto de los individuos que han tomado el test.

Entonces se dice que el test ha sido sometido a una *selección explícita*. Tal es el caso, por ejemplo, de pruebas que se aplican para seleccionar candidatos a empleos. Los scores de cualquier otro test correlacionado (criterio) con el anterior se dicen estar sometidos a una *selección incidental*.

Para tomar una decisión acerca de la utilidad del test como herramienta de selección, no debe perderse de vista que el coeficiente de validez que debería considerarse es el correspondiente al grupo completo y no el que surge de los cálculos que se restringen al subgrupo seleccionado.

En consecuencia, la cuestión a investigar en este punto es cómo puede lograrse una estimación del coeficiente de validez para el grupo completo (sin selección) cuando no se dispone de registros sobre el criterio para aquellos individuos que no han sido incluidos en el subgrupo seleccionado.

Designemos como X y X' los scores de una variable que ha sido sometida a una selección explícita (test), antes y después de haber procedido a la selección de individuos y como Y e Y' los scores sobre un criterio de validación para X, antes y después de la selección.

Si se utiliza X para predecir Y y a X' para predecir Y', entonces podemos escribir las ecuaciones de regresión poblacionales para ambos:

$$\hat{Y} = \rho_{YX} \cdot \frac{\sigma_Y}{\sigma_X} \cdot (X - \mu_X) + \mu_Y,$$

$$\hat{Y} = \beta_{Y.X} \cdot (X - \mu_X) + \mu_Y$$

$$\hat{Y'} = \rho_{Y'X'} \cdot \frac{\sigma_{Y'}}{\sigma_{X'}} \cdot (X' - \mu_{X'}) + \mu_{Y'}, \qquad (4.39)$$

$$\hat{Y'} = \beta_{Y'.X'} \cdot (X' - \mu_{X'}) + \mu_{Y'}$$

El costo de investigar el efecto de la selección de individuos en el coeficiente de validez es el establecimiento de dos supuestos:

✓ Los coeficientes de regresión en ambas ecuaciones (para el grupo completo y el subgrupo seleccionado) son iguales:

$$\beta_{Y.X} = \beta_{Y'.X'}$$

$$\rho_{YX} \cdot \frac{\sigma_Y}{\sigma_X} = \rho_{Y'X'} \cdot \frac{\sigma_{Y'}}{\sigma_{X'}}. \qquad (4.40)$$

✓ La varianza poblacional alrededor de la línea de regresión es la misma en ambos grupos (homocedasticidad):

$$\sigma^2_{Y.X} = \sigma^2_{Y'.X'},$$

$$\sigma^2_Y \cdot (1 - \rho^2_{YX}) = \sigma^2_{Y'} \cdot (1 - \rho^2_{Y'X'}) \qquad (4.41)$$

Debe advertirse que la presencia de un *efecto techo* (o un *efecto piso*) en el test implicaría la violación de estos supuestos. Al combinar las expresiones (4.40) y (4.41) se llega a que:

$$\frac{\sigma^2_Y - \sigma^2_{Y'}}{\sigma^2_{Y'}} = \rho^2_{Y'X'} \cdot \left[\frac{\sigma^2_X - \sigma^2_{X'}}{\sigma^2_{X'}} \right], \qquad (4.42)$$

donde el primer miembro representa el cambio relativo en la varianza de la variable que sido sometida a una selección incidental, cuando se procede a la selección de individuos y la expresión entre corchetes del segundo miembro es dicha medida pero para el predictor.

De la ecuación (4.42) se ve que los cambios serían equivalentes sólo en el caso en que ambas variables (X e Y) estuvieran perfectamente correlacionadas, mientras que no se produciría ningún cambio relativo en la varianza del criterio si ambas variables no tuvieran ninguna correlación.

Combinando las expresiones anteriores es fácil mostrar que:

$$\rho_{TX} = \frac{\rho_{T \cdot X \cdot}}{\sqrt{\rho^2_{T \cdot X \cdot} + \frac{\sigma^2_{X \cdot}}{\sigma^2_X} \cdot \left(1 - \rho^2_{T \cdot X \cdot}\right)}}, \qquad \textbf{(4.43)}$$

$$\sigma^2_T = \sigma^2_{T \cdot} \left[1 - \rho^2_{T \cdot X \cdot} \left(1 - \frac{\sigma^2_X}{\sigma^2_{X \cdot}}\right)\right], \qquad \textbf{(4.44)}$$

A partir de la ecuación (4.43) se puede estimar el coeficiente de validez para el grupo completo y mediante la (4.44) la varianza de la variable sujeta a selección incidental para el grupo completo.

A manera de ejemplo, imaginemos que en una cierta situación la correlación entre el predictor y el criterio para el subgrupo seleccionado es 0,60, la varianza del test para el grupo completo es 15 y para la selección 10 mientras que la varianza del criterio para el grupo seleccionado es 8. Entonces,

$$r_{TX} = \frac{0{,}60}{\sqrt{0{,}60^2 + \frac{10}{15} \cdot \left(1 - 0{,}60^2\right)}} = 0{,}68,$$

$$s^2_T = 8 \cdot \left[1 - 0{,}60^2 \left(1 - \frac{15}{10}\right)\right] = 9{,}44.$$

CAPITULO 5

Análisis de Item

El objetivo fundamental que se persigue al construir un test, es lograr que para una confiabilidad y validez predeterminadas, su extensión sea mínima. El proceso por el cual se intenta alcanzar este resultado se conoce como *Análisis de Item* y que involucra el cálculo y la consideración de las propiedades estadísticas de las respuestas de los individuos a un cierto item que integra la prueba.

Este capítulo repasa brevemente los parámetros que describen estas propiedades estadísticas y presenta el Análisis de Item mediante diferentes técnicas corrientes, como el uso de Indices de Dificultad y Discriminación, Indices de Confiabilidad y Validez, Curvas Características y el Análisis Factorial Finalmente se considera el Análisis de Items para tests referenciados en un criterio.

5.1. ETAPAS EN LA CONSTRUCCION DE UN TEST

Un proceso tendiente a construir un test debería, al menos, incluir los siguientes pasos para lograr su cometido:

- ✓ Establecer el dominio de contenidos que se intenta cubrir con la prueba. Esta primera fase del proceso resulta de suma importancia, puesto que de ella depende que el test presente finalmente "validez de contenido", que pueda ser acreditada por expertos que evalúen este aspecto.

- ✓ Producir items para cada una de las áreas de interés. Algunos autores sugieren que se proponga una cantidad entre dos y tres veces el número de items que incluirá la versión final del test.

- ✓ Administrar esta prueba a una muestra de individuos representativa de la población para la que se piensa el test. Buena parte de la bibliografía específica en este tema recomienda una regla práctica consistente en incluir una cantidad de individuos por lo menos entre cinco a diez veces el número de items que contenga el test.

- ✓ Elaborar un Análisis de Item en orden a seleccionar los mejores items y/o refinar algunos otros. La consideración en detalle de esta etapa constituye el núcleo de este capítulo. Básicamente, un Análisis de Item debería implicar los siguientes puntos:

  - Determinar qué propiedades deberían exhibir los scores del test.
  - Identificar los parámetros de los items que describen estas propiedades.
  - Estimar estos parámetros a partir de la muestra de individuos seleccionada.
  - Establecer los criterios para la selección y/o refinamiento de los items que compondrán la versión final del test.

- ✓ Conducir un estudio de "Validación cruzada", es decir aplicar la versión final del test a una nueva muestra de individuos para determinar si se han alcanzado los resultados deseados.

5.2. ANALISIS DE ITEM MEDIANTE INDICES DE DIFICULTAD E INDICES DE DISCRIMINACION DEL ITEM

5.2.1. El INDICE DE DIFICULTAD DE UN ITEM

Cuando se elabora un test, frecuentemente se busca diseñar un instrumento que permita explorar las diferencias entre los individuos que toman la prueba en relación a algún dominio de contenidos específico o bien en relación a algún criterio externo que se vincula estrechamente con los scores del test. De esta forma el interés reside en establecer la capacidad de la prueba como herramienta de discriminación entre los individuos con diferentes niveles de desempeño en el dominio que se evalúa.

El *Indice de Dificultad de un Item* , designado como p_i, se define como la proporción de individuos que responden correctamente al item. En consecuencia, un item cuya dificultad es 0,2 es más difícil que otro cuya dificultad sea 0,9 y aunque esta definición resulta algo contraria al sentido común, históricamente se ha mantenido en el contexto de la Teoría de Tests.

La dificultad de un item provee una forma sencilla y eficaz de elaborar un Análisis de Item. Debería observase que un item cuya dificultad se localice en alguno de los extremos del intervalo [0 , 1] debería eliminarse, o al menos modificarse sustancialmente, puesto que tanto en la situación en la que prácticamente ninguno de los individuos logra contestar correctamente el test o en la que casi todos lo hacen, no es posible conseguir la discriminación que se busca.

Para el caso de un item dicotómico, el máximo poder de discriminación se logra cuando la dificultad del item es igual a 0,5 puesto que este valor maximiza la varianza del item que resulta ser: $p_i (1- p_i)$. Sin embargo, la validez de esta proposición está asociada a los niveles de asociación que muestren los items. Supongamos que todos los items exhiben una correlación perfecta y sus dificultades son todas 0,5. Esto implicaría inmediatamente que la mitad de los individuos han obtenido 0 puntos en tanto que la otra mitad lograron el máximo score total, situación en la que tampoco se alcanza a distinguir adecuadamente entre los diferentes desempeños individuales.

Cuanto más homogéneos sean los items, esto es cuanto mayor es la correlación entre ellos, más ancho debería ser el rango del índice de dificultad, mientras que en la medida que los items sean más heterogéneos se puede reducir el recorrido de estos valores. En la tabla siguiente se muestran los valores sugeridos por Henryssen (1971) para el índice de dificultad de acuerdo al nivel de asociación entre los items medido por el promedio de la correlación biserial entre los scores de los items y el score total:

Correlación Biserial items-total	Rango del índice de Dificultad
<0,3	0,45 a 0,55
0,3 a 0,4	0,4 a 0,6
>0,4	0,3 a 0,7

Tabla 5.1 Rangos sugeridos para el Indice de Dificultad según la Correlación Biserial items-total

La elección de los niveles de dificultad también se relaciona con el formato de los items del test. Supongamos que se ha construido un test cuyos items son de la forma *Verdadero – Falso*. Podría ocurrir que un nivel de dificultad de 0,5 en un item se presente a causa que los individuos seleccionaron sus respuestas en forma aleatoria. En tal caso, la probabilidad de elegir la respuesta correcta es justamente 0,5 y no sería posible la discriminación que se pretende con la prueba. Lord (1953) sugirió que para items con formato *Elección Múltiple*, un nivel óptimo de dificultad se obtiene en

el punto medio del segmento cuyos extremos son la probabilidad de elegir al azar la respuesta correcta y 1. A manera de ejemplo, si cada item presenta cinco respuestas posibles (y sólo una es la correcta), el intervalo se extendería desde 1/5 a 1 y su punto medio es 0,6.

En pruebas en las que se establece un score de corte (*cutting score*) en orden a seleccionar un subgrupo de individuos en algún extremo de la escala, debería producirse un esfuerzo para lograr el máximo poder de discriminación alrededor del punto de corte:

- En un examen de admisión a un programa de entrenamiento para seleccionar el 5% de los individuos en el extremo superior de la escala, deberían incluirse items de suma dificultad.

- Para distinguir alumnos que necesitan de un programa de apoyo escolar se requeriría una prueba cuyos items sean bastante accesibles.

Otro índice de dificultad se presentará más adelante.

5.2.2. INDICES DE DISCRIMINACION DE UN ITEM

No debe perderse de vista que el interés está centrado en lograr una discriminación de los individuos que toman la prueba en función de los diferentes desempeños que cada uno muestra en el dominio de interés. En principio, una medida sencilla que facilita tal discriminación es el mismo score total del test que da una idea de la posición relativa de cada individuo.

En un momento posterior, la pregunta que se plantea es cuáles son aquellos items de la prueba que colaboran con esta tarea de identificación. Más propiamente se trata de localizar los items del tests que muy probablemente sean correctamente respondidos por individuos con altos scores totales y erróneamente contestados por quienes muestran totales bajos. Es evidente que un item que pudiera ser correctamente completado tanto por personas con altos o bajos puntajes no sería de gran utilidad y menos aún aquellos items que mayoritariamente son respondidos acertadamente por individuos con totales pobres, lo que constituiría una "discriminación negativa".

Los índices de discriminación intentan medir el grado en el que las respuestas a un determinado item se relacionan con las respuestas a otros items del test. Entre los más populares se cuentan el Indice Discriminante de un Item y otras medidas de asociación como la Correlación Biserial por Puntos, el Coeficiente de Correlación Biserial, el Coeficiente Phi y el Coeficiente de Correlación Tetracórica, que se comentan brevemente a continuación.

Es claro que estas medidas de asociación resultan adecuadas para situaciones en las que se trabaja con items puntuados dicotómicamente, en tanto que para casos donde sea posible debería optarse por el Coeficiente de Correlación de Pearson.

5.2.1.1. ***Indice Discriminante de un Item***

Este índice, denotado como "d_i", se define como la diferencia entre la proporción de individuos con scores totales altos que respondieron correctamente el i-ésimo item de la prueba y la proporción de individuos con scores totales bajos que lo contestaron en forma correcta.

Simbólicamente,

$$d_i = \frac{U_i}{n_{iU}} - \frac{L_i}{n_{iL}} \qquad \textbf{(5.1)}$$

donde

U_i es el número de individuos con scores totales altos que han respondido correctamente el item "i".
n_{iU} es el número de individuos con scores totales altos.
L_i es el número de individuos con scores totales bajos que han respondido correctamente el item "i".
n_{iL} es el número de individuos con scores totales bajos.

Usualmente, entre el 10% y el 33% de los puntajes de la muestra que se localizan en el extremo superior (inferior) de la escala se asumen como scores *altos* (*bajos*). Para el caso que la distribución de scores observados sea aproximadamente normal, Kelley (1939) mostró que es óptimo considerar el 27% superior e inferior en la escala. Sin embargo, tomando entre 25% y 33% de las observaciones en cada extremo se logran estimaciones muy próximas entre si para el valor de d_i.

Un inconveniente asociado al uso del Indice Discriminante es el hecho que su distribución de muestreo es muy compleja y no resulta sencillo decidir si es significativamente diferente de cero o si existe una diferencia significativa entre dos valores de éstos.

Sin embargo, a manera de regla práctica para interpretar el valor de este índice, se proponen los siguientes criterios:

Valor del Indice Discriminante d_i	Decisión sobre el item
$d_i \geq 0,40$	*Se mantiene sin revisión*
$0,39 \geq d_i \geq 0,30$	*Requiere "poca o ninguna" revisión*
$0,29 \geq d_i \geq 0,20$	*Requiere revisión*
$0,19 \geq d_i$	*Debería ser eliminado o completamente revisado*

Tabla 5.2 Decisiones posibles sobre un item según el valor del Indice Discriminante.

5.2.1.2. ***El Coeficiente de Correlación Biserial por Puntos***

Esta medida resulta útil cuando se intenta explorar la magnitud de la asociación entre un item dicotómico (puntuado como 0 ó 1) y el score total del test o algún otro criterio (variable) de naturaleza continua.

El coeficiente de Correlación Biserial por Puntos se define como:

$$\rho_{pbis} = \frac{\mu_+ - \mu_X}{\sigma_X} \cdot \sqrt{\frac{P}{1-P}} \qquad (5.2)$$

donde

μ_+ es la media poblacional de los scores totales (u otra variable continua) del subconjunto de individuos que responden correctamente el item.

μ_X: es la media poblacional de los scores totales.

σ_X: es el desvío estándar poblacional de los scores totales.

P : es la dificultad del item

En la práctica se reemplazan los parámetros por estimadores:

$$r_{pbis} = \frac{\overline{X_+} - \overline{X}}{s_X} \cdot \sqrt{\frac{p}{1-p}} \qquad (5.3)$$

Si se desea probar la significación de este coeficiente tomamos:

$$t = \sqrt{\frac{N-2}{1-r_{pbis}^2}} \qquad (5.4)$$

y comparamos este valor con el percentil adecuado de una distribución t de Student con N-2 grados de libertad.

Para muestras grandes, una aproximación conveniente para el error estándar de este estadístico es:

$$s_{r_{pbis}} = \frac{1}{\sqrt{N-1}} \qquad \textbf{(5.5)}$$

y un valor crítico puede tomarse a dos desvíos estándares a la derecha de 0, para decidir si el coeficiente es significativamente diferente de 0.

Supongamos, para ejemplificar que consideramos una muestra de 300 individuos, entonces,

$$s_{r_{pbis}} = \frac{1}{\sqrt{300-1}} = 0{,}06$$

por lo que el valor crítico sería

$$0{,}00 + 2 \cdot 0{,}06 = 0{,}12 .$$

Debe tenerse en cuenta que un aspecto cuestionable de esta medida es el hecho que el item que se considera integra el score total con el cual se lo asocia. Esto no constituye un problema serio si se trata con una escala compuesta por una gran cantidad de items. Cuando se dispone de pocos items, una expresión corregida es:

$$\rho_{i(x-i)} = \frac{\rho_{xi} \cdot \sigma_x - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_x^2 - 2 \cdot \rho_{xi} \cdot \sigma_i \cdot \sigma_x}} \qquad \textbf{(5.6)}$$

donde

$\rho_{i\,(x-i)}$: es la correlación entre el item "i" y el score total sin incluir dicho item.

σ_i : es el desvío estándar poblacional del item.

5.2.1.3. ***El Coeficiente de Correlacion Biserial***

Si se puede sostener el supuesto que la aptitud para resolver el item *i* es una variable continua que por convención se ha dicotomizado como *acierto – fracaso* (o en términos numéricos 0 ó 1) y su distribución de probabilidad es normal, entonces una medida de asociación entre el item y el score total (o algún otro criterio continuo) se logra mediante el Coeficiente de Correlación Biserial:

$$\rho_{pbis} = \frac{\mu_{+} - \mu_{X}}{\sigma_{X}} \cdot \frac{P}{Y} \qquad \textbf{(5.7)}$$

donde todas los parámetros se interpretan en forma similar a la ecuación (5.2) e *Y* es la ordenada de la curva normal estándar del percentil asociado al valor de la dificultad del item.

A manera de ejemplo supongamos que la dificultad del item es $P = 0{,}60$, entonces la ordenada de la curva normal del punto $z = 0{,}60$ es 0,3867.

Una estimación de este parámetro se logra reemplazando estimadores adecuados en la ecuación (5.6):

$$r_{pbis} = \frac{\overline{X_{+}} - \overline{X}}{s_{X}} \cdot \frac{p}{y} \qquad \textbf{(5.8)}$$

El error estándar para este estimador se puede aproximar mediante:

$$s_{r_{bis}} = \frac{\sqrt{\frac{p \cdot (1-p)}{N-1}}}{Y} \qquad \textbf{(5.9)}$$

Como en el caso anterior, es posible utilizar en forma análoga esta aproximación del error estándar para establecer un valor crítico que permita decidir si el coeficiente es significativamente diferente de cero.

5.2.1.4. ***El Coeficiente Phi***

Si el caso es que el item es dicotómico, pero el criterio con el que se lo asocia también lo es (el individuo es clasificado por género o como *éxito – fracaso*, etc.) es posible organizar la información en una tabla de doble entrada compuesta por dos filas y dos columnas que corresponden a las categorías del item y a las del criterio, ambos dicotómicos:

		Criterio		
		Éxito	Fracaso	Total
Item	0	A	B	A+B
	1	C	D	C+D
	Total	A+C	B+D	A+B+C+D

Tabla 5.3 Frecuencias de casos de un item y un criterio dicotómicos

Entonces puede computarse el Coeficiente Phi:

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(B+D)(A+C)(C+D)}} \qquad \textbf{(5.10)}$$

La significación de este coeficiente de decide comparando el valor de:

$$\chi^2 = N \cdot \phi^2 \qquad \textbf{(5.11)}$$

con el percentil adecuado de una distribución chi – cuadrado con 1 grado de libertad.

Para muestras grandes, la ecuación (5.5) ofrece una aproximación del error estándar de este estimador que puede utilizarse para determinar el valor crítico que permite decidir la significancia del valor observado.

Para ilustrar el uso de este coeficiente supongamos que los datos disponibles relacionados con los resultados en un item de una evaluación en matemática y el género del alumno se han agrupado en la siguiente tabla:

		Género		
		varones	mujeres	Total
Item	0	19	11	30
	1	5	15	20
	Total	24	26	50

Tabla 5.4 Ejemplo hipotético de los resultados en un item clasificados por género.

El valor del Coeficiente Phi en este caso es:

$$\phi = \frac{19 \cdot 15 - 11 \cdot 5}{\sqrt{30 \cdot 20 \cdot 24 \cdot 26}} = 0{,}376.$$

Calculando:

$$\chi^2 = 50 \cdot 0{,}376^2 = 7$$

que resulta significativo al 1% pues $\chi^2_{(0{,}99\ ;\ 1g.l.)}$ = 6,635, lo que indica una asociación entre las categorías de la tabla.

Finalmente, es oportuno establecer que si el item que se investiga discrimina razonablemente bien, se espera una asociación positiva significativa con el score total. Un valor negativo para este coeficiente podría ser una señal que se ha cometido un error al asignar los puntajes del item o bien que los individuos tuvieron problemas en su interpretación y debería eliminarse o revisarse profundamente este item.

5.2.1.5. ***El Coeficiente de Correlación Tetracórica***

Es útil cuando se dispone de datos en una tabla de doble entrada con dos filas y dos columnas y es razonable suponer que ambas variables dicotomizadas son en realidad continuas, con distribución normal y linealmente relacionadas.

El cálculo de este coeficiente es bastante laborioso e incluye una serie de potencias del mismo. Algunos términos de esta serie, en su versión muestral, son:

$$r_t + r_t^2 \cdot \frac{z \cdot z'}{2} + r_t^3 \cdot \frac{\left(z^2 - 1\right) \cdot \left(z'^2 - 1\right)}{6} + \ldots + \frac{ad - bc}{y \cdot y' \cdot N^2} \qquad \textbf{(5.12)}$$

donde

r_t : es el coeficiente de correlación tetracórica muestral.

z y z': son los valores en puntajes z correspondientes a la abscisa para los puntos correspondientes a "p" y "1-p", donde p es la dificultad del item.

y e *y'* : son los valores correspondientes a las ordenadas de la curva normal

a,b,c,d : representan las frecuencias en las celdas de la Tabla 5.3 para la muestra seleccionada.

Para evitar la complejidad del cómputo de la serie anterior, se han propuesto varios métodos que estiman el valor del Coeficiente de Correlación Tetracórica. Uno de éstos se conoce como "Fórmula del Coseno de Pi" y consiste en calcular:

$$r_{\cos\pi} = \cos\left(\frac{180°}{1+\sqrt{\frac{ad}{bc}}}\right) \qquad (5.13)$$

Como ejemplo supongamos que los resultados de dos items dicotómicos, para 500 alumnos, se han organizado en la siguiente tabla de doble entrada:

		Item 1		
		1	**0**	**Total**
Item 2	**1**	187	127	314
	0	85	101	186
	Total	272	228	500

Tabla 5.4 Ejemplo hipotético de los resultados en dos items dicotómicos.

Con la fórmula anterior,

$$r_{\cos\pi} = \cos\left(\frac{180°}{1+\sqrt{\frac{187\cdot 101}{127\cdot 85}}}\right) = 0{,}22$$

En la práctica, el uso de este coeficiente de correlación se limita a casos en los que no resulta apropiado calcular el Coeficiente Phi, como por ejemplo cuando estos valores de correlación integran una matriz de correlaciones que será utilizada en un Análisis Factorial. En tal situación se sugiere utilizar Coeficientes de Correlación Tetracórica.

5.2.3. UN EJEMPLO DE USO DE LOS INDICES DE DISCRIMINACION DE ITEMS

Imaginemos por un momento que se dispone de los resultados de un test compuesto por items con formato *elección múltiple* y se requiere proceder a la realización de un Análisis de Item. Como se vio, resultan de gran importancia como guías para la selección del subconjunto de items que logra la mayor discriminación entre los individuos el índice de dificultad y el índice discriminante de un item (o alguna otra medida de correlación adecuada).

Por lo general, también resulta provechoso considerar el porcentaje de individuos de cada subgrupo, cuando éstos son clasificados en categorías según su rendimiento general en la prueba, que eligen cada respuesta posible en un item. Las respuestas incorrectas de un item se conocen como *distractores* y en muchas ocasiones cuando se construye un test, los distractores que se utilizan auxilian en la tarea de identificar el aspecto específico del contenido que se investiga en el que el individuo presenta sus mayores dificultades.

En las tablas siguientes se muestran cuatro casos hipotéticos, correspondientes a distintas situaciones que podrían tener lugar durante un análisis de item, habiendo clasificado a los individuos en alguno de los tres grupos según el rendimiento general del test y de acuerdo a la opción seleccionada.

ITEM 1				
		Rendimiento Bajo	**Rendimiento Medio**	**Rendimiento Alto**
Opciones de respuestas	**a**	0,29	0,40	0,12
	b	0,50	0,32	0,09
	#c	0,15	0,20	0,72
	d	0,06	0,08	0,07
$p_i = 0,36$ $d_i = 0,57$ $r_{pbis} = 0,52$				

Tabla 5.5 Ejemplo hipotético de resultados en un item aceptable por grupo de rendimiento y opciones

ITEM 2				
		Rendimiento Bajo	Rendimiento Medio	Rendimiento Alto
Opciones de respuestas	#a	0,80	0,31	0,05
	b	0,00	0,00	0,00
	c	0,19	0,43	0,51
	d	0,01	0,26	0,44
$p_i = 0,39$ $d_i = -0,75$ $r_{pbis} = -0,42$				

Tabla 5.6 Ejemplo hipotético de resultados en un ítem "pobre" por grupo de rendimiento y opciones

ITEM 3				
		Rendimiento Bajo	Rendimiento Medio	Rendimiento Alto
Opciones de respuestas	a	0,13	0,04	0,01
	b	0,10	0,06	0,00
	c	0,02	0,10	0,01
	#d	0,75	0,80	0,98
$p_i = 0,84$ $d_i = 0,23$ $r_{pbis} = 0,20$				

Tabla 5.7 Ejemplo hipotético de resultados en un ítem "muy fácil" por grupo de rendimiento y opciones

ITEM 4				
		Rendimiento Bajo	Rendimiento Medio	Rendimiento Alto
Opciones de respuestas	#a	0,00	0,02	0,05
	b	0,56	0,48	0,15
	c	0,23	0,03	0,60
	d	0,21	0,47	0,20
$p_i = 0,02$ $d_i = 0,05$ $r_{pbis} = 0,06$				

Tabla 5.8 Ejemplo hipotético de resultados en un ítem "muy difícil" por grupo de rendimiento y opciones

En cada caso, la respuesta correcta se identifica con el símbolo cardinal, las tres restantes son los distractores del item.

El ITEM 1 podría ser juzgado como razonable, puesto que su dificultad es 0,36, su índice discriminante es 0,57 y su coeficiente de Correlación Biserial por puntos con el score total es 0,52 (mucho mayor que el valor crítico de 0,12 para una muestra de 300 individuos). Los distractores *a* y *b* cumple su función adecuadamente y en menor medida el distractor *d*.

El signo negativo del Indice Discriminante y del Coeficiente de Correlación Biserial del ITEM 2 evidencian que se trata de un item bastante pobre. El distractor *b* no cumple ninguna función. Los alumnos con scores totales altos muestran una tendencia hacia los distractores *c* y *d*. Este item requiere de una profunda revisión o de lo contrario simplemente debería ser eliminado.

El ITEM 3 resulta demasiado accesible para el grupo. Esto se refleja en que un gran porcentaje de individuos en los tres grupos de rendimiento han seleccionado la respuesta correcta. En consecuencia, este item no es útil en la discriminación de los participantes de la prueba.

Como contraparte al anterior, el ITEM 4 parece ser muy difícil para el nivel de desempeño de los individuos que tomaron el test, lo que se deduce de los escasos porcentajes de alumnos que lo han contestado correctamente en los tres grupos. Como el anterior, estos item son inservibles como instrumentos de discriminación y deberían ser modificados sustancialmente o eliminados del test.

5.3. ANALISIS DE ITEM MEDIANTE INDICES DE CONFIABILIDAD Y VALIDEZ PARA UN ITEM

Tanto el Indice de Confiabilidad del Item como su Indice de Validez son funciones de su variabilidad propia y de la magnitud de su asociación con el criterio.

En el supuesto que el criterio no sea otro que el score total de la prueba, entonces el *Indice de Confiabilidad* del Item *i* se define como:

$$\rho_{ix} = \sigma_i \cdot \rho_{ix} \qquad \textbf{(5.14)}$$

donde

σ_i : es el desvío estándar del item

ρ_{Xi} : es el Coeficiente de Correlación Biserial por Puntos entre el item *i* y el score total de la prueba.

El *Indice de Validez* del Item *i* se define como:

$$\rho_v = \sigma_i \cdot \rho_{Yi} \qquad \textbf{(5.15)}$$

donde

ρ_{Yi} : es el Coeficiente de Correlación Biserial por Puntos entre el item *i* y el criterio de validación externo.

En ambos casos, en la práctica, se obtienen estimaciones de estos parámetros sustituyendo estimadores adecuados en el segundo miembro de (5.14) y (5.15).

Supongamos que se dispone de los resultados de una prueba de N items. Se discutirá ahora cómo seleccionar los k mejores items de forma que se maximice la confiabilidad (consistencia interna) o la validez referenciada en un criterio de la versión final del test.

Cuatro estadísticos relacionados con cada item son de gran utilidad:

- ✓ La dificultad del item p_i.
- ✓ La desviación estándar del item s_i.
- ✓ El Indice de Confiabilidad del Item $s_i r_{Xi}$.
- ✓ El Indice de Validez del Item $s_i r_{Yi}$.

Entonces, es posible calcular, para la versión final del test compuesta por "k" items los siguientes estadísticos:

- La *Media* del test:

$$\overline{X} = \sum_{i=1}^{k} p_i \qquad \textbf{(5.16)}$$

- El *Desvío Estándar* del test:

$$s_X = \sum_{i=1}^{k} s_i \cdot r_{Xi} \qquad \textbf{(5.17)}$$

- La *Confiabilidad* del test:

$$r_{xx'} = \frac{k}{k-1} \cdot \left[1 - \frac{\sum_{i=1}^{k} s_i^2}{\left(\sum_{i=1}^{k} s_i \cdot r_{xi} \right)} \right] \qquad (5.18)$$

- La *Validez* del test:

$$r_{xT} = \frac{\sum_{i=1}^{k} s_i \cdot r_{Ti}}{\sum_{i=1}^{k} s_i \cdot r_{xi}} \qquad (5.19)$$

A partir de la ecuación (5.18) es claro que si se desea maximizar la confiabilidad del test compuesto por los k items seleccionados, los coeficientes de correlación biserial por puntos entre los items y el score total deben ser máximos, puesto que de esta forma se haría mínimo el cociente dentro de los corchetes.

Para auxiliar en la selección de los items que exhiben máxima correlación biserial por puntos con los scores totales, es posible graficar en un par de ejes cartesianos las desviaciones estándar versus los índices de confiabilidad de los items como se ve en la figura 5.1. Ahora es claro que deben seleccionarse los items marcados con puntos más intensos, toda vez que para un dado valor del desvío estándar, éstos puntos muestran los máximos valores observados del coeficiente de confiabilidad del item.

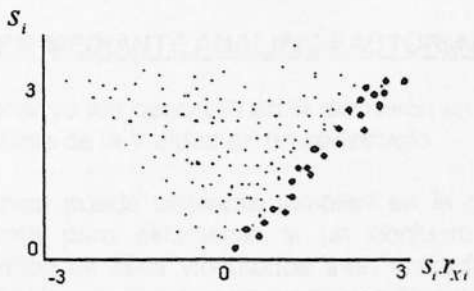


Figura 5.1. Selección de Items para lograr máxima confiabilidad

Por otra parte, la ecuación (5.19) muestra que para maximizar la validez del test, la suma de los Indices de Validez es próxima a la suma de los Indices de Confiabilidad de la prueba (notar que se trata de un coeficiente de correlación y por lo tanto es un número comprendido en el intervalo [-1,1] lo que significa que el valor absoluto del numerador es siempre menor o igual que el del denominador).

Los items que logran este efecto son los que se destacan en la figura 5.2., en la que se grafican los Indices de Validez versus los de confiabilidad de los items, y no son otros que los se aproximan a la primera bisectriz del gráfico.

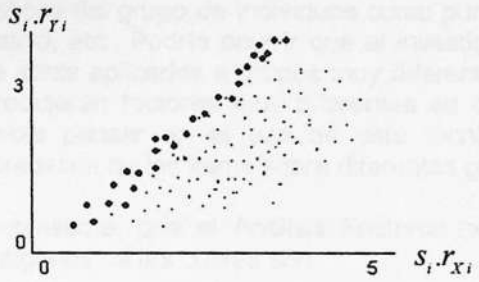


Figura 5.2. Selección de Items para lograr máxima validez

Si el objetivo es lograr una maximización simultánea de la confiabilidad y de la validez del test pueden generarse algunas dificultades, puesto que en muchas ocasiones la selección de un subconjunto de items que logra una validez máxima también implica una disminución en la medida de la confiabilidad (de consistencia interna) respecto de la que se podría obtener mediante otro subconjunto. Sin embargo es posible en este caso conseguir una medida aceptable de confiabilidad de formas alternas o paralelas.

5.4. ANALISIS DE ITEM MEDIANTE ANALISIS FACTORIAL

El Análisis Factorial ya fue descripto en la discusión en la Sección 4.3.2.2. en el contexto del Análisis de la Validez de un *constructo.*

Esta misma técnica puede utilizarse también en la construcción de un test, como herramienta para establecer si un conjunto de items resulta *homogéneo* en el sentido de estar vinculados a un único factor subyacente y facilitar la identificación de aquellos que no participan de esta homogeneidad.

Por lo general esta técnica es útil cuando se trabaja con una gran cantidad de items y la inspección visual de la matriz de correlaciones se hace engorrosa y complicada.

En algunas situaciones, como consecuencia del Análisis Factorial se generan un par de factores, uno de los cuales exhibe saturaciones altas para todos los items, salvo unos pocos que muestras coeficientes elevados en el otro factor. En tal caso, debería considerarse un profundo análisis de los items marginales para su eliminación o revisión sustancial.

Los factores emergentes de un Análisis Factorial pueden ser influenciados por distintas características del grupo de individuos como por ejemplo la edad, el sexo, el nivel educativo, etc.. Podría ocurrir que al investigar los resultados del mismo conjunto de items aplicados a grupos muy diferentes en algunos de estos aspectos, se produjeran factores muy diferentes en cada caso. Como consecuencia es posible pensar en el uso de esta técnica para analizar diferencias en la interpretación de los items sobre diferentes grupos.

Debería tenerse presente, que el Análisis Factorial no está exento de dificultades prácticas, algunas de las cuales son:

- ✓ El tipo de coeficiente de correlación que se utilice para construir la Matriz de Correlaciones de la que se parte, puede condicionar el número de factores significativos que se producen.

- ✓ La adición de un nuevo item o la eliminación de otros puede provocar alteraciones importantes en las saturaciones de los factores o en el número de factores comunes.

- ✓ La aplicación de diferentes técnicas para la extracción del espacio factorial puede conducir a diferentes resultados y en todo caso el investigador debería estar en condiciones de justificar teóricamente la elección de un determinado procedimiento.

5.5. ANALISIS DE ITEM MEDIANTE SU CURVA CARACTERISTICA

Una *Curva Característica de un Item* (ICC) es un gráfico en el que se representa la probabilidad de responder correctamente ese item versus los verdaderos scores de cada individuo en relación al constructo que se analiza.

Pero dado que usualmente no se dispone de estos últimos, la Curva Característica se estima localizando las proporciones de individuos que responden correctamente ese item (en el eje vertical) versus los scores totales del test.

La Figura 5.3. muestra una curva característica estimada de un item. Se ve que la poligonal presenta una pendiente positiva, esto es a mayor score total sobre el test, mayor es la probabilidad de responder con acierto al item en cuestión.

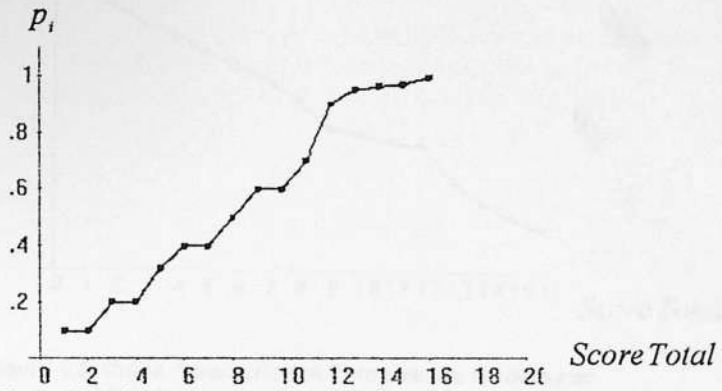


Figura 5.3. Curva Característica Estimada de un item

La Curva Característica de la Figura 5.4 corresponde a un item con escaso poder de discriminación: para cualquier puntaje total sobre el test, existe la misma probabilidad de responder correctamente este item.

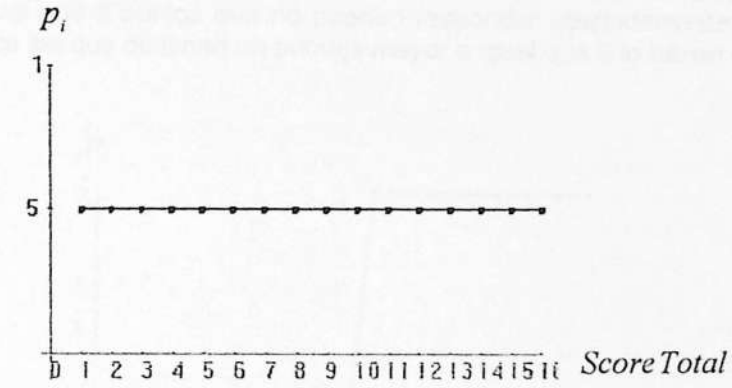


Figura 5.4. Curva Característica Estimada de un item con muy poco poder de discriminación.

La Curva Característica siguiente, de la Figura 5.5. está asociada a un item con una discriminación negativa. Obviamente tal item debería eliminarse del test.

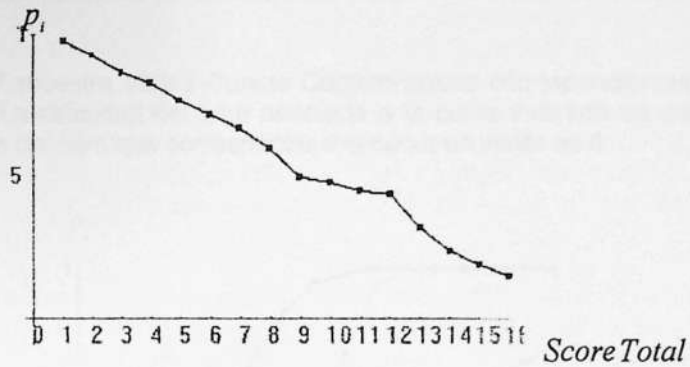


Figura 5.5. Curva Característica Estimada de un item con discriminación negativa.

La pendiente de cada tramo de la poligonal puede considerarse una medida del poder discriminativo del item, en el sentido que a mayor capacidad de discriminación, la pendiente debe ser mayor. El caso extremo se muestra en la Figura 5.6. Este item distingue perfectamente los individuos con un score total menor o igual que 8 puntos que no pueden responder acertadamente el item, mientras que los que obtienen un puntaje mayor o igual que 9 lo hacen en forma segura.

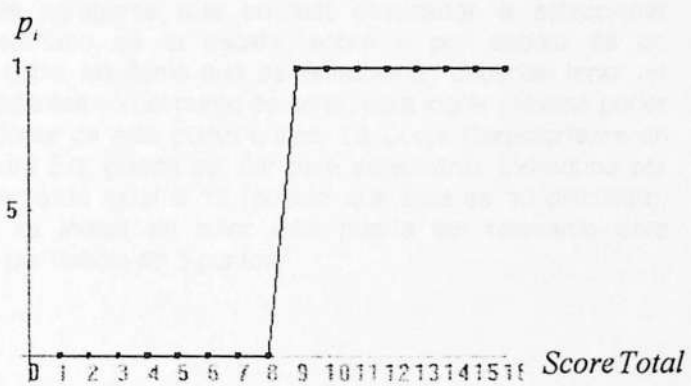


Figura 5.6. Curva Característica Estimada de un item con gran poder de discriminación.

A partir de una Curva Característica se puede obtener una medida de la dificultad del item, si se define ésta última como el score total que corresponde a una probabilidad de contestar correctamente el item igual a 0,5. Esta medida aumenta con el incremento de la dificultad del item, a diferencia de la medida p_i.

La Figura 5.7 muestra varias Curvas Características correspondientes a diferentes items. La dificultad del item asociado a la curva indicada en color azul es igual a 5, la del item que corresponde a la curva en verde es 8.

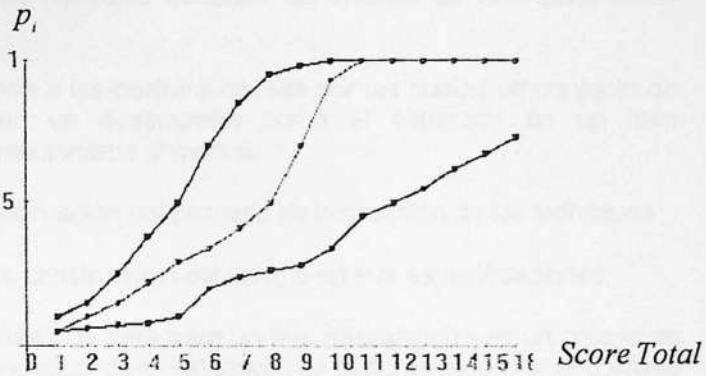


Figura 5.6. Curvas Características Estimadas de tres items

Finalmente, puede agregarse que en test destinados a seleccionar individuos en algún extremo de la escala, sobre o por debajo de un determinado punto de corte, los items que se seleccionen deberían tener un nivel de dificultad coincidentes con el punto de corte, para lograr máximo poder de discriminación alrededor de este punto crítico. La Curva Característica en color violeta de la Figura 5.6. puede ser útil para seleccionar individuos por encima de un punto de corte igual a 12 (puesto que éste es su dificultad), mientras que la que se indica en color azul podría ser relevante para seleccionar candidatos por debajo de 5 puntos.

5.6. ANALISIS DE ITEM PARA TESTS REFERENCIADOS EN UN CRITERIO

Como se explicó en la introducción de este trabajo, en el caso de un test referenciado en un criterio, interesa medir el grado en que un individuo desarrolló cierta aptitud como producto de un proceso de instrucción, sin hacer referencias a los desempeños que exhiben otros individuos. Una cuestión central está vinculada a la validez de su contenido, puesto que no debe perderse de vista que una prueba de este tipo consiste en una muestra representativa de un dominio específico de contenidos. Es claro entonces que la opinión de expertos sobre la validez del test juega un papel muy importante. Sin embargo es también relevante conducir un análisis de item para estas pruebas.

Si se hace referencia a las posibles causas por las cuales un conjunto de individuos no presentan un desempeño como el esperado en un item determinado, debería mencionarse al menos:

- ✓ Falta de adecuación del proceso de instrucción de los individuos
- ✓ Fallas en la construcción del item, o en sus especificaciones.

La meta de un Análisis de Item para un test referenciado en un criterio es la identificación de la magnitud de la influencia de factores externos al dominio específico de contenidos en el desempeño de los individuos sobre un item. A partir de este punto se explica que los estadísticos que describen el comportamiento en un item no deberían estar afectados por la varianza de los scores, ya que no se persigue una discriminación entre los individuos que participan de la prueba.

5.6.1. LA DIFICULTAD DE UN ITEM

La definición de la Dificultad del Item *i* es análoga a la expuesta en la Sección 5.2.1., esto es la proporción de individuos que responden correctamente ese item.

Lo importante en este momento es que se abandona el criterio de seleccionar un subconjunto de items que maximice la varianza de los scores de los individuos.

En muchas ocasiones es de utilidad calcular la media o mediana de las dificultades de los items que integran un grupo (*cluster*) que en teoría deberían medir un mismo objetivo, ya que esta medida puede dar una idea razonable sobre la efectividad del proceso de instrucción y/o la adecuación de la construcción de estos items.

Es muy común la práctica de aplicar un pretest, como instancia previa a la instrucción y luego un posttest para evaluar la efectividad del entrenamiento. Si se detecta en un pretest que ciertos items son demasiado fáciles, probablemente el instructor debería cuestionarse la necesidad real de incluir estos objetivos en el programa instruccional o si los resultados de un posttest indican que determinados items tienen una dificultad muy alta, podría ser una evidencia seria de una falla durante el proceso de entrenamiento o de la inclusión en la prueba de contenidos que no fueron cubiertos durante el programa de enseñanza.

5.6.2. LA SENSITIVIDAD AL PROCESO DE INSTRUCCION

Una medida de la sensitividad al proceso de instrucción, esto es la discriminación entre individuos que han recibido entrenamiento de los que no lo han hecho, está dada por la diferencia entre las dificultades del posttest y el pretest:

$$D = p_{post} - p_{pre} \qquad \textbf{(5.20)}$$

Un procedimiento alternativo (Brenan, 1.972) en el que se debe especificar un score de corte para identificar a los individuos que han alcanzado una suficiencia en la destreza de interés viene dado por:

$$B = \frac{U}{n_1} - \frac{L}{n_2} \qquad \textbf{(5.21)}$$

donde

- U es el número de individuos sobre el score de corte que respondieron correctamente el item.
- n_1 es el número de individuos sobre el score de corte.
- L es el número de individuos bajo el score de corte que respondieron correctamente el item.
- n_2 es el número de individuos bajo el score de corte.

Debe notarse que tanto para D como para B cuanto mayor sea su valor, mayor será la efectividad del proceso de instrucción.

Una cuestión que debería advertirse es la posible contradicción que se produciría al seleccionar items a través de un estudio de la sensitividad al proceso de instrucción y la finalidad básica de un test referenciado en un criterio. Se supone que una prueba de este tipo, cuyos contenidos han sido

evaluados por expertos y en principio incluye contenidos que el individuo debería conocer puede verse reducida por la selección de item, descartándose de esta forma algunos items que podrían ser de importancia. De hecho la eliminación de algunos items a través del estudio de sensitividad a la instrucción no mejora la validez de contenido de la prueba.

De lo expuesto se deduce entonces que la finalidad de computar un índice de sensitividad a la instrucción es medir la efectividad del programa de instrucción y no la selección de items del test. Se trata entonces de un análisis de items que presupone la buena calidad de los items del test.

5.6.3. INDICES DE ACUERDO

Puede resultar de gran importancia comparar las respuestas a dos pares de items, puesto que si éste es el caso, los items serían *intercambiables*. Las respuestas a items dicotómicos se pueden organizar en tablas de doble entrada, con dos filas y dos columnas.

Si la pregunta de interés es si los items *miden lo mismo*, se puede usar el Coeficiente Phi o el estadístico chi-cuadrado, dados por las ecuaciones (5.10) y (5.11) respectivamente. Si no se encontrara significación para cualquiera de estos estadísticos, surgiría una seria duda si las especificaciones son apropiadas o existe un problema con la calidad técnica de alguno o ambos items.

El grado de concordancia entre las respuestas a un par de items, en una tabla como la siguiente: :

		Item 1		
		1	**0**	**Total**
Item 2	**1**	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	Total	*a+c*	*b+d*	*a+b+c+d*

Tabla 5.4 Resultados en dos items dicotómicos.

puede medirse sencillamente mediante la suma de las proporciones correspondientes a las celdas de la diagonal de la tabla que representan un acuerdo

$$\frac{a+d}{n} \qquad \textbf{(5.22)}$$

Por último si la cuestión a dilucidar es si las dificultades de los items son similares en la población de individuos, se sugiere utilizar:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \qquad \textbf{(5.23)}$$

que debe compararse con el percentil adecuado a un cierto valor α de la distribución chi-cuadrado con 1 grado de libertad.

CAPITULO 6

Estudio de un caso real

Durante los cinco primeros capítulos de este trabajo se revisaron los aspectos más relevantes de la Teoría Clásica de los Scores Verdaderos, conocida como *Teoría Débil de Tests*. La intención de este sexto capítulo es aplicar los principios expuestos en un caso concreto.

Para tal finalidad, se tomará en consideración una batería de tests que fue diseñada por el Ministerio de Educación de la Nación en oportunidad del Operativo Nacional de Evaluación durante el mes de Noviembre de 1.999 para aplicarse en los séptimos años de la Educación General Básica (E.G.B.) de las escuelas públicas de nuestro país.

En este caso particular, una Prueba de Matemática y una Prueba de Lengua, se aplicó en la Escuela *Docencia Tucumana* en la ciudad de Villa Mariano Moreno, provincia de Tucumán, en el mes de Junio de 2.000. Se tomaron ocho (de diez) divisiones de 7° año de la escuela, de las cuales cuatro corresponden al turno mañana y cuatro corresponden al turno tarde, donde la distribución de alumnos por turno y división es la que sigue:

	Turno Mañana					*Turno Tarde*				
Curso	**7°1°**	**7°3°**	**7°4°**	**7°5°**	**Total**	**7°1°**	**7°3°**	**7°4°**	**7°5°**	**Total**
Alumnos	31	27	27	23	108	20	20	18	20	78

Tabla 6.1 Distribución de alumnos por turno y división en el caso considerado.

En ambos turnos, el criterio de las autoridades del establecimiento en cuanto a la asignación de alumnos a los distintos cursos se basa en la edad del estudiante, correspondiendo los de divisiones más bajas a los de menor edad. De esta forma, dado que las dos primeras divisiones reúnen a los alumnos entre 11 y 12 años se optó por excluir las segundas divisiones por una limitación de los recursos económicos disponibles para llevar adelante el estudio, reteniéndose las primeras divisiones como representativas de este subgrupo de alumnos.

6.1. DESCRIPCION DE LAS PRUEBAS

Las pruebas consisten de 35 ejercicios de elección múltiple con cuatro opciones para las respuestas posibles, siendo exactamente una de ellas la opción correcta. Cada ejercicio posee un score de 1 ó 0 punto de acuerdo a que la opción elegida sea o no la correcta. No se preveen scores intermedios como medidas de la proximidad de la información que maneja el alumno a la respuesta correcta.

Las pruebas aplicadas pueden consultarse en el Apéndice 1.

6.1.1. LAS COMPETENCIAS EN MATEMATICA Y LENGUA AL TERMINO DE LA E.G.B. 2 (6° GRADO DE LA ESCOLARIDAD PRIMARIA)

En teoría, las pruebas se diseñan para permitir la medida en que cada alumno desarrolló las competencias previstas por la currícula en Matemática y Lengua.

Una *Competencia* se concibe como una capacidad compleja que posee diferentes grados de integración y se pone de manifiesto en una gran variedad de situaciones correspondientes a los diversos ámbitos de la vida humana, personal y social. Se expresan en un desempeño eficaz cognitivo – afectivo – valorativo – operativo.

- ***Competencias en Matemática***

En Matemática se toman en consideración cuatro competencias:

✓ RECONOCER:

Hace referencia a la identificación de un dato dentro de un conjunto dado de información y al establecimiento de la una relación de similitud o igualdad con datos pertenecientes a conocimientos previos.

✓ CONCEPTUALIZAR:

Refiere a los procesos de agrupar diferentes objetos, datos o hechos particulares, en conjuntos más amplios o generales. También remite al establecimiento de las relaciones entre diferentes conceptos, a un mapeo conceptual.

✓ RESOLVER PROBLEMAS:

Hace referencia a la aplicación de diferentes estrategias, recursos o métodos para intentar soluciones a diferentes situaciones problemáticas.

✓ APLICAR ALGORITMOS:

Se vincula con la aplicación de pasos secuenciales, fijos y rutinarios previamente establecidos.

- ***Competencias en Lengua***

En Lengua, dos son las competencias que se evalúan:

✓ COMPRENSION LECTORA:

Hace referencia a la capacidad de procesar la información que un texto ofrece y de construir su significado, poniendo en juego conocimientos y saberes previos.

La Comprensión Lectora incluye, al menos, los procesos de:

- Reconocer la información textual explícita.
- Reconocer la información textual implícita.
- Coherencia.
- Cohesión.
- Reconocer relaciones de causa – efecto.
- Comprensión de Vocabulario relacionado con el texto.

✓ NOCIONES Y REGLAS GRAMATICALES

Se refiere a la aplicación de reglas o normas que rigen el aspecto gramatical del idioma e incluye:

- Relaciones Morfológicas.
- Relaciones Semánticas.
- Normativa.

Debería advertirse que cada uno de los aspectos que comprenden las dos competencias en Lengua, podrían ser consideradas como competencias en sí mismas.

6.1.2. LONGITUD DE CADA TEST

Cada competencia se evalúa mediante un test específico que se integra por un determinado número de items. Las tablas 6.2. y 6.3. exhiben la longitud de cada test.

	MATEMATICA			
Competencia	*Reconocer*	*Conceptualizar*	*Resolver Problemas*	*Aplicar algoritmos*
Nº de ejercicios	6	10	15	4

Tabla 6.2. Longitud del test empleado para medir las competencias en Matemática.

	LENGUA	
Competencia	*Comprensión Lectora*	*Nociones y Reglas Gramaticales*
Nº de ejercicios	25	10

Tabla 6.3. Longitud del test empleado para medir las competencias en Lengua

6.1.3. ITEMS QUE COMPONEN CADA TEST

Las Tablas 6.4 y 6.5. indican la distribución de items en cada test.

	MATEMATICA			
Competencia	*Reconocer*	*Conceptualizar*	*Resolver Problemas*	*Aplicar algoritmos*
Items que lo integran	1-14-15-16-34-35	3-4-10-24-25-26-27-28-30-31	2-5-9-11-12-17-18-19-20-21-22-23-29-32-33	6-7-8-13

Tabla 6.4. Longitud del test empleado para medir las competencias en Matemática.

	LENGUA	
Competencia	*Comprensión Lectora*	*Nociones y Reglas Gramaticales*
Items que lo integran	1 a 25	26 a 35

Tabla 6.5. Longitud del test empleado para medir las competencias en Lengua

6.1.4. CONSIDERACIONES GENERALES

Un aspecto importante a tener presente es la cuestión acerca de si estos tests deberían considerarse como *referenciados en una norma* o *referenciados en un criterio.*

Si se acepta que estas pruebas están dirigidas a estimar una medida del grado en que los estudiantes aprenden los contenidos, procedimientos y otros aspectos previstos en los currículums diseñados para tal finalidad, sin interesarse por la construcción de una escala comparativa entre los alumnos, es evidente que debería asumirse que se trata de tests *referenciados en un criterio.*

Este es el enfoque básico que se mantendrá en el resto del capítulo, aunque de manera lateral, podrían ser interesantes algunas conclusiones *referenciadas en la norma* que se expondrán.

6.2. TEST SOBRE LA COMPETENCIA *RESOLUCION DE PROBLEMAS*

Año tras año parece confirmarse que la gran dificultad en los procesos de enseñanza – aprendizaje en el área de Matemática se relaciona en buena medida con el desarrollo de la Competencia *Resolución de Problemas* . Se trata de un aspecto clave en la calidad de la educación que se imparte puesto que se relaciona con la aplicación de diferentes estrategias, recursos o métodos para hallar soluciones a diferentes situaciones problemáticas, no sólo en el ámbito de la Matemática sino también en cualquier situación de la vida real.

Esta es una razón de peso para seleccionar el test que evalúa esta competencia en orden a ejercitar los conceptos y principios teóricos expuestos en los cinco capítulos previos.

6.2.1. ANALISIS DE CONFIABILIDAD DEL TEST

Dado que se trata de items dicotómicos los que integran el test que se analiza, podemos recurrir a la expresión (3.16) para computar el coeficiente KR_{20}, que no es otro que el valor del α de Cronbach para este tipo de items

Es posible solicitar a S.P.S.S. el cálculo de este valor (de hecho se lo requiere posteriormente) pero con el fin de ilustrar la técnica para su cómputo procedemos a obtener en primer lugar los valores de la Dificultad de cada item y la varianza del test completo. La salida de S.P.S.S. se exhibe en el Apéndice 2.1. y a partir de esta podemos calcular la suma de productos $p_i q_i$. Debe advertirse que los números de cada item en la tabla siguiente corresponden al número de orden de los items 2; 5; 9; 11; 12; 17; 18; 19; 20; 21; 22; 23; 29; 32 y 33 respectivamente.

Item	p_i	q_i	$p_i q_i$
1	0,5145	0,4855	0,2498
2	0,5723	0,4277	0,2448
3	0,8613	0,1387	0,1195
4	0,526	0,474	0,2493
5	0,4277	0,5723	0,2448
6	0,3468	0,6532	0,2265
7	0,7283	0,2717	0,1979
8	0,4624	0,5376	0,2486
9	0,4566	0,5434	0,2481
10	0,3064	0,6936	0,2125
11	0,5318	0,4682	0,2490
12	0,6994	0,3006	0,2102
13	0,3121	0,6879	0,2147
14	0,2197	0,7803	0,1714
15	0,2312	0,7688	0,1777
			3,2649

Ahora podemos calcular:

$$\mathbf{KR}_{20} = \frac{k}{k-1} \cdot \left(1 - \frac{\sum p_i q_i}{\sigma_X^2}\right) = \frac{15}{15-1} \cdot \left(1 - \frac{3,2649}{7,6937}\right) = 0,6167$$

que coincide, salvo por errores de redondeo, con el valor calculado por S.P.S.S. que se muestra en el Apéndice 2.2.

Es posible también recurrir al Método de Hoyt para estimar el valor del coeficiente de confiabilidad para lo cual necesitamos llevar a cabo un Análisis de la Varianza en dos direcciones: items y alumnos para estar en condiciones de aplicar la expresión dada en (3.18).

La salida de S.P.S.S. que se ofrece en el Anexo 2.3 provee al ANOVA requerido, a partir de lo cual podemos calcular:

$$r_{xx'} = 1 - \frac{MS(r)}{MS(p)} = 1 - \frac{0{,}1979}{0{,}5129} = 0{,}6141$$

que nuevamente coincide con los resultados precedentes. Como quedó explicitado en el Capítulo 3, este método permite estimar de manera sencilla el valor del coeficiente de confiabilidad a partir de los resultados de un Análisis de la Varianza en dos direcciones, en particular cuando no se dispone de un software que lo provea. Sin embargo no debe perderse de vista que a partir de los resultados del Análisis de la Varianza no es posible estimar otros parámetros del análisis de item.

6.2.3. ANALISIS DE VALIDEZ DEL TEST

- ***Validez de Contenido***

Como se explicitó precedentemente, este test se diseño en ocasión del Operativo Nacional de Evaluación, organizado por la Dirección Nacional de Evaluación, dependiente de la Secretaría de Programación y Evaluación Educativa del Ministerio de Cultura y Educación de la Nación.

Considerando que la prueba fue elaborada por un equipo técnico compuesto por especialistas en la materia, debería descontarse la validez de contenido del test.

No se dispone de información adicional sobre el desempeño de los alumnos en otras instancias u otras variables de interés, de manera que no es posible elaborar otros tipos de análisis acerca de la validez del test.

6.2.4. ANALISIS DE ITEMS

- ***Análisis de Items como test referenciado en un criterio***

Interesa ahora obtener una medida de la dificultad de cada item. Se ha explicado ya que la *dificultad del item i* se define como la proporción p_i de alumnos que responden adecuadamente ese item.

El Apéndice 2.5. muestra el Output de S.P.S.S. con estadísticos descriptivos para cada item del test, indicando los valores de los Indices de

Dificultad y sus desvíos estándares. También se muestran estadísticos para describir la escala completa, las medias y varianzas de los items y las correlaciones entre ellos.

La dificultad media del test es 0,48, mientras los índices recorren valores desde 0,22 a 0,86.

Se ha indicado en el capítulo anterior que para un test referenciado en un criterio, el análisis de la dificultad de cada item no se hace sobre el principio de discriminar los individuos en función del desempeño que muestran en la prueba. Más bien, si se sostiene el supuesto que no hay fallas en las especificaciones de los items, lo cual es muy probable en este caso puesto que se trata de planteo de situaciones problemáticas sencillas, este valor medio debería asumirse como una medida de la efectividad de la instrucción que reciben los estudiantes. Es claro que existe una fuerte evidencia que deberían revisarse los contenidos, metodologías y otros aspectos de la planificación, conducción y evaluación de los currículums vigentes relacionados con esta competencia.

* ***Análisis de Item como test referenciado en norma***

Si los puntajes del test fueran a utilizar para construir inferencias comparativas entre los desempeños de los estudiantes que completaron el test, podemos extender el análisis de item, ahora girando en torno a la idea de medir el grado de discriminación que se logra.

Como primera cuestión, el Apéndice 2.7. muestra la salida de S.P.S.S. en la que se indica la media y la varianza de la escala completa (test) en el supuesto que se eliminara cada uno de los items que componen la prueba.

Se exhibe además los valores de la correlación (corregida) entre cada uno de los items y el score total del test y el valor de la confiabilidad de la prueba, dada por el α de Cronbach si se decidiera eliminar cada item.

En particular, si el item 17 se eliminara de la prueba la confiabilidad de la prueba se incrementaría a 0,6305, mientras que la exclusión del item 23 provocaría un descenso en la confiabilidad hasta 0,5666. En general la eliminación de cualquiera de los items no provocaría una mejora sustancial en la eficiencia de la escala completa.

Podemos también computar el Indice Discriminante de cada item de esta prueba. Como se recordará esta cantidad se define como la diferencia entre dos proporciones: los que han respondido acertadamente y pertenecen al grupo de mayor rendimiento en toda la prueba y quienes han contestado correctamente pero componen el grupo cuyos scores totales son bajos respecto del resto.

En cuanto a la forma de la distribución del score total del test, el gráfico de línea y el histograma en el que se ha ajustado una curva normal, sugieren una forma de campana. Estos se puede consultar en los Apéndices 3.1 y 3.2. Asociado a estos resultados, el Apéndice 2.6 presenta la salida de S.P.S.S. sobre la distribución de frecuencias de los scores totales del test y una prueba de normalidad de la cual se infiere que hay evidencia suficiente para no descartar el supuesto de normalidad de los datos.

Siguiendo la sugerencia de Kelley (1939) para obtener un índice discriminante más estable y con mayor poder discriminativo en el caso de una distribución de frecuencias aproximadamente normal, construimos los grupos extremos con el 27% de las observaciones en cada cola de la distribución. El gráfico de línea con porcentajes acumulados que aparece en el Apéndice 3.3. nos auxilia en esta tarea y se puede ver que el grupo de bajo rendimiento se integra por alumnos que hayan obtenido un score total menor o igual a 5 puntos, mientras que el grupo de alto rendimiento se constituye por alumnos cuyos puntajes son mayores o iguales a 9 puntos.

El paso siguiente es requerir de S.P.S.S., para cada item del test, una tabla de doble entrada en la que las filas localicen los dos valores posibles de cada item y las columnas representen los grupos de rendimiento. En cada celda se registrará la cantidad de casos y el porcentaje correspondiente a cada intersección de filas y columnas. La salida de S.P.S.S. con las tablas descriptas aparecen en el Apéndice 2.8 y un resumen de estos resultados se ofrece en la Tabla 6.6.

Además, simultáneamente se presentan los índices de dificultad y los valores de correlación con el score total para cada item de la prueba, de forma de facilitar una decisión sobre el item, en el caso que los resultados se utilizaran en la construcción de una versión definitiva del test.

Los items 2, 5, 11, 12, 18, 19, 20, 22 y 23 se aceptarían sin ningún tipo de restricciones, puesto que sus índices discriminantes son mayores a 0,40. Los items 17, 21, 29 y 33 presentan índices discriminantes entre 0,323 y 0,353 por lo que. en principio requerirían alguna revisión. En cuanto a los items 9 y 32, con índices discriminantes de 0,254 y 0,205 deberían ser inspeccionados con mayor detalle.

Sin embargo, si se hace intervenir en el análisis el valor del Coeficiente de Correlación Biserial por Puntos, podemos pensar en que para el caso de una muestra de tamaño 173 como la que se tiene, el error estándar del coeficiente, dado por la ecuación (5.5) es igual a 0,0762 y tomando dos desvíos sobre cero, el valor crítico para decidir su significación es 0,1525. Como se ve, todos los items del test superan este valor y por lo tanto ameritan su inclusión en la prueba.

Item	% respuestas correctas Grupo de Rendimiento		Indice Discriminante	Dificultad	Correlación Item - total
	Bajo	Alto			
2*	24,4	80,7	56,3	0,51	0,46
5*	26,7	82,5	55,8	0,57	0,44
9	71,1	96,5	25,4	0,86	0,36
11*	31,1	71,9	40,8	0,53	0,37
12*	13,3	77,2	63,9	0,43	0,51
17	15,6	49,1	33,5	0,35	0,21
18*	44,4	87,7	43,3	0,73	0,44
19*	13,3	70,2	56,9	0,46	0,50
20*	20,0	70,2	50,2	0,46	0,44
21	13,3	45,6	32,3	0,31	0,28
22*	22,2	77,2	55,0	0,53	0,45
23*	28,9	91,2	62,3	0,70	0,56
29	15,6	50,9	35,3	0,31	0,29
32	11,1	31,6	20,5	0,22	0,27
33	06,7	40,4	33,7	0,23	0,34

Tabla 6.6. Indices Discriminante, Dificultad y Correlación con el total para cada item del test que mide la competencia "Resolver Problemas".

Por último también puede plantearse la cuestión del análisis de item a través de las curvas características. Debe recordarse que la curva característica de un item es un gráfico cartesiano en el que sobre el eje horizontal se localizan los scores totales observados del test (en realidad como estimadores de los verdaderos scores desconocidos) y sobre el eje vertical las proporciones de individuos que contestaron correctamente ese item.

Las curvas características de los quince items que integran la prueba que evalúa la competencia *Resolver Problemas* se muestran en los Apéndices 3.4. a 3.17. En general todos los items presentan una curva con pendiente positiva, requisito indispensable para decidir la utilidad de un item, aunque estas pendientes exhiben muy distintas inclinaciones.

Los items 2, 12, 18, 19, 20, 22 y 23 son razonables, confirmando las conjeturas hechas a partir del análisis de los indices discriminantes y correlaciones con los totales.

Sin embargo, algunos de ellos son de mayor utilidad para discriminar con scores de cortes bajos como el 18, cuyo score de corte es próximo a 4; el 23 con score de corte próximos a 5 puntos y el item 5 con score de corte próximo a 6 puntos, en tanto que el resto pueden son más eficientes para discriminar con cortes más elevados: el 19 cuyo score de corte igual a 7; el 2 y 20 con scores de corte de 8; el 12 con corte de 8,5; los items 21 y 22 que tienen scores de corte próximos a 10 puntos y finalmente el item 33 con un score de corte mayor a 11 puntos.

El item 5 muestra gran inestabilidad, al igual que el 11, al punto que no es posible determinar un único score de corte.

El item 9 aparenta *ser muy fácil*, discriminando adecuadamente para un score de corte de apenas 2 puntos.

Por último los items 17, 29 y 32 vuelven a mostrar que necesitan de una revisión en detalle, presentan gran inestabilidad, sin un score de corte definido, lo que resulta coherente con el cuestionamiento hecho en el análisis previo.

6.3. ANALISIS DISCRIMINANTE DE LA BATERIA DE TESTS APLICADA

6.3.1. FINALIDAD

A manera de cierre de este capítulo destinado a la aplicación del cuerpo teórico expuesto sobre la Teoría Clásica de Tests, probablemente sería adecuado evaluar la capacidad de las variables *Competencias adquiridas por los alumnos en matemática y lengua al término de la E.G.B. 2* (Educación General Básica, segundo nivel, equivalente al 6º grado de escolaridad del sistema educativo anterior), en la discriminación de los individuos clasificados según el rendimiento general que hayan alcanzado (alto, medio y bajo) .

Se busca entonces, establecer las diferencias y similitudes de los alumnos que comparten una misma categoría de clasificación en términos de la asimilación de competencias de las dos asignaturas más importantes: Matemática y Lengua, desarrolladas hasta esa instancia educativa.

6.3.2. LAS COMPETENCIAS

Mediante la técnica del Análisis Discriminante, descripta en detalle en el Capítulo 4, se posibilitará identificar aquellas variables (competencias) que mejor discriminan a los individuos de las diferentes categorías de rendimiento.

Para este fin se han de considerar como competencias (entre paréntesis el nombre de la variable que la identifica)

✓ En Matemática:

- Reconocer ("RECONOCE")
- Conceptualizar ("CONCEPTU")
- Resolver Problemas ("RESOLVER")
- Aplicar Algoritmos ("ALGORIT")

✓ En Lengua:

- Comprensión Lectora ("COMPLECT")
- Nociones y Reglas Gramaticales ("GRAMATIC")

6.3.3. LOS GRUPOS DE RENDIMIENTO

Como primera medida, resulta necesario definir un criterio de formación de los grupos a partir de los rendimientos alcanzados de 159 alumnos, para los que se dispone de los scores en los tests sobre las competencias en Matemática y Lengua.

Una primera idea que puede surgir es la de dividir la distribución empírica de los datos por sus terciles o cuartiles, de tal forma de tener igual cantidad de individuos en cada grupo. Sin embargo este punto de vista presenta algunas limitaciones importantes como criterio de construcción de grupos, ya que no necesariamente la clasificación que se logra distribuye en los mismos grupos a individuos con similares características, pues los cortes se producen simplemente cuando se alcanza la cantidad de datos correspondiente. Esto puede provocar que dos individuos con scores muy diferentes entre si, se incluyan finalmente en una misma categoría, cuando en realidad sus desempeños son bastante disímiles.

Un criterio alternativo, y quizás más razonable, es construir los grupos a partir de la observación de la distribución de frecuencias acumuladas (ojiva) de los datos. La idea consiste en hacer cortes en los extremos de los intervalos donde la función (siempre creciente) permanece aproximadamente constante (mesetas).

A partir de la distribución de frecuencias acumuladas de la Calificación Final obtenida por cada alumno en el conjunto de tests administrados, que puede consultarse en el Apéndice 3.18, podemos construir tres grupos bien diferenciados:

- el 25% inferior de los individuos con un score menor que 41 sobre 100.
- el 53% central obtuvo un score entre 41 y 59 sobre 100 puntos
- el 22% superior logro un score de 60 puntos sobre 100 posibles o superior.

Creamos entonces una variable de clasificación, *Nivel,* en la base de datos de S.P.S.S. que asume el valor 1 para el primer grupo (rendimiento bajo), 2 para el segundo (rendimiento medio) y 3 para el tercero (rendimiento alto). Luego podemos utilizar esta variable auxiliar como variable de agrupamiento en un análisis discriminante que nos permita decidir cuáles son las competencias que más o mejor distinguen a estos grupos de diferente rendimiento.

6.3.4. ANALISIS DE RESULTADOS

Las salida de S.P.S.S. se pueden apreciar en el Apéndice 2.11.

✓ El programa informa que se utilizarán 159 casos en el análisis y que 13 casos han sido excluidos mediante una variable de selección (es decir un 92% de los casos se incluyen en la muestra), el resto se utilizará para la validación posterior.

✓ S.P.S.S. muestra los valores de las medias y desvíos estándares de cada variable (competencias) en cada grupo y a continuación los resultados de un Análisis de Varianza univariado (test F univariado) para contrastar las hipótesis de que las medias de cada competencia es igual en todos los grupos: todos los valores son significativos, por lo que se rechazan estas hipótesis.

✓ Luego de presentar los criterios de entrada, salida, tolerancia, máximo número de iteraciones, el máximo número de funciones de discriminación a obtener e indicar que las probabilidades *a priori* se toman en función de los tamaños de los grupos, en el primer paso selecciona la variable "COMPLECT" que cumple con las exigencias de entrada.

✓ En los pasos siguientes, 2 al 4, se incorporan, en orden, las variables "RESOLVER", "GRAMATIC" y "CONCEPTU", sin proceder a eliminar ninguna de ellas ya que no cumplen con los criterios de salida.

✓ S.P.S.S. indica en el paso siguiente que ninguna otra variable cumple con las condiciones de entrada, por lo que quedan sin seleccionar las competencias "ALGOR" y "RECONOCE".

✓ A continuación se presenta un calendario de ingresos de variables, exhibiéndose el valor del estadístico Lambda de Wilks y su significación, como resumen del procedimiento anterior.

- ✓ Luego se presentan los coeficientes de las funciones discriminantes canónicas: pueden observarse los dos autovalores calculados, siendo el primero de ellos mucho más grande que el segundo, lo que nos refiere el hecho que la primera función de discriminación captura la mayor parte de la variación total (un 99,28%). En concordancia con esta situación los dos valores de las correlaciones canónicas son muy diferentes (0,8937 vs. 0,1676).

- ✓ Sólo la primera función de discriminación presenta un valor p-observado menor que 0,05, por lo que podemos concluir que ésta es la que aporta información de valor en la discriminación de estos tres grupos clasificados según rendimiento alcanzado.

- ✓ A continuación S.P.S.S. presenta los valores de los coeficientes de las funciones de discriminación canónica estandarizados.

  En este caso, las funciones de discriminación logradas son:

$$D_{i1} = 0{,}86334 \cdot COMPLECT_i + 0{,}61353 \cdot RESOLVER_i + 0{,}42901 \cdot GRAMATIC_i + \\ + 0{,}33152 \cdot CONCEPTU_i$$

$$D_{i2} = -0{,}48345 \cdot COMPLECT_i + 0{,}01649 \cdot RESOLVER_i + 0{,}32668 \cdot GRAMATIC_i + \\ + 0{,}78377 \cdot CONCEPTU_i$$

- ✓ Debe notarse que la competencia que mejor discrimina los alumnos según sus rendimientos es la *Comprensión Lectora* de Lengua, seguida por la competencia *Resolver Problemas* de Matemática. Estas dos competencias presentan los coeficientes más altos en la primera función de discriminación canónica estandarizada. Con menor importancia relativa aparecen luego *Nociones y Reglas Gramaticales* y *Conceptualizar*.

- ✓ El resultado confirma una vez más la importancia del desarrollo en los estudiantes de la capacidad para interpretar lo que se lee y la estrecha conexión con la aptitud para buscar estrategias de resolución ante situaciones problemáticas. El sentido íntegro y la consistencia del *contexto* no son aspectos menores puesto que permiten al alumno comprender el escenario en el que se encuentra o que se le transmite. Esta práctica de identificar lo importante y nuclear de un mensaje se extiende a otros aspectos y entonces se entiende la presencia de la competencia *Conceptualizar* de Matemática. Parece ser que el conocimiento de las reglas ortográficas debe considerarse un aspecto de gran poder discriminante entre alumnos que logran alcanzar ciertos niveles de destreza y los que no.

- ✓ Notamos además que no se han seleccionado el resto de las competencias. Sin embargo, esto no debe interpretarse en el sentido que se trata de cuestiones de menor cuantía. Probablemente la exclusión de estas competencias del subconjunto con poder discriminativo se deba al tamaño reducido de los tests que se aplicaron para su registro.

- ✓ Es oportuno advertir que, lateralmente, estas funciones de discriminación construidas pueden utilizarse en la predicción de la membresía de un alumno particular, para el que se conocen sus scores en estos tests, a un grupo de rendimiento, esto es la probabilidad de pertenecer a dicho subconjunto.

- ✓ Cuando se evalúan las funciones de discriminación en las medias de los grupos, es claro que la primera función logra una gran separación entre los tres subconjuntos de alumnos, mientras que la segunda no lo hace adecuadamente. Esto puede verse claramente en el Diagrama de Dispersión de ambas funciones de discriminación que se muestra en el Anexo 3.19, donde los centroides de los grupos están distanciados en el sentido horizontal (de la primera función de discriminación), pero en la dirección vertical.

- ✓ Finalmente las matrices de confusión para los casos seleccionados en la muestra y para los que no nos permiten apreciar la efectividad de estas funciones de discriminación: en el caso de los individuos seleccionados en la muestra, el porcentaje de casos clasificados correctamente asciende a 90,41% (se nota el peso de la distribución sobre la diagonal de la tabla que corresponde a clasificaciones correctas). Cuando las funciones de discriminación se aplican a los casos no seleccionados, se logra un más que aceptable porcentaje de 84,62%.

- ✓ A manera de conclusión, digamos que si no se hubiera llevado a cabo un análisis discriminante, y tuviésemos que clasificar un individuo al azar, lo asignaríamos al grupo 2, dado que es el grupo con mayor probabilidad a priori. De esta forma hubiéramos clasificado correctamente un 54,79% de los datos. Contando con el Análisis Discriminante la capacidad de clasificación mejora sensiblemente hasta alcanzar un valor de 84,62%, mejorando sensiblemente los resultados.