Programa Internacional de Evaluación de Alumnos (PISA)

PISA 2003

Manual de análisis de datos

Usuarios de SPSS®

OCDE

ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICOS

ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICOS

La OCDE es un foro singular donde los gobiernos de 30 democracias trabajan juntos para enfrentarse a los retos económicos, sociales y medioambientales de la globalización. La OCDE también se encuentra en primera línea de los esfuerzos para comprender y ayudar a los gobiernos en su respuesta a los nuevos desarrollos y preocupaciones, tales como la dirección empresarial, la economía de la información y los retos de una población que envejece. La Organización proporciona un espacio en el que los gobiernos pueden comparar sus experiencias políticas, buscar respuestas a los problemas comunes, identificar las buenas prácticas y trabajar para coordinar las políticas internas e internacionales.

Los países miembros de la OCDE son: Alemania, Australia, Austria, Bélgica, Canadá, Corea, Dinamarca, España, los Estados Unidos, Finlandia, Francia, Grecia, Hungría, Irlanda, Islandia, Italia, Japón, Luxemburgo, México, Noruega, Nueva Zelanda, los Países Bajos, Polonia, Portugal, el Reino Unido, la República Checa, la República Eslovaca, Suecia, Suiza y Turquía. La Comisión de las Comunidades Europeas participa en los trabajos de la OCDE.

OECD Publishing difunde ampliamente los resultados de las recopilaciones estadísticas y de las investigaciones sobre temas económicos, sociales y medioambientales de la Organización, así como las convenciones, directrices y estándares acordados por sus miembros.

Editado originalmente en inglés por la OCDE con el título: **PISA 2003 Data Analysys Manual: SPSS Users**

Traducción y maquetación de Mercedes Polledo Carreño
Revisión de Rosario Martínez Arias y Ramón Pajares Box

Ministerio de Educación y Ciencia
Calle San Fernando del Jarama, 14
28002 Madrid, España
www.ince.mec.es

Prefacio

Las encuestas del Programa Internacional de Evaluación de Alumnos (PISA) de la OCDE, que tienen lugar cada tres años, se han diseñado para recabar información acerca de los estudiantes de 15 años en los países participantes. PISA examina hasta qué punto los estudiantes están preparados para afrontar los retos del futuro, más que su dominio de currículos determinados. Los datos recogidos durante cada ciclo de PISA son una fuente de información valiosísima para investigadores, responsables políticos, educadores, padres y alumnos. Hoy en día se reconoce que el futuro bienestar económico y social de los países se relaciona estrechamente con los conocimientos y destrezas de sus poblaciones. La información proporcionada por PISA permite a los países establecer comparaciones internacionales y evaluar cómo están preparados para la vida sus jóvenes de quince años en un contexto más amplio, así como contrastar sus relativos puntos fuertes y débiles.

La base de datos de PISA 2003, en la que se centra este manual, contiene información acerca de más de un cuarto de millón de alumnos de 41 países. No sólo incluye información sobre su rendimiento en las cuatro áreas principales de evaluación (lectura, matemáticas, ciencias y solución de problemas), sino también las respuestas al cuestionario del alumno, que forman parte de la evaluación. También se incluyen datos de los directores de los centros.

El MANUAL DE ANÁLISIS DE DATOS DE PISA 2003 ha surgido a partir de los talleres analíticos celebrados en Sydney, Viena, París y Bratislava, que permitieron a los participantes familiarizarse con las distintas técnicas necesarias para analizar correctamente las complejas bases de datos. Permite a los analistas replicar con seguridad los procedimientos empleados para la producción de los informes iniciales de PISA 2003, *Learning for Tomorrow's World – First Results from PISA 2003* (OCDE, 2004*a*) (existe traducción española: *Informe PISA 2003, Aprender para el Mundo de Mañana*, Madrid, Santillana, 2005) y *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003* (OCDE, 2004*b*), así como emprender nuevos análisis con precisión en áreas de interés especial. Además de incluir las técnicas necesarias, el manual también incorpora una descripción detallada de las variables construidas a partir de los cuestionarios de los alumnos y de los centros. Esta información se publicó anteriormente en el *Manual for the PISA 2000 Database* (OCDE, 2002*a*).

El MANUAL DE ANÁLISIS DE DATOS DE PISA 2003 consta de cuatro partes. Las dos primeras secciones aportan una base teórica detallada e instrucciones para analizar los datos; la tercera sección detalla los códigos de programación (sintaxis y macros) necesarios para llevar a cabo los análisis; la cuarta sección [no incluida en esta traducción] contiene una descripción detallada de la base de datos.

PISA supone un esfuerzo de colaboración por parte de los países participantes, guiado por sus gobiernos sobre la base de intereses de política educativa compartidos. El Consejo de gobierno de PISA está formado por representantes de todos los países y decide sobre la evaluación y la presentación de los resultados de PISA.

Existen dos versiones de este manual: una para usuarios de SPSS® y otra para usuarios de SAS®. La OCDE reconoce la labor creativa de Christian Monseur en la preparación del texto para am-

bas versiones del manual, en colaboración con Sheila Krawchuk y Keith Rust, así como en la preparación del código de programación para la versión SAS® del manual. El código para la versión SPSS® del manual fue preparado por Wolfram Schulz y Eveline Gebhardt. El principal trabajo de corrección fue llevado a cabo en la secretaría de la OCDE por Miyako Ikeda, Sophie Vayssettes, John Cresswell, Claire Shewbridge y Kate Lancaster. Las evaluaciones de PISA y los datos sobre los que se basan los manuales fueron preparados por el Consorcio de PISA bajo la dirección de Raymond Adams.

ÍNDICE

Instrucciones

Preparación de los archivos de datos

Todos los archivos de datos (en formato de texto) y los archivos de sintaxis de SPSS® están disponibles en el sitio web de PISA (*www.pisa.oecd.org*).

Usuarios de SPSS®

Los archivos de datos de alumnos y de centros de PISA 2003 deben estar depositados en la carpeta `C:\PISA\Data2003` antes de ejecutar la sintaxis de los siguientes capítulos.

Nombre del archivo de datos de alumnos: `C:\PISA\Data2003\INT_stui_2003.sav`

Nombre del archivo de datos de centros: `C:\PISA\Data2003\INT_schi_2003.sav`

Las macros de SPSS® presentadas en el capítulo 15 deben guardarse en `C:\PISA\macros`.

Sintaxis y macros de SPSS®

Todas las sintaxis y macros que se emplean en este manual pueden copiarse del sitio web de PISA (*www.pisa.oecd.org*). Cada capítulo del manual contiene un conjunto completo de sintaxis, que deben recorrerse consecutivamente para que todas se ejecuten con corrección dentro del capítulo.

Redondeo de cifras

En las tablas y fórmulas, las cifras se han redondeado a un número conveniente de decimales, aunque los cálculos siempre se han realizado con el número total de decimales.

Abreviaciones de países empleadas en este manual

AUS	Australia	FRA	Francia	KOR	Corea	PRT	Portugal
AUT	Austria	GBR	Reino Unido	LIE	Liechtenstein	RUS	Federación de Rusia
BEL	Bélgica	GRC	Grecia	LUX	Luxemburgo	SVK	Eslovaquia
BRA	Brasil	HKG	Hong Kong-China	LVA	Letonia	SWE	Suecia
CAN	Canadá	HUN	Hungría	MAC	Macao-China	THA	Tailandia
CHE	Suiza	IDN	Indonesia	MEX	México	TUN	Túnez
CZE	R. Checa	IRL	Irlanda	NLD	Países Bajos	TUR	Turquía
DEU	Alemania	ISL	Islandia	NOR	Noruega	URY	Uruguay
DNK	Dinamarca	ITA	Italia	NZL	Nueva Zelanda	USA	Estados Unidos
ESP	España	JPN	Japón	POL	Polonia	YUG	Serbia
FIN	Finlandia						

Estatus socioeconómico

A lo largo de todo este manual, el estatus profesional más elevado de cualquiera de los padres (HISEI) figurará como estatus socioeconómico del alumno. Debe advertirse que el estatus profesional es tan sólo uno de los aspectos del estatus socioeconómico, que también puede incluir los estudios y el patrimonio. La base de datos de PISA 2003 también incluye una medida socioeconómica más amplia, llamada *índice de estatus económico, social y cultural* (ESCS), que procede del estatus profesional más elevado de los padres, el nivel de estudios más elevado y una estimación relacionada con los bienes del hogar.

Documentación adicional

Para obtener más información acerca de los resultados de PISA 2003, véanse los informes iniciales de PISA 2003: *Learning for Tomorrow's World– First Results from PISA 2003* (OCDE, 2004*a*) (existe una traducción española: *Informe PISA 2003 - Aprender para el Mundo de Mañana*, Madrid, Santillana, 2005) y *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003* (OCDE, 2004*b*). Para obtener más información sobre los instrumentos de evaluación y los métodos usados en PISA, véase *PISA 2003 Technical Report* (OCDE, 2005) y el sitio web de PISA (*www.pisa.oecd.org*).

Capítulo 1

El Programa Internacional de Evaluación de Alumnos de la OCDE

Perspectiva general de PISA

El Programa Internacional de Evaluación de Alumnos (PISA) de la OCDE es un esfuerzo realizado en colaboración, que implica a todos los países de la OCDE y a un número considerable de países asociados, para medir hasta qué punto los estudiantes de 15 años están preparados para afrontar los desafíos de las actuales sociedades del conocimiento. La evaluación está orientada al futuro: se centra en la capacidad de los jóvenes de utilizar sus conocimientos y destrezas para afrontar los retos de la vida real, más que en el dominio de currículos educativos concretos. Esta orientación refleja un cambio en los objetivos de los mismos currículos, que cada vez más se ocupan de la aplicación del conocimiento en lugar de limitarse a su adquisición. Se escoge a alumnos de 15 años porque en la mayoría de países de la OCDE se trata de la edad en que aquellos se acercan al final de su escolarización obligatoria.

PISA es el esfuerzo internacional más amplio y riguroso hasta la fecha para evaluar el rendimiento estudiantil y reunir datos acerca de los alumnos, así como acerca de los factores familiares e institucionales que pueden influir en el rendimiento. Las decisiones sobre el alcance y la naturaleza de la evaluación y la información de contexto que había de recogerse fueron tomadas por destacados expertos de los países participantes y dirigidas en conjunto por sus gobiernos sobre la base de intereses compartidos y dirigidos hacia la aplicación política. Se dedicaron notables esfuerzos y recursos a cubrir un amplio espacio cultural y lingüístico en los materiales de evaluación. Se aplicaron mecanismos estrictos de control de calidad en la traducción, el muestreo y la recogida de datos. Como consecuencia, los resultados de PISA presentan gran validez y fiabilidad; además, pueden mejorar significativamente la comprensión de los resultados educativos en un gran número de países del mundo.

PISA se basa en un modelo dinámico de formación permanente en el que los nuevos conocimientos y destrezas necesarios para adaptarse a un mundo cambiante se adquieren de forma continua a lo largo de la vida. PISA se centra en las destrezas que los quinceañeros necesitarán en el futuro y pretende evaluar su capacidad de ponerlas en práctica. PISA sí evalúa los conocimientos de los estudiantes, pero también su potencial para reflexionar acerca de sus conocimientos y experiencias y aplicar estos a situaciones del mundo real. Por ejemplo, si quisiera entender y evaluar las indicaciones científicas sobre seguridad de los alimentos, un adulto no sólo necesitaría poseer algunas nociones básicas sobre la composición de las sustancias nutritivas, sino que debería ser capaz de aplicar esa información. El término *literacy* (competencia) se utiliza para condensar este concepto más amplio de conocimientos y destrezas.

PISA es un estudio continuo en el que se recogen datos cada tres años. En la primera encuesta de PISA, llevada a cabo en el 2000 en 32 países, se utilizaron tareas escritas en los centros bajo condiciones de examen que seguían normas aplicadas sistemáticamente. Otros 11 países participaron en la misma encuesta a finales del 2001 o a comienzos del 2002. La segunda encuesta se realizó en el 2003 en 41 países. La tabla 1.1 proporciona la lista de países participantes en PISA 2000 y PISA 2003.

Tabla 1.1. Países participantes en PISA 2000 y PISA 2003

	PISA 2000	PISA 2003
Países de la OCDE	Alemania, Australia, Austria, Bélgica, Canadá, Corea, Dinamarca, España, Estados Unidos, Finlandia, Francia, Grecia, Holanda[a], Hungría, Irlanda, Islandia, Italia, Japón, Luxemburgo, México, Noruega, Nueva Zelanda, Polonia, Portugal, Reino Unido, República Checa, Suecia, Suiza.	Alemania, Australia, Austria, Bélgica, Canadá, Corea, Dinamarca, España, Estados Unidos. Finlandia, Francia, Grecia, Holanda, Hungría, Irlanda, Islandia, Italia, Japón, Luxemburgo, México, Noruega, Nueva Zelanda, Polonia, Reino Unido,[c] República Checa, República Eslovaca, Suecia, Suiza, Turquía.
Países asociados	Albania, Argentina, Brasil, Bulgaria, Chile, Hong Kong-China, Indonesia, Israel, Letonia, Liechtenstein, Macedonia, Perú, Rumanía, Federación Rusa, Tailandia.	Brasil, Hong Kong-China, Indonesia, Liechtenstein, Letonia, Macao-China, Federación Rusa, Tailandia, Túnez, Uruguay, Serbia.[d]

PISA evalúa principalmente la competencia en lectura, matemáticas y ciencias. Para cada recogida de datos, se escoge una de estas tres áreas de evaluación como la principal, mientras que las otras se consideran áreas menores. PISA 2000 hizo hincapié en la lectura, mientras que en PISA 2003 el área principal fue la competencia matemática. Alrededor de un 70% del tiempo del examen se dedica al área principal y las otras áreas se reparten el tiempo restante.

Tabla 1.2. Áreas de evaluación cubiertas por cada recogida de datos

	Área de evaluación principal	Áreas de evaluación menores
PISA 2000	Lectura	Matemáticas Ciencias
PISA 2003	Matemáticas	Lectura Ciencias Solución de problemas
PISA 2006	Ciencias	Matemáticas Lectura

[a] La tasa de respuestas es demasiado baja para garantizar la comparabilidad. Véase el apéndice 3 en *Literacy Skills for the World of Tomorrow – Further Results From PISA 2000* (OCDE, 2003*a*).

[b] La tasa de respuestas es demasiado baja para garantizar la comparabilidad. Véase el apéndice 3 en *Literacy Skills for the World of Tomorrow – Further Results From PISA 2000* (OCDE, 2003*a*).

[c] La tasa de respuestas es demasiado baja para garantizar la comparabilidad. Véase el apéndice 3 en *Learning for Tomorrow's World – First Results from PISA* 2003 (OCDE, 2004*a*).

[d] Para el país Serbia y Montenegro, en PISA 2003 no se encuentran disponibles los datos de Montenegro, que representa el 7,9% de la población nacional. Se usa el nombre *Serbia* como abreviación de la parte serbia de Serbia y Montenegro.

En 2009, el área de evaluación principal volverá a ser la lectura.

¿Por qué PISA es único?

PISA no es el primer estudio internacional comparativo del rendimiento de los alumnos. A lo largo de los últimos cuarenta años se han llevado a cabo otros estudios, desarrollados sobre todo por la *International Association for the Evaluation of Educational Achievement* (IEA, Asociación Internacional para la Evaluación del Rendimiento Educativo) y la *Education Testing Service's International Assessment of Educational Progress* (IAEP, Asociación Internacional para la Evaluación del Progreso Educativo del ETS, Servicio de Evaluación Educativa).

Estas encuestas se han centrado en resultados directamente relacionados con las materias del currículo escolar que son básicamente comunes en los países participantes. Los aspectos del currículo particulares de un país o de un pequeño número de países no han sido normalmente tenidas en cuenta en las evaluaciones, pese a la importancia que los países implicados atribuyan a esas partes del currículo.

Entre las características fundamentales de PISA se encuentran:

- su orientación hacia la política educativa, con métodos de diseño e informes determinados por la necesidad de los gobiernos de obtener lecciones para su aplicación política;
- su concepto innovador de la *competencia*, relacionado con la capacidad de los alumnos para aplicar conocimientos y destrezas en materias clave y para analizar, razonar y comunicarse con eficacia mientras plantean, resuelven e interpretan problemas en situaciones diversas;
- su relevancia para la formación a lo largo de la vida, ya que PISA no se limita a evaluar las competencias curriculares y transcurriculares de los estudiantes, sino que también les pide que aporten información sobre su propia motivación para aprender, su concepto acerca de sí mismos y sus estrategias de aprendizaje;
- su regularidad, que permitirá a los países controlar los progresos que realizan para cumplir objetivos educativos clave;
- su amplitud geográfica y su condición de proyecto en colaboración; los 47 países que han participado en una evaluación de PISA hasta el momento y los 13 países adicionales que se unirán a la evaluación de PISA 2006 representan un tercio de la población mundial y casi nueve décimas partes del producto interior bruto mundial[1];
- su selección, basada en la edad, de jóvenes que se acercan al final de la escolarización obligatoria, lo que permitirá a los países evaluar el rendimiento de sus sistemas educativos. Si bien la mayoría de jóvenes de los países de la OCDE continúan su educación más allá de los 15 años, esta edad suele estar cercana al final del período inicial de escolarización básica en el que todos los alumnos suelen seguir un currículo común. Es útil determinar, en esta etapa, hasta qué punto han adquirido conocimientos y destrezas que los ayuden en el futuro, lo que incluye los derroteros particulares de la educación posterior que quizá sigan.

Esta insistencia en plantear las pruebas en términos de dominio y de amplios conceptos es espe-

cialmente significativa si se considera la preocupación de las naciones por desarrollar capital humano, que la OCDE define como los conocimientos, destrezas, competencias y otros atributos que poseen los individuos y son relevantes para el bienestar personal, social y económico.

Las estimaciones de capital humano han tendido, en el mejor de los casos, a obtenerse mediante variables indirectas, como el nivel de estudios alcanzado. Cuando el interés en el capital humano se extiende para incluir atributos que permitan la completa participación social y democrática en la vida adulta y que preparen a las personas para formarse a lo largo de toda la vida, la insuficiencia de estas variables sustitutivas se muestra con mayor claridad.

Al realizar pruebas directamente sobre conocimientos y destrezas cerca del final de la educación básica, PISA examina el grado de preparación de los jóvenes para la vida adulta y, hasta cierto punto, la eficacia de los sistemas educativos. El propósito de PISA es evaluar los logros en relación con los objetivos subyacentes (como los define la sociedad) de los sistemas educativos, no en relación con la enseñanza y el aprendizaje de un *corpus* de conocimientos. Esta visión de los resultados educativos es necesaria si se pretende que los centros y sistemas educativos hagan hincapié en los retos de hoy en día.

PISA define las áreas evaluadas de este modo:

- *Competencia matemática:* capacidad de una persona para identificar y comprender el papel que las matemáticas desempeñan en el mundo, realizar razonamientos bien fundados y utilizar y relacionarse con las matemáticas de formas que satisfagan las necesidades de esa persona como ciudadano constructivo, comprometido y reflexivo.
- *Competencia lectora:* capacidad de una persona para comprender y utilizar textos escritos y reflexionar sobre ellos, con el propósito de alcanzar objetivos propios, desarrollar sus conocimientos y potencial y participar en la sociedad.
- *Competencia científica:* capacidad de utilizar el conocimiento científico, identificar preguntas y obtener conclusiones basadas en datos, con objeto de entender el mundo natural y los cambios que ha sufrido debido a la actividad humana y contribuir a tomar decisiones acerca de él.
- *Solución de problemas:* capacidad de una persona de usar procesos cognitivos para afrontar y resolver situaciones reales e interdisciplinarias en las que la solución no sea inmediatamente obvia, y donde las áreas de competencia o conocimiento que podrían aplicarse no pertenecen a un ámbito único de matemáticas, ciencia o lectura.

En las siguientes publicaciones de PISA puede encontrarse más información sobre las áreas de evaluación:

- *Measuring Student Knowledge and Skills – A New Framework for Assessment* (OCDE, 1999*a*);
- *Sample Tasks from the PISA 2000 Assessment – Reading, Mathematical and Scientific Literacy* (OCDE, 2002*b*);
- *Literacy Skills for the World of Tomorrow – Further Results from PISA 2000* (OCDE, 2003*a*);
- *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (OCDE, 2003*b*) (existe traducción española: *Marcos teóricos de PISA 2003:*

Conocimientos y destrezas en Matemáticas, Lectura, Ciencias y Solución de problemas, Madrid: Ministerio de Educación y Ciencia, Instituto Nacional de Evaluación y Calidad del Sistema Educativo, 2004; ver *www.ince.mec.es/pub*);

- *Learning for Tomorrow's World – First Results from PISA 2003* (OCDE, 2004*a*) (existe traducción española: *Informe PISA 2003, Aprender para el Mundo de Mañana*, Madrid, Santillana, 2005);
- *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies* (OCDE, 2004*b*).

Cómo se realiza la evaluación

La evaluación del rendimiento académico

Las evaluaciones PISA 2000 y 2003 consistieron en pruebas realizadas con lápiz y papel. El formato de las preguntas de la evaluación es variado. Algunas preguntas requieren que los alumnos seleccionen o propongan respuestas sencillas que pueden compararse directamente con una respuesta correcta única, como preguntas de elección múltiple o preguntas de respuesta construida cerrada. Otras piden más elaboración: los estudiantes tienen que desarrollar su propia respuesta, diseñada para medir conceptos más generales que los recogidos en estudios más tradicionales; estas preguntas permiten aceptar una gama más amplia de respuestas y un sistema de calificación más complejo que puede considerar respuestas parcialmente correctas.

En PISA, la competencia se evalúa a través de unidades encabezadas por un estímulo (por ejemplo, un texto, una tabla, un gráfico, una ilustración, etcétera), seguido de ciertas tareas asociadas con él. Esta es una característica importante: permite que las preguntas alcancen mayor profundidad que si cada una de ellas introdujera un contexto totalmente nuevo. Así se da tiempo al estudiante para asimilar material que puede utilizarse a continuación para evaluar diversos aspectos del rendimiento.

Se encuentran ejemplos de ítems de la evaluación PISA 2000 en *Sample Tasks from the PISA 2000 Assessment – Reading, Mathematical and Scientific Literacy* (OCDE, 2002*b*).

Se encuentran ejemplos de ítems de la evaluación PISA 2003 en *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (OCDE, 2003*b*) (existe traducción española: *Marcos teóricos de PISA 2003: Conocimientos y destrezas en Matemáticas, Lectura, Ciencias y Solución de problemas,* Madrid: Ministerio de Educación y Ciencia, Instituto Nacional de Evaluación y Calidad del Sistema Educativo, 2004; ver *www.ince.mec.es/pub*);

Los cuestionarios de contexto y su uso

Para obtener información contextual, PISA pide a los alumnos y a los directores de los centros participantes que respondan a unos cuestionarios acerca de su entorno que duran entre veinte y treinta minutos. Estos cuestionarios son decisivos para el análisis de los resultados, ya que proporcionan información sobre un abanico de características de los alumnos y centros.

Los cuestionarios buscan información sobre:

- los alumnos y su contexto familiar, incluido el capital económico, social y cultural de los alumnos y sus familias;
- aspectos de la vida de los alumnos, como su actitud ante el aprendizaje, sus hábitos y su vida en el centro y en el ambiente familiar;
- aspectos de los centros, como la calidad de los recursos humanos y materiales, la financiación pública y privada, los procesos de toma de decisiones y las prácticas de contratación de personal;
- el contexto de la enseñanza, como los tipos y estructuras de la enseñanza, el tamaño de las clases y el grado de implicación de los padres;
- estrategias de aprendizaje autorregulado, preferencias de motivación y orientaciones sobre objetivos, mecanismos cognitivos, estrategias de control de actividad, preferencias por distintos tipos de situaciones educativas, estilos de aprendizaje y destrezas sociales (estos aspectos formaban parte de una opción internacional en la evaluación PISA 2000, pero se incluyeron en el cuestionario obligatorio de los alumnos en PISA 2003);
- aspectos del aprendizaje y la enseñanza, incluidos la motivación de los estudiantes, su compromiso y su grado de confianza en relación con el área de evaluación principal evaluada, así como la influencia de las estrategias de aprendizaje sobre el rendimiento en esta área.

Tanto en PISA 2003 como en PISA 2000, se ofreció como opción internacional un cuestionario sobre tecnología de la información y la comunicación. Dicho cuestionario se centraba en: 1) la disponibilidad y el uso de tecnología de la información (TI), así como el lugar donde se utiliza más la TI y el tipo de uso que se hace de ella; 2) el grado de confianza con la TI y la actitud hacia ella, incluida la autoeficacia y la actitud frente a los ordenadores; y 3) el aprendizaje previo de TI, haciendo hincapié en dónde habían aprendido los alumnos a usar los ordenadores y la Internet.

En PISA 2003, también se ofreció como opción internacional un cuestionario de estudios futuros, que recogió datos sobre la formación educativa de los alumnos en tres áreas: 1) la formación que habían recibido en el pasado, lo que incluía la repetición de curso, las interrupciones de la escolarización, los cambios de colegio y los cambios de programas educativos; 2) la formación actual de los alumnos en aspectos relacionados con las matemáticas, centrándose en el tipo de clases de matemáticas y su actual grado de rendimiento; y 3) el futuro de los estudiantes y su profesión, incidiendo en las expectativas de nivel de formación y la profesión esperada a los 30 años de edad.

Los cuestionarios de PISA 2003 están disponibles en los apéndices 2 a 5 de este volumen [no incluidos en esta traducción], así como en el sitio web de PISA (*www.pisa.oecd.org*).

Varios índices relativos a los alumnos y a los centros fueron elaborados a partir de los datos de los cuestionarios. Estos índices combinan varias respuestas aportadas por los alumnos o directores para establecer un concepto más amplio que no puede observarse directamente. Por ejemplo, no se puede comprobar directamente el compromiso del estudiante con la lectura, pero sí es

posible hacer varias preguntas como «me gusta hablar de libros con otras personas» que reflejen el grado de compromiso con la lectura.

Puede encontrarse más información sobre la elaboración de estos índices y sus propiedades psicométricas en el apéndice 9, así como en *PISA 2003 Technical Report* (OCDE 2005).

Acerca de este manual

PISA puso en marcha complejos procedimientos metodológicos para garantizar la fiabilidad de las estimaciones poblacionales y sus correspondientes errores típicos. Más exactamente, PISA 2000 y PISA 2003 utilizaron valores plausibles para obtener las estimaciones poblacionales del rendimiento de la población y pesos replicados para el cálculo de los errores típicos respectivos.

Además de estas dos complejidades metodológicas, PISA recopila datos con regularidad, en un entorno definido y con procedimientos normalizados.

Este manual está diseñado para explicar estas metodologías complejas mediante ejemplos que utilizan los datos de PISA. El manual no detalla todos los aspectos de las metodologías; sin embargo, se describen para garantizar que los posibles usuarios de la base de datos de PISA puedan comprenderlas y utilizar los datos de modo apropiado.

El análisis de datos de PISA es un proceso que se ha simplificado al utilizar procedimientos de programación dentro de paquetes de *software* de estadística, como SAS® y SPSS®. Como consecuencia, este manual también contiene ejemplos de estos procedimientos. Es más, existen dos versiones del manual: una para usuarios de SAS® y otra para usuarios de SPSS®. Cada versión del manual consta de cuatro partes.

La primera parte, que abarca desde el capítulo 1 hasta el capítulo 5, es idéntica en ambas versiones del manual. Presenta conceptos y teorías utilizados en PISA. Estos capítulos son:

1. El Programa Internacional de Evaluación de Alumnos de la OCDE
2. Pesos de las muestras
3. Pesos replicados
4. El modelo de Rasch
5. Valores plausibles

La segunda parte, que comprende los capítulos 6 al 14, es distinta para los dos manuales. En cada una, se describe cómo analizar correctamente los datos de PISA y contienen la sintaxis necesaria: o SAS® o SPSS®. Estos capítulos son:

6. Cálculo de errores típicos
7. Análisis con valores plausibles
8. Uso de niveles de rendimento
9. Análisis con variables a nivel de centro
10. Error típico de una diferencia
11. Media de la OCDE y total de la OCDE
12. Tendencias
13. Análisis multinivel
14. Otras cuestiones estadísticas

La tercera parte también es diferente en cada manual: consta del capítulo 15, que presenta las macros, o bien de SAS® o bien de SPSS®, que facilitan el cálculo de las estimaciones y los errores típicos.

La cuarta parte es idéntica en ambas versiones del manual. Consiste en apéndices que describen los detalles de los archivos de datos de PISA 2003[2]. [Esta parte no se incluye en la presente traducción].

Mientras que los capítulos se organizan según el tipo de análisis, el manual avanza progresivamente sobre la base del conocimiento estadístico y de la sintaxis de SAS® o SPSS® ya presentada. Por tanto, se recomienda leer los capítulos en orden, comenzando por el capítulo 1.

También existen paquetes especializados de *software* que están configurados para tratar con muestras complejas y valores plausibles. Entre ellos, se encuentran WesVar®, de Westat Inc. (*www.westat.com/wesvar*), AM, de American Institutes for Research (*www.am.air.org*), y SUDAAN, del Research Triangle Institute (*www.rti.org/sudaan*).

Además, la OCDE ha desarrollado un sitio web interactivo que realiza automáticamente análisis estadísticos sencillos (sobre todo, cálculo de medias y porcentajes), haciendo uso de las metodologías de valores plausibles y de pesos replicados:

http://PISAweb.acer.edu.au/oecd_2003/oecd_PISA_data.html

Este sitio también contiene las bases de datos completas de PISA 2003 en formato ASCII.

---

[1] La población combinada de todos los países (excluido Taiwán) que han participado o participarán en las evaluaciones de PISA 2000, 2003 y 2006 equivale a un 32% de la población mundial en el 2002. El PIB de estos países equivale a un 87% del PIB mundial. Los datos sobre PIB y tamaños de población proceden de la base de datos Indicadores de Desarrollo Mundial, de las Naciones Unidas.

[2] La descripción de los archivos de datos de PISA 2000 se recoge en *Manual for the PISA 2000 Database* (OCDE, 2002*a*).

Capítulo 2

Los pesos muestrales

Introducción

Las encuestas nacionales o internacionales suelen recopilar datos a partir de una muestra. Es preferible trabajar con una muestra en vez de con la población entera, por varias razones.

En primer lugar, para un estudios censal, es necesario identificar a todos los miembros de la población. El proceso de identificación no presenta grandes dificultades para las poblaciones humanas en algunos países, donde pueden estar disponibles bases de datos nacionales con los nombres y direcciones de todos o casi todos los ciudadanos. Sin embargo, en otros países no es posible que el investigador pueda identificar a todos los miembros o unidades muestrales de la población objetivo, principalmente porque llevaría demasiado tiempo o también debido a la naturaleza de esa población objetivo.

En segundo lugar, incluso si todos los miembros de una población son fáciles de identificar, los investigadores también pueden seleccionar una muestra, ya que trabajar con la población total:

- podría exigir presupuestos muy elevados;
- llevaría mucho tiempo y no se cumplirían, por tanto, los plazos de publicación;
- no necesariamente ayuda a obtener información adicional o necesaria.

Una muestra puede seleccionarse de distintas formas, según las características de la población y las preguntas de la encuesta. Todos los diseños muestrales procuran evitar el sesgo en el procedimiento de selección y alcanzar la máxima precisión a la vista de los recursos disponibles. Sin embargo, pueden surgir sesgos en la selección:

- si el muestreo se realiza mediante un método no aleatorio, lo que en general significa que la selección se ve influida consciente o inconscientemente por preferencias humanas. No debería subestimarse la importancia de la aleatoriedad en el procedimiento de selección;
- si el marco de muestreo (lista, índice u otro registro de población) que sirve como base de la selección no cubre la población adecuadamente, por completo o con exactitud.

Los sesgos también pueden surgir si algunas partes de la población son imposibles de localizar o rehúsan cooperar. En las encuestas educativas, los centros quizá rechacen participar y, dentro de los centros participantes, algunos alumnos podrían negarse a ello o sencillamente estar ausentes el día de la evaluación. El tamaño del sesgo introducido por la falta de respuesta de centros o alumnos es proporcional a la correlación entre la disposición del centro o del estudiante a participar y las medidas de la encuesta. Por ejemplo, quizá sea más probable que los alumnos con bajo rendimiento estén ausentes el día de la evaluación que aquellos cuyo rendimiento es alto. Por esta razón, las encuestas internacionales sobre educación exigen un índice mínimo de participación de alumnos. Para PISA, este mínimo es el 80%.

Por último, si las unidades muestrales no tienen las mismas oportunidades de ser seleccionadas y si los parámetros poblacionales se calculan sin tener en cuenta estas probabilidades diferenciales, quizá los resultados también queden sesgados. Para compensar estas distintas posibilidades, es necesario ponderar o asignar pesos a los datos. La ponderación consiste en reconocer que algunas unidades de la muestra son más importantes que otras y deben contribuir más que las restantes al cálculo de cualquier estimación poblacional. Una unidad de muestra con proba-

bilidad de selección muy pequeña se considerará más importante que una unidad con gran probabilidad de selección. Por tanto, los pesos son inversamente proporcionales a la probabilidad de selección.

De todas formas, una muestra sólo es útil en la medida en que permite la estimación de algunas características de la población completa. Esto significa que los índices estadísticos calculados sobre la muestra —como una media, una desviación típica, una correlación, un coeficiente de regresión, etcétera— pueden generalizarse a la población. Esta generalización es más fiable si se han cumplido los requisitos del muestreo.

Según el diseño muestral, las probabilidades de selección y los procedimientos para calcular las ponderaciones serán diferentes. Estas diferencias se discuten en las próximas secciones.

Pesos para muestras aleatorias simples

Seleccionar miembros de una población mediante muestreo aleatorio simple es el procedimiento más directo. Existen diversos métodos de obtener una muestra así; por ejemplo:

- se numera a los *N* miembros de una población[1] y se seleccionan *n* de ellos mediante números aleatorios sin reemplazamiento;
- en un recipiente se meten *N* fichas numeradas, se mezclan bien y se seleccionan *n* de ellas al azar;
- los *N* miembros de la población se disponen en orden aleatorio y después se selecciona por saltos sistemáticos a cada miembro que se encuentra en una posición múltiplo de *N/n*;
- a cada miembro de la población *N* se le asigna un número aleatorio. Los números aleatorios se ordenan de menor a mayor o mayor a menor. Los primeros *n* números forman una muestra aleatoria.

La muestra aleatoria simple proporciona una probabilidad igual de selección a cada miembro de la población. Si se seleccionan *n* miembros a partir de una población de *N* miembros según un procedimiento aleatorio simple, la probabilidad de cada miembro *i* de formar parte de la muestra es igual a:

$$p_i = \frac{n}{N}.$$

Por ejemplo, si 40 alumnos se seleccionan al azar a partir de una población de 400 estudiantes, la probabilidad de cada alumno *i* de formar parte de la muestra es igual a:

$$p_i = \frac{n}{N} = \frac{40}{400} = 0{,}1.$$

Dicho de otro modo, cada estudiante tiene una probabilidad entre diez de ser seleccionado.

Como ya se ha dicho, los pesos suelen definirse como la inversa de la probabilidad de selección. En el caso de una muestra aleatoria simple, el peso será igual a:

$$w_i = \frac{1}{p_1} = \frac{N}{n}.$$

El peso de cada uno de los 40 estudiantes seleccionados a partir de una población de 400 estudiantes será, por tanto, igual a:

$$w_i = \frac{1}{p_i} = \frac{N}{n} = \frac{400}{40} = 10.$$

Esto significa que cada alumno de la muestra se representa a sí mismo y además a otros nueve alumnos. Puesto que cada unidad tiene la misma probabilidad de selección en una muestra aleatoria simple, el peso asociado a cada unidad seleccionada también será idéntico. Por tanto, la suma de los pesos de las unidades seleccionadas será igual al tamaño de la población, es decir, *N:*

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{N}{n} = N.$$

En el ejemplo,

$$\sum_{i=1}^{40} 10 = 400.$$

Además, puesto que todas las unidades seleccionadas para la muestra tienen el mismo peso, la estimación de cualquier parámetro poblacional no se verá afectada por los pesos. Por ejemplo, consideremos la media de alguna característica *X*. La media ponderada es equivalente a la suma del producto del peso por *X* dividido por la suma de los pesos.

$$\hat{\mu}_{(X)} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Puesto que w_i es una constante, la media ponderada y la media no ponderada serán iguales.

$$\hat{\mu}_{(X)} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 \sum_{i=1}^{n} x_i}{w_1 \sum_{i=1}^{n} 1} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Diseños de muestreo para encuestas educativas

El muestreo aleatorio simple se usa muy raras veces en encuestas educativas , porque:

- Es demasiado caro. Es más, según el tamaño de la población de los centros, es muy posible que los alumnos seleccionados estén escolarizados en muchos centros distintos. Esto exigiría la formación de un gran número de administradores de pruebas, el pago de gran cantidad de gastos de viaje, etcétera.

- No es práctico: habría que entrar en contacto con demasiados centros.
- Sería imposible relacionar, desde un punto de vista estadístico, las variables de los alumnos y las variables de centros, clases o profesores. Las encuestas de educación, por lo general, intentan comprender la variabilidad estadística de la medida de los resultados escolares según variables del centro o del aula. Con sólo uno o unos pocos alumnos por centro, esta relación estadística no tendría estabilidad.

Por tanto, las encuestas de educación suelen seleccionar una muestra de alumnos en dos etapas. En primer lugar, se selecciona una muestra de centros a partir de la lista completa de los centros que contengan la población estudiantil de interés. Luego se toma una muestra aleatoria simple de alumnos o de aulas dentro de los centros seleccionados. En PISA, normalmente se escogen al azar 35 alumnos de la población de alumnos de 15 años escolarizados en los centros seleccionados. Si en algún centro seleccionado se escolariza a menos de 35, se invita a participar a todos los alumnos de 15 años del centro.

Este procedimiento de muestreo en dos etapas influirá sobre el cálculo de los pesos y, de modo parecido, el procedimiento de selección de los centros afectará a las características y propiedades de la muestra de alumnos.

Supongamos que la población de 400 alumnos se distribuye entre diez centros, cada uno de ellos con 40 alumnos. Se seleccionan cuatro centros al azar y, dentro de cada centro, se seleccionan diez alumnos según un procedimiento similar. Cada centro *i* tiene una probabilidad de selección igual a:

$$p_{1\_i} = \frac{n_c}{N_c} = \frac{4}{10} = 0{,}4\,.$$

Dentro de los cuatro centros seleccionados, cada alumno *j* tiene una probabilidad de selección igual a:

$$p_{2\_ij} = \frac{n_i}{N_i} = \frac{10}{40} = 0{,}25\,.$$

donde N_i es el número de alumnos en el centro *i* y n_i, el número de alumnos pertenecientes a la muestra del centro *i*. Significa que, dentro de cada centro seleccionado, la probabilidad de que un alumno sea seleccionado es de uno entre cuatro.

La probabilidad final de selección para el alumno *j* que asiste al centro *i* es igual al producto de la probabilidad de selección del centro por la probabilidad de selección del alumno dentro del centro, esto es:

$$p_{ij} = p_{1\_i} p_{2\_ij} = \frac{n_c n_i}{N_c N_i}\,.$$

En el ejemplo, la probabilidad final del alumno es igual a:

$$p_{ij} = p_{1\_i} p_{2\_ij} = \frac{n_c n_i}{N_c N_i} = \frac{4 \cdot 10}{10 \cdot 40} = 0{,}4 \cdot 0{,}25 = 0{,}10\,.$$

El peso del centro w_{1_i}, el peso dentro del centro w_{2_ij} y el peso final del alumno w_{ij} son, respectivamente, iguales a:

$$w_{1\_i} = \frac{1}{p_{1\_i}} = \frac{1}{0{,}4} = 2{,}5;$$

$$w_{2\_ij} = \frac{1}{p_{2\_ij}} = \frac{1}{0{,}25} = 4;$$

$$w_{ij} = \frac{1}{p_{ij}} = \frac{1}{0{,}1} = 10.$$

La tabla 2.1 presenta la probabilidad de selección entre centros, dentro del centro y la probabilidad final de selección para los alumnos escogidos, así como la ponderación para estos distintos niveles. Se han seleccionado los centros 2, 5, 7 y 10.

Tabla 2.1. Probabilidad de selección entre centros, dentro del centro y final y pesos correspondientes para una muestra aleatoria simple en dos etapas, cuyas unidades de la primera etapa son centros con el mismo tamaño

Número del centro	Tamaño del centro N_i	Prob. del centro p_{1_i}	Peso del centro w_{1_i}	Prob. dentro del centro p_{2_ij}	Peso dentro del centro w_{2_ij}	Prob. final del alumno p_{ij}	Peso final del alumno w_{ij}	Suma de pesos finales $n_i w_{ij}$
1	40							
2	40	0,4	2,5	0,25	4	0,1	10	100
3	40							
4	40							
5	40	0,4	2,5	0,25	4	0,1	10	100
6	40							
7	40	0,4	2,5	0,25	4	0,1	10	100
8	40							
9	40							
10	40	0,4	2,5	0,25	4	0,1	10	100
Total			**10**					**400**

Como se muestra en la tabla 2.1, la suma de los pesos de centros se corresponde con el número de centros de la población, es decir, 10, y la suma de las ponderaciones finales se corresponde con el número de alumnos de la población, es decir, 400.

Por supuesto, en la práctica los centros tienen un tamaño diferente. Los centros suelen escolarizar a más alumnos en las áreas urbanas y menos en las rurales. Si se selecciona a los centros por medio de un muestreo aleatorio simple, la probabilidad del centro no cambia pero dentro de los

centros la probabilidad de selección del alumno variará de acuerdo con el tamaño del centro. En un centro pequeño esta probabilidad será grande mientras que en un centro grande la probabilidad será pequeña. La tabla 2.2 muestra un ejemplo de los resultados que se obtienen cuando los centros no tienen el mismo tamaño.

Tabla 2.2. Probabilidad de selección entre centros, dentro del centro y final y pesos correspondientes para una muestra aleatoria simple en dos etapas, cuyas unidades de la primera etapa son centros de tamaños desiguales

Número del centro	Tamaño del centro	Prob. del centro	Peso del centro	Prob. dentro del centro	Peso dentro del centro	Prob. final del alumno	Peso final del alumno	Suma de pesos finales
1	10							
2	15	0,4	2,5	0,66	1,5	0,27	3,75	37,5
3	20							
4	25							
5	30	0,4	2,5	0,33	3	0,13	7,5	75
6	35							
7	40	0,4	2,5	0,25	4	0,1	10	100
8	45							
9	80							
10	100	0,4	2,5	0,1	10	0,04	25	250
Total			**10**					**462,5**

Con una muestra aleatoria simple de centros de distinto tamaño, todos los centros tendrán la misma probabilidad de selección y, como antes, la suma de los pesos del centro será igual al número de centros de la población. Por desgracia, la suma de pesos finales de alumno no será forzosamente igual al número de alumnos de la población. Además, el peso final del alumno será diferente de un centro a otro, según el tamaño de cada centro. Esta variabilidad reducirá la fiabilidad de todas las estimaciones de parámetros de la población.

Las tablas 2.3 y 2.4 presentan las distintas probabilidades y pesos si se seleccionan los cuatro centros más pequeños o los cuatro más grandes. Como se muestra en estas dos tablas, las sumas de los pesos finales del alumno difieren sustancialmente del valor esperado de 400. La suma de pesos del centro, en cambio, será siempre igual al número de centros de la población.

Tabla 2.3. Probabilidad de selección entre centros, dentro del centro y final y pesos correspondientes para una muestra aleatoria simple de centros de tamaños desiguales (centros más pequeños)

Número del centro	Tamaño del centro	Prob. del centro	Peso del centro	Prob. dentro del centro	Peso dentro del centro	Prob. final del alumno	Peso final del alumno	Suma de pesos finales
1	10	0,4	2,5	1	1	0,4	4	40
2	15	0,4	2,5	0,66	1,5	0,27	3,75	37,5
3	20	0,4	2,5	0,5	2	0,2	5	50
4	25	0,4	2,5	0,4	2,5	0,16	6,25	62,5
Total			**10**					**190**

Tabla 2.4. Probabilidad de selección entre centros, dentro del centro y final y pesos correspondientes para una muestra aleatoria simple de centros de tamaños desiguales (centros más grandes)

Número del centro	Tamaño del centro	Prob. del centro	Peso del centro	Prob. dentro del centro	Peso dentro del centro	Prob. final del alumno	Peso final del alumno	Suma de pesos finales
7	40	0,4	2,5	0,250	4	0,10	10,00	100,0
8	45	0,4	2,5	0,222	4,5	0,08	11,25	112,5
9	80	0,4	2,5	0,125	8	0,05	20,00	200,0
10	100	0,4	2,5	0,100	10	0,04	25,00	250,0
Total			**10**					**662,5**

Las encuestas internacionales de educación, como PISA, se centran mucho más en la muestra de alumnos que en la muestra de centros. Muchos autores consideran incluso que tales estudios no obtienen una muestra de centros en sí. Se limitan a considerar la muestra de centros como un estadio operativo para obtener la muestra de alumnos. Por tanto, un diseño de muestreo que consista en una muestra aleatoria simple de centros es inadecuado, ya que calcularía un tamaño de población de alumnos demasiado pequeño o demasiado grande. También provocaría una variabilidad notable de pesos finales y, como consecuencia, aumentaría la varianza muestral.

Con objeto de evitar estas desventajas, los centros se seleccionan con probabilidades proporcionales a su tamaño (PPT). Los centros más grandes, por tanto, tendrán una probabilidad superior de selección que los centros más pequeños, pero los alumnos de los centros más grandes tendrán una probabilidad menor de ser seleccionados dentro del centro que los alumnos de los centros más pequeños. Con tales procedimientos, la probabilidad de que un centro sea seleccionado es igual a la razón del tamaño del centro multiplicada por el número de centros que entrarán en la muestra y dividida por el número total de alumnos de la población:

$$p_{1\_i} = \frac{N_i \cdot n_c}{N}.$$

Las fórmulas para calcular las probabilidades y pesos dentro del centro permanecen invariables. La probabilidad y el peso finales siguen siendo el producto de las probabilidades o pesos del centro y dentro del centro. Por ejemplo, la probabilidad del centro para el centro 9 es igual a:

$$p_{1\_9} = \frac{N_9 \cdot n_c}{N} = \frac{80 \cdot 4}{400} = \frac{4}{5} = 0{,}8\,.$$

La probabilidad de selección para un alumno dentro del centro 9 es igual a:

$$p_{2\_9j} = \frac{n_9}{N_9} = \frac{10}{80} = 0{,}125\,.$$

La probabilidad final es igual a:

$$p_{9j} = 0{,}8 \cdot 0{,}125 = 0{,}1\,.$$

Tabla 2.5. Probabilidad de selección entre centros, dentro del centro y final y pesos correspondientes para una muestra de centros con probabilidades proporcionales a su tamaño y de tamaños desiguales

Número del centro	Tamaño del centro	Prob. del centro	Peso del centro	Prob. dentro del centro	Peso dentro del centro	Prob. final del alumno	Peso final del alumno	Suma de pesos finales
1	10							
2	15							
3	20	0,2	5,00	0,500	2,0	0,1	10	100
4	25							
5	30							
6	35							
7	40	0,4	2,50	0,250	4,0	0,1	10	100
8	45							
9	80	0,8	1,25	0,125	8,0	0,1	10	100
10	100	1	1,00	0,100	10,0	0,1	10	100
Total	**400**		**9,75**					**400**

Como se muestra en la tabla 2.5, los pesos del centro y dentro del centro difieren entre los centros, pero los pesos finales del alumno no varían. Por tanto, los pesos no aumentarán la variabilidad del muestreo. Además, la suma de pesos finales se corresponde con el número total de alumnos de la población. Sin embargo, la suma de los pesos de centros difiere del valor esperado de 10, pero esto no supone un problema grave, ya que las encuestas educativas suelen interesarse principalmente por la muestra de alumnos.

Con una muestra de centros PTT (con probabilidades proporcionales a su tamaño) y un número igual de alumnos seleccionados en cada centro seleccionado, la suma de los pesos finales del alumno será siempre igual al número total de alumnos de la población (en esta etapa, no se

tiene en cuenta la ausencia de respuestas). Esto seguirá siendo así incluso si los centros más pequeños o más grandes resultan seleccionados. Sin embargo, la suma de los pesos del centro no será igual al número de centros en la población. Si se seleccionan los cuatro centros más pequeños, la suma de los pesos del centro será igual a 25,666. Si se seleccionan los cuatro centros más grandes, será igual a 6,97.

Con el propósito de mantener una diferencia mínima entre el número de centros de la población y la suma de los pesos del centro de la muestra, los centros se seleccionan según un procedimiento sistemático. Este consiste, primero, en seleccionar los centros según su tamaño. Se calcula un intervalo muestral como la razón entre el número total de alumnos de la población y el número de centros de la muestra; es decir,

$$Int = \frac{N}{n_c} = \frac{400}{4} = 100$$

Tabla 2.6. Selección de centros según un procedimiento PTT y sistemático

Número del centro	Tamaño del centro	Primer número de alumno	Último número de alumno	¿Forma parte de la muestra?
1	10	1	10	No
2	15	11	25	No
3	20	26	45	No
4	25	46	70	No
5	30	71	100	Sí
6	35	101	135	No
7	40	136	175	No
8	45	176	220	Sí
9	80	221	300	Sí
10	100	301	400	Sí

Se toma un número aleatorio a partir de una distribución uniforme [0,1]. Digamos 0,752. Este número aleatorio se multiplica entonces por el intervalo muestral, es decir, 0,752 × 100 = 75,2. El centro que contiene el primer número de alumno mayor de 75,2 resulta seleccionado. Luego, el intervalo muestral se añade al valor 75,2. El centro donde estudia el alumno que tiene el primer número de alumno superior a 175,2 será seleccionado. Este procedimiento sistemático se aplica hasta alcanzar el número de centros necesarios en la muestra. En el ejemplo, los cuatro números de selección serán los siguientes: 75,2; 175,2; 275,2 y 375,2.

Ordenar el marco muestral de centros según el tamaño de estos y utilizar después un procedimiento sistemático de selección evita obtener una muestra con sólo centros pequeños o (más probablemente) con sólo centros grandes. Esto, por tanto, reduce la varianza muestral de la

suma de los pesos del centro, que es una estimación del tamaño de la población de centros.

¿Por qué varían los pesos de PISA?

Como se ha demostrado en la sección anterior, un diseño muestral en dos etapas con una muestra PPT de centros debería garantizar que todos los alumnos tengan la misma probabilidad de selección y, por tanto, el mismo peso. Sin embargo, los datos de PISA aún deben ser sometidos a ponderación.

La tabla 2.7 muestra claramente que los pesos finales de PISA 2003 presentan cierta variabilidad. Esta variabilidad es bastante pequeña en países como Islandia, Luxemburgo y Túnez, pero parece ser mayor en países como Canadá, Italia y el Reino Unido.

La tabla 2.8 presenta las medias ponderadas y no ponderadas por país en la escala de matemáticas de PISA 2003. Las diferencias entre las medias ponderadas y no ponderadas son pequeñas para los países con pequeña variabilidad en los pesos, como Islandia, Luxemburgo y Túnez. Por el contrario, el efecto de los pesos sobre la media podría ser considerable en los países que presenten una gran variabilidad en los pesos. Por ejemplo, no usar los pesos sobreestimaría el rendimiento en matemáticas de los alumnos italianos alrededor de 30 puntos en la escala de matemáticas de PISA e infraestimaría el rendimiento medio de los alumnos canadienses en casi 11 puntos.

Tabla 2.7. Los percentiles 10º, 25º, 50º, 75º y 90º de los pesos finales de PISA

	Percentil 10	Percentil 25	Percentil 50	Percentil 75	Percentil 90
AUS	4,70	11,86	19,44	25,06	29,55
AUT	13,00	14,92	17,24	20,33	25,53
BEL	4,09	10,48	12,96	15,32	19,22
BRA	222,44	309,68	407,59	502,14	627,49
CAN	1,16	2,18	5,09	13,17	36,28
CHE	1,35	2,88	6,70	15,55	21,76
CZE	5,19	12,55	17,77	23,77	27,33
DEU	140,10	160,05	180,05	208,72	243,21
DNK	8,86	10,07	11,73	13,29	16,22
ESP	3,97	4,38	15,50	48,73	83,84
FIN	2,80	9,94	11,60	12,24	13,29
FRA	142,51	148,21	159,98	177,56	213,43
GBR	7,73	10,71	23,12	136,69	180,64
GRC	15,07	17,18	21,71	27,56	30,90
HKG	13,31	14,26	15,15	16,60	19,36
HUN	16,13	19,27	22,25	25,37	29,41
IDN	21,82	42,47	106,18	272,23	435,96
IRL	11,33	12,01	13,51	15,31	17,99
ISL	1,06	1,12	1,16	1,20	1,36
ITA	2,56	14,93	20,65	66,11	108,66
JPN	217,14	248,47	258,13	281,97	314,52
KOR	80,82	89,60	96,72	107,86	117,81
LIE	1,00	1,00	1,01	1,03	1,06
LUX	1,00	1,01	1,03	1,06	1,09
LVA	4,26	5,17	6,47	7,40	8,92
MAC	1,14	3,12	4,80	6,60	8,09
MEX	3,09	6,36	13,00	27,49	67,09
NLD	24,84	35,41	43,80	52,42	65,60
NOR	11,11	11,59	12,47	13,53	14,76
NZL	7,41	8,99	10,77	12,34	13,98
POL	103,73	110,45	118,72	130,28	144,73
PRT	13,90	16,33	18,70	22,66	28,82
RUS	172,98	245,92	326,11	426,26	596,07
SVK	4,39	6,98	8,64	11,02	16,79
SWE	17,95	19,54	22,03	24,47	28,81
THA	74,96	101,57	119,35	130,48	154,26
TUN	31,27	31,41	32,19	33,32	34,62
TUR	22,06	50,49	109,69	135,98	152,65
URY	1,81	2,79	4,43	8,06	11,66
USA	296,10	418,79	554,25	704,78	885,84
YUG	8,68	12,83	16,62	18,20	19,73

Tabla 2.8. Medias ponderada y no ponderada de los países en la escala de matemáticas de PISA 2003

	Media ponderada	Media no ponderada	Diferencia
AUS	524,27	522,33	1,94
AUT	505,61	511,86	−6,25
BEL	529,29	533,19	−3,90
BRA	356,02	360,41	−4,40
CAN	532,49	521,40	11,09
CHE	526,55	518,24	8,31
CZE	516,46	534,95	−18,50
DEU	502,99	508,41	−5,43
DNK	514,29	513,69	0,60
ESP	485,11	494,78	−9,67
FIN	544,29	542,81	1,48
FRA	510,80	514,73	−3,93
GBR	508,26	514,44	−6,18
GRC	444,91	440,88	4,04
HKG	550,38	555,86	−5,48
HUN	490,01	488,59	1,42
IDN	360,16	361,51	−1,35
IRL	502,84	504,68	−1,84
ISL	515,11	515,05	0,05
ITA	465,66	496,00	−30,34
JPN	534,14	533,51	0,62
KOR	542,23	540,60	1,62
LIE	535,80	536,46	−0,67
LUX	493,21	493,48	−0,27
LVA	483,37	486,17	−2,80
MAC	527,27	522,79	4,48
MEX	385,22	405,40	−20,18
NLD	537,82	542,12	−4,29
NOR	495,19	495,64	−0,46
NZL	523,49	525,62	−2,13
POL	490,24	489,00	1,24
PRT	466,02	465,23	0,79
RUS	468,41	472,44	−4,03
SVK	498,18	504,12	−5,94
SWE	509,05	507,95	1,09
THA	416,98	422,73	−5,75
TUN	358,73	359,34	−0,61
TUR	423,42	426,72	−3,30
URY	422,20	412,99	9,21
USA	482,88	481,47	1,41
YUG	436,87	436,36	0,51

A la variabilidad de los pesos contribuyen distintos factores:

- *Sobremuestreo o inframuestreo en algunos estratos de la población:* Normalmente, la población de centros se divide en subgrupos diferentes, llamados *estratos*. Por ejemplo, un país podría decidir, por razones de conveniencia, separar los centros urbanos de los rurales en la lista de centros. En la mayoría de los casos, el número de alumnos seleccionados en el estrato rural y en el urbano será proporcional a lo que estos dos estratos representen en la población total. Este proceso de estratificación garantiza, por ejemplo, que se seleccione un número predefinido de centros dentro de cada estrato. Sin embargo, debido al propósito de obtener información nacional, un país podría decidir ampliar la muestra, en alguna parte de la población de alumnos, a más alumnos de los que se hubieran seleccionado basándose en una asignación proporcional. Supongamos que el 90% de la población estudiantil de un país sigue un itinerario de formación académica y el 10% uno de formación profesional. Si se desea comparar el rendimiento de los alumnos según el itinerario de formación, será necesario muestrear más alumnos de formación profesional de los que se hubieran seleccionado basándose en una asignación proporcional.
- *Inexactitud o falta de actualización en el tamaño de los centros en el marco muestral:* Cuando los centros se seleccionan con probabilidad proporcional a su tamaño, debe incluirse una medida del tamaño en la lista de centros. En PISA, esta medida de tamaño es el número de alumnos de 15 años en cada centro de la población, pero no siempre están disponibles estadísticas nacionales que consignen la fecha de nacimiento. Por tanto, la medida del tamaño puede ser el número de alumnos en aquel curso que tenga mayor número de quinceañeros (curso modal) o el número total de alumnos del centro dividido por el número de cursos. Además, incluso si las estadísticas nacionales consignan la fecha de nacimiento, estos datos podrían estar defasados uno o dos años. Por tanto, las incoherencias entre el número de quinceañeros en el momento de la prueba y la medida del tamaño usada en el marco muestral generan cierta variabilidad en las ponderaciones finales. Supongamos que el centro 9 de la tabla 2.5 tiene 100 estudiantes de 15 años en el momento de la prueba. Cuando los centros se seleccionaron a partir de la lista de centros, la medida del tamaño estaba fijada en 80. El peso del centro estaba fijado en 1,25. El peso dentro del centro será igual a 100 entre 10, es decir, 10 en lugar de 8. Por lo tanto, el peso final será igual a 12,5 en lugar de la cifra esperada, 10.
- *Ajuste de pesos entre centros y dentro del centro debido a la falta de respuesta de centros y alumnos:* Algunos centros y, dentro de los centros seleccionados, algunos alumnos quizá se nieguen a participar. Para compensar esta falta de respuesta, se aplica un ajuste de peso en cada nivel en que se produce la falta de respuesta. Por ejemplo, si sólo 25 de los 35 alumnos seleccionados están presentes el día de la evaluación, entonces el peso de los alumnos participantes se multiplicará por una razón de 35 por 25.[2] El índice de participación de los alumnos variará de un centro a otro, por lo que los pesos finales serán diferentes. Un procedimiento similar se aplica también para compensar la falta de respuesta de un centro. Sobre estos factores de ajuste puede encontrarse más información en *PISA 2003 Technical Report* (Informe técnico de PISA 2003, OCDE 2005).

Conclusiones

Este capítulo ha descrito brevemente: *a*) qué es un peso y cómo calcularlo; *b*) cuál es el diseño muestral de PISA y por qué se considera el más adecuado; *c*) por qué los pesos de PISA muestran variabilidad; y *d*) la influencia de los pesos sobre las estimaciones poblacionales.

Todos los procedimientos o análisis estadísticos sobre los datos de PISA deben ponderarse. Los análisis sin ponderar proporcionarán estimaciones sesgadas de los parámetros de la población.

---

[1] *N* suele representar el tamaño de la población y *n*, el tamaño de la muestra.

[2] En PISA 2003, el ajuste de ponderación de alumnos debido a la falta de respuesta de los alumnos también podría diferir en un centro determinado.

Capítulo 3

Los pesos replicados

Introducción

En la mayoría de los casos, como se ha mencionado en el capítulo 2, las encuestas nacionales o internacionales recopilan datos a partir de una muestra en lugar de llevar a cabo un censo completo. Sin embargo, para una población determinada, existen miles, incluso millones de muestras posibles, y cada una de ellas no proporciona necesariamente las mismas estimaciones de estadísticos poblacionales. Toda generalización realizada a partir de una muestra, es decir, cada estimación de un estadístico poblacional, tiene asociada cierta incertidumbre o riesgo de error. La varianza muestral se corresponde con la medida de esta incertidumbre debida al muestreo.

Este capítulo explica los procedimientos estadísticos utilizados para calcular la varianza muestral y su raíz cuadrada, el error típico. Más concretamente, este capítulo expone cómo estimar las varianzas muestrales mediante pesos replicados para las estimaciones de población derivadas de un diseño de muestra complejo. En primer lugar, se examinará el concepto de varianza muestral mediante un ejemplo ficticio para el muestreo aleatorio simple. En segundo lugar, se investigará el cálculo del error típico para el muestreo en dos etapas. En tercer lugar, se introducirán los métodos de replicación para estimar las varianzas muestrales en muestras aleatorias simples y en muestras de dos etapas, respectivamente.

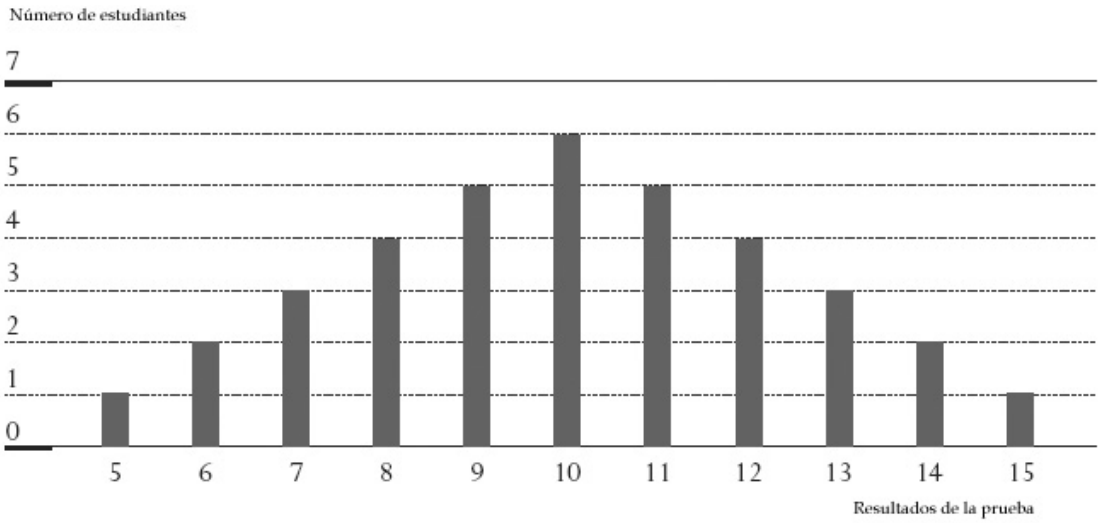
Varianza muestral para el muestreo aleatorio simple

Supongamos que un profesor decide aplicar el modelo *mastery learning* en su clase. Esta metodología exige que a cada lección le siga un examen. En el ejemplo dado, la clase del profesor tiene 36 alumnos. El profesor se da cuenta enseguida de que le llevaría demasiado tiempo corregir todos los exámenes, de modo que decide seleccionar una muestra de ellos para comprobar si el material impartido ha sido asimilado (Bloom, 1979).

Sin embargo, el muestreo aleatorio de algunos exámenes puede provocar que sólo se seleccionen aquellos con mejor o con peor resultado, lo que introduciría un error importante en la estimación del rendimiento medio de la clase. Estas situaciones son ejemplos extremos, pero seleccionar una muestra al azar siempre genera cierta incertidumbre.

En la misma situación de ejemplo, antes de proceder a seleccionar algunos exámenes, el profesor decide corregir todos los de la primera lección y analizar sus resultados. La figura 3.1 presenta la distribución de los resultados de los 36 alumnos. Un alumno obtiene una puntuación de 5, dos alumnos obtienen una puntuación de 6, y así sucesivamente.

Figura 3.1. Distribución de los resultados de los 36 alumnos



La distribución de las puntuaciones de los alumnos se corresponde con una distribución normal. La media de la población y la varianza de la población son, respectivamente, iguales a:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{(5+6+6+7+\cdots+14+14+15)}{36} = \frac{360}{36} = 10\,,$$

$$\sigma^2 = \frac{1}{N}\sum_{i\_1}^{N}(x_i-\mu)^2 = \frac{\left[(5-10)^2+(6-10)^2+\cdots+(14-10)^2+(15-10)^2\right]}{36} = \frac{240}{36} = 5{,}8333$$

La desviación típica es, por tanto, igual a:

$$\sigma = \sqrt{\sigma^2} = \sqrt{5{,}833} = 2{,}415\,.$$

Entonces, el profesor decide seleccionar aleatoriamente una muestra de dos alumnos después de la siguiente lección, para ahorrar tiempo de corrección. El número de muestras posibles de 2 alumnos en una población de 36 es igual a:

$$C_{36}^{2} = \frac{36!}{(36-2)!2!} = 630\,.$$

Existen 630 muestras posibles de 2 alumnos en una población de 36. La tabla 3.1 describe estas 630 muestras posibles. Por ejemplo, existen dos muestras posibles que proporcionan una estimación de la media de 5,5 para el rendimiento del alumno. Estas dos muestras son: *a*) el alumno con nota de 5 y el primer alumno con nota de 6; y *b*) el alumno con un 5 y el segundo alumno con un 6. Del mismo modo, existen dos formas de seleccionar una muestra que produzca una puntuación media de 6: *a*) los dos alumnos de la muestra obtienen una nota de 6; o *b*) un alumno recibe un 5 y el segundo recibe un 7. Como sólo dos alumnos obtuvieron una nota de 6 (figura 3.1), sólo existe una muestra posible con dos notas de 6. Puesto que la figura 3.1 muestra que

sólo hay un alumno que recibió nota de 5 y tres alumnos que recibieron nota de 7, existen tres muestras posibles de alumnos con nota de 5 y de 7.

Tabla 3.1. Descripción de las 630 muestras posibles de 2 alumnos seleccionados entre 36 según su media

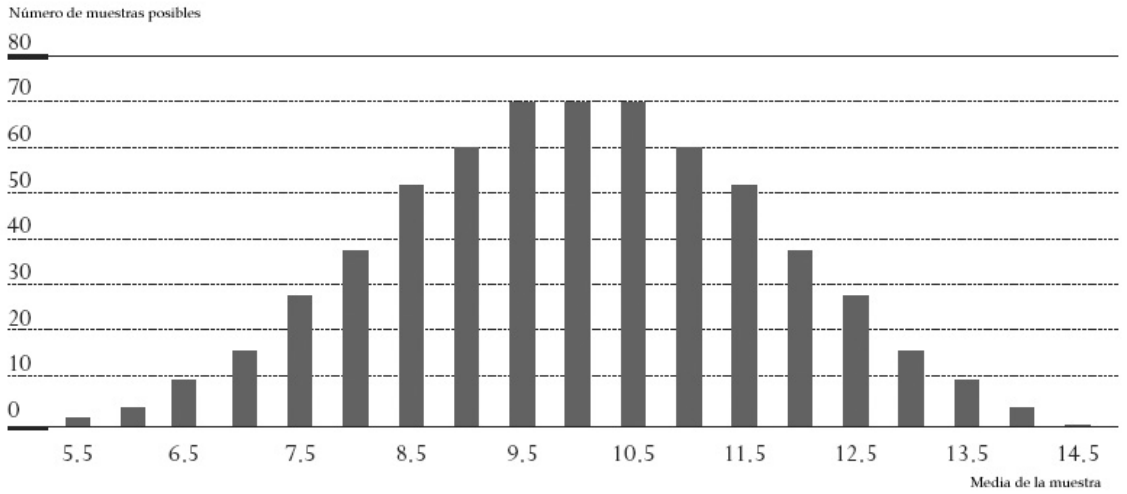
Media de la muestra	Resultados de los dos alumnos de la muestra	Número de combinaciones de los dos resultados	Número de muestras
5,5	5 y 6	2	2
6	6 y 6	1	4
	5 y 7	3	
6,5	5 y 8	4	10
	6 y 7	6	
7	7 y 7	3	16
	5 y 9	5	
	6 y 8	8	
7,5	5 y 10	6	28
	6 y 9	10	
	7 y 8	12	
8	8 y 8	6	38
	5 y 11	5	
	6 y 10	12	
	7 y 9	15	
8,5	5 y 12	4	52
	6 y 11	10	
	7 y 10	18	
	8 y 9	20	
9	9 y 9	10	60
	5 y 13	3	
	6 y 12	8	
	7 y 11	15	
	8 y 10	24	
9,5	5 y 14	2	70
	6 y 13	6	
	7 y 12	12	
	8 y 11	20	
	9 y 10	30	
10	10 y 10	15	70
	5 y 15	1	
	6 y 14	4	
	7 y 13	9	
	8 y 12	16	
	9 y 11	25	
10,5	6 y 15	2	70
	7 y 14	6	
	8 y 13	12	
	9 y 12	20	
	10 y 11	30	
11	7 y 15	3	60
	8 y 14	8	
	9 y 13	15	
	10 y 12	24	
	11 y 11	10	

11,5	8 y 15 9 y 14 10 y 13 11 y 12	4 10 18 20	52
12	9 y 15 10 y 14 11 y 13 12 y 12	5 12 15 6	38
12,5	10 y 15 11 y 14 12 y 13	6 10 12	28
13	11 y 15 12 y 14 13 y 13	5 8 2	16
13,5	12 y 15 13 y 14	4 6	10
14	13 y 15 14 y 14	3 1	4
14,5	14 y 15	2	2
			630

Como se observa en la tabla 3.1, existen dos muestras posibles con una media de 5,5, cuatro muestras posibles con una media de 6, diez muestras posibles con una media de 6,5, dieciséis muestras posibles con una media de 7, y así sucesivamente.

La figura 3.2 es un gráfico de la distribución de frecuencias de las estimaciones de las medias para todas las muestras posibles de 2 alumnos entre de 36.

Figura 3.2. Media de la distribución de medias



Como para todas las distribuciones, esta distribución de las medias de todas las muestras posibles puede resumirse mediante índices de tendencia central e índices de dispersión, como la media y la varianza (o su raíz cuadrada, la desviación típica).

$$\mu_{(\hat{\mu})} = \left[(2\times 5{,}5)+(4\times 6)+(10\times 6{,}5)+(16\times 7)+(28\times 7{,}5)+(38\times 8)+\ldots+(2\times 14{,}5)\right]/630 = 10$$

La media de todas las medias muestrales posibles es igual a la media de la población de alumnos, es decir, 10. Este resultado no es una coincidencia, sino una propiedad fundamental de la media de una muestra aleatoria simple; es decir, la media de las medias de todas las muestras posibles es igual a la media de la población. En lenguaje más formal, la media de la muestra es una estimación no sesgada de la media de la población. Dicho de otro modo, el valor esperado de la media de la muestra es igual a la media de la población.

Sin embargo, debería advertirse que existe una importante variación alrededor de este supuesto. En el ejemplo considerado, las medias de las muestras varían entre 5,5 y 14,5. La varianza de esta distribución, normalmente conocida como varianza muestral de la media, puede calcularse como:

$$\sigma^2_{(\hat{\mu})} = \left[(5{,}5-10)^2 + (5{,}5-10)^2 + (6-10)^2 + \cdots + (14{,}5-10)^2 + (14{,}5-10)^2\right] / 630 = 2{,}833$$

Su raíz cuadrada, conocida como error típico, es igual a:

$$\sigma_{(\hat{\mu})} = \sqrt{\sigma^2_{(\hat{\mu})}} = \sqrt{2{,}833} = 1{,}68$$

Sin embargo, ¿qué información aporta el error típico de la media o, más concretamente, qué nos dice el valor 1,68? La distribución de las medias de todas las muestras posibles sigue aproximadamente una distribución normal. Por tanto, basándose en las propiedades matemáticas de la distribución normal, puede decirse que:

- el 68,2% de las medias de todas las muestras posibles se encuentra entre –1 y +1 de error típico en torno a la media;
- el 95% de las medias de todas las muestras posibles se encuentra entre –2 y +2 de errores típicos.

Comprobemos las propiedades matemáticas de la distribución normal utilizando la distribución muestral de las medias y su varianza. Recordemos que la media de la distribución muestral de medias es igual a 10 y su desviación típica, conocida con el término *error típico,* es igual a 1,68.

¿Cuántas muestras tienen una media entre $\mu_{(\hat{\mu})} - \sigma_{(\hat{\mu})}$ y $\mu_{(\hat{\mu})} + \sigma_{(\hat{\mu})}$, es decir, entre (10 − 1,68) y (10 + 1,68) o entre 8,32 y 11,68?

Tabla 3.2. Distribución de todas las muestras posibles con una media entre 8,32 y 11,68

Media de muestra	Número de muestras	Porcentaje de muestras	Porcentaje acumulado de muestra
8,5	52	0,0825	0,0825
9	60	0,0952	0,1777
9,5	70	0,1111	0,2888
10	70	0,1111	0,4
10,5	70	0,1111	0,5111
11	60	0,0952	0,6063
11,5	52	0,0825	0,6888
	434		

La tabla 3.2 ilustra que hay 434 muestras tomadas de 630 con una media comprendida entre 8,32 y 11,68; estas representan el 68,8% de todas las muestras. También puede demostrarse que el porcentaje de muestras con medias entre $\mu_{(\hat{\mu})} - 2\sigma_{(\hat{\mu})}$ y $\mu_{(\hat{\mu})} + 2\sigma_{(\hat{\mu})}$, es decir, entre 6,64 y 13,36, es igual a 94,9.

Para estimar el error típico de la media, se ha calculado la media de todas las muestras posibles. En realidad, sin embargo, sólo se conoce la media de una muestra. Esto, como se demostrará, es suficiente para calcular una estimación de la varianza muestral. Por ello, es importante identificar los factores responsables de la varianza muestral a partir de la única muestra seleccionada.

El primer factor determinante es el tamaño de la muestra. Si el profesor, en nuestro ejemplo, decide seleccionar cuatro exámenes en lugar de dos, entonces la distribución de la media muestral abarcará desde 6 (los 4 resultados más bajos son 5, 6, 6 y 7) a 14 (los 4 resultados más altos son 13, 14, 14 y 15). Recordemos que la distribución muestral iba de 5,5 a 14,5 con muestras de dos unidades. Aumentar el tamaño de la muestra reduce la varianza de la distribución.

Existen 58.905 muestras posibles de 4 alumnos en una población de 36 alumnos. La tabla 3.3 ilustra la distribución de todas las muestras posibles de 4 alumnos para una población de 36.

Tabla 3.3. Distribución de la media de todas las muestras posibles de cuatro alumnos en una población de 36 alumnos

Media de muestra	Número de muestras posibles
6,00	3
6,25	10
6,50	33
6,75	74
7	159
7,25	292
7,50	510
7,75	804
8	1213
8,25	1700
8,50	2288
8,75	2896
9	3531
9,25	4082
9,50	4553
9,75	4830
10	4949
10,25	4830
10,50	4553
10,75	4082
11	3531
11,25	2896
11,50	2288
11,75	1700
12	1213
12,25	804
12,50	510
12,75	292
13	159
13,25	74
13,50	33
13,75	10
14	3

Puede demostrarse fácilmente que esta distribución tiene una media de 10 y una desviación típica, llamada *error típico,* de 1,155.

Esto prueba que el tamaño de la muestra no afecta al valor esperado de la media de muestra, aunque sí reduce la varianza de la distribución de las medias muestrales: cuanto mayor sea el tamaño de la muestra, tanto más baja será la varianza muestral de la media.

El segundo factor que contribuye a la varianza muestral es la varianza de la población misma. Por ejemplo, si los resultados son relativos a una puntuación total de 40 en vez de 20 (es decir, los resultados de los alumnos se multiplican por dos), entonces la media de los resultados de los alumnos será 20, la varianza será 23,333 (es decir, cuatro veces 5,8333) y la desviación típica será igual a 4,83 (es decir, dos veces 2,415).

Puede demostrarse que la varianza muestral a partir de una muestra de dos alumnos será igual

a 11,333, y que el error típico de la media será igual a 3,3665 (es decir, dos veces 1,68).

El error típico de la media es, por tanto, proporcional a la varianza de población. Basándose en estos ejemplos, puede establecerse que la varianza muestral de la media es igual a:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

y el error típico de la media muestral es igual a:

$$\sigma_{(\hat{\mu})} = \sqrt{\sigma^2_{(\hat{\mu})}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Donde:

σ^2 = varianza de la población;

σ = desviación típica de la población;

n = tamaño de la muestra;

N = tamaño de la población.

Esta fórmula puede comprobarse con el ejemplo:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{5{,}833}{2}\left(\frac{36-2}{36-1}\right) = 2{,}8333$$

Conforme aumenta el tamaño de la población, la razón $\left(\frac{N-n}{N-1}\right)$ tiende hacia 1. En tales casos, una aproximación cercana de la varianza muestral de la media se obtiene mediante:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n}$$

Sin embargo, en la práctica, la varianza de la población es desconocida y se estima a partir de una muestra. La estimación de la varianza muestral de la media, al igual que la estimación de la media, puede variar según la muestra. Por tanto, al estar basada en una muestra, sólo puede calcularse una estimación de la varianza muestral de la media (o de cualquier otra estimación).

En las restantes páginas de este manual, los conceptos de *varianza muestral* y *estimaciones de la varianza muestral* se unificarán, para simplificar el texto y las notaciones matemáticas. Es decir, los símbolos que muestren las estimaciones de la varianza muestral no tendrán un circunflejo (^) para diferenciarlos de los valores verdaderos, pero debe entenderse que, de hecho, son estimaciones.

Varianza muestral para el muestreo en dos etapas

Las encuestas educativas y, entre ellas, sobre todo las encuestas internacionales rara vez realizan muestras de alumnos a partir de una muestra aleatoria. Primero se seleccionan los centros y, dentro de cada centro seleccionado, se realiza un muestreo aleatorio de clases o de alumnos.

Una de las diferencias entre el muestreo aleatorio simple y el muestreo en dos etapas es que, para este último, los alumnos seleccionados escolarizados en el mismo centro no pueden considerarse como observaciones independientes. Esto es así porque los alumnos de un mismo centro tienen entre sí normalmente más características en común que con los alumnos de otras instituciones educativas. Por ejemplo, todos disponen de los mismos recursos del centro, quizá tengan los mismos profesores y estudian según un mismo plan de estudios, etcétera. Las diferencias entre los alumnos de centros distintos también son más grandes si en todos los centros no se imparten los mismos programas educativos. Por ejemplo, esperaríamos observar más diferencias entre los alumnos de un centro de formación profesional y los de un centro de formación académica que las que se observarían entre alumnos de dos centros de formación profesional.

Además, es bien conocido que, dentro de un país, dentro de sus regiones y dentro de sus ciudades, las personas tienden a vivir agrupadas en áreas determinadas según sus recursos económicos. Puesto que los alumnos suelen acudir a colegios cercanos a sus hogares, es probable que los alumnos que estudien en el mismo centro provengan de entornos sociales y económicos parecidos.

Por tanto, una muestra aleatoria simple de 4000 alumnos cubrirá probablemente la diversidad de la población mejor que una muestra de 100 centros, con 40 alumnos observados en cada centro. De esto se deduce que la incertidumbre asociada con cualquier estimación de parámetros poblacionales (es decir, el error típico) será superior para una muestra en dos etapas que para una muestra aleatoria simple del mismo tamaño.

El aumento de la incertidumbre debido al muestreo en dos etapas es directamente proporcional a las diferencias entre las unidades de la primera etapa, conocidas como unidades muestrales primarias, es decir, los centros escolares en las encuestas educativas. Las consecuencias de esta incertidumbre para dos situaciones extremas y ficticias se dan a continuación:

- Todos los alumnos de la población se asignan al azar a centros. Por tanto, no debería haber diferencias entre los centros. Seleccionar al azar 100 centros y, después, dentro de ellos, seleccionar al azar 40 alumnos sería similar, desde un punto de vista estadístico, a seleccionar directamente 4000 alumnos al azar, ya que no hay diferencias entre los centros. La incertidumbre asociada con toda estimación de parámetros poblacionales sería igual a la incertidumbre obtenida a partir de una muestra aleatoria simple de 4000 alumnos.
- Todos los centros son distintos pero, dentro de los centros, todos los alumnos son idénticos. en este caso, observar a un solo alumno o a 40 proporcionaría la misma cantidad de información. Por tanto, si se seleccionan 100 centros y se observan 40 alumnos por cada centro seleccionado, el tamaño muestral efectivo de esta muestra sería igual a 100. Por tanto, la incertidumbre asociada con cualquier estimación de parámetros poblacionales sería igual a la incertidumbre obtenida a partir de una muestra aleatoria simple de 100 alumnos.

Por supuesto, no existe ningún sistema educativo en el mundo que pueda identificarse con estas situaciones extremas y ficticias. Sin embargo, en algunos sistemas educativos (al menos por lo que se refiere a lo que la encuesta mide: por ejemplo, el rendimiento académico), las diferencias entre centros resultan ser muy pequeñas, mientras que en otros sistemas educativos, las diferencias pueden ser considerables.

El rendimiento académico de cada alumno puede representarse mediante una puntuación en una prueba o mediante la diferencia entre su puntuación y la puntuación media del país. En investigación educativa, es corriente dividir la diferencia entre la puntuación del alumno y la puntuación media del país en tres partes: 1) la distancia entre el rendimiento del alumno y la media de la clase correspondiente; 2) la distancia entre esta media de la clase y la media del centro correspondiente; y 3) la distancia entre esta media del centro y la media del país. La primera diferencia se relaciona con la varianza dentro de la clase (o la varianza residual en términos de análisis de la varianza). Indica hasta qué punto las notas de los alumnos pueden variar dentro de una clase determinada. La segunda diferencia (la distancia entre la media de la clase y la media del centro) se relaciona con la varianza entre clases dentro del mismo centro. Esta diferencia refleja la gama de diferencias entre clases dentro de los centros. Esta varianza entre clases dentro del mismo centro podría ser considerable en los centros docentes que ofrezcan a la vez formación profesional y formación académica. La tercera distancia (la diferencia entre la media del centro y la media del país) se llama varianza entre centros. Esta diferencia indica hasta qué punto varía el rendimiento de los alumnos entre unos centros y otros.

Para obtener una estimación de estos tres componentes de la varianza, sería necesario muestrear varios centros, al menos dos clases por centro y varios alumnos por clase. PISA selecciona aleatoriamente estudiantes de 15 años directamente a partir de listas de alumnos dentro de los centros participantes. Por tanto, hablando en general, es imposible distinguir las varianzas entre las clases y dentro de ellas. PISA sólo puede proporcionar estimaciones de las varianzas entre centros y dentro de los centros.

La tabla 3.4 proporciona las varianzas entre centros y dentro de los centros en la escala de matemáticas de PISA 2003. En los países del norte de Europa, las varianzas entre centros son muy pequeñas comparadas con sus estimaciones de varianzas dentro de los centros. En estos países, la varianza de los alumnos se produce principalmente en este último nivel. En cuanto al rendimiento escolar, por tanto, los centros en tales países no varían en gran medida. Sin embargo, en Austria, Bélgica, Alemania, Hungría y Turquía, por ejemplo, más del 50% de las diferencias de alumnos en el rendimiento corresponden al nivel de los centros. Esto significa que el rendimiento escolar difiere considerablemente entre centros. Así pues, la incertidumbre asociada con todo parámetro de población será superior para estos países al compararla con la incertidumbre para los países del norte de Europa, dado un tamaño de muestra comparable de centros y alumnos.

Como advirtió Kish (1987):

> Los métodos estándar para el análisis estadístico se han desarrollado a partir de supuestos de muestreo aleatorio simple. Suponer la independencia de los elementos (u observaciones) individuales facilita mucho las matemáticas usadas para las teorías de distribución de las fórmulas de estadística avanzada. [...] Sin embargo, la selección independiente de elementos rara vez se ejecuta en la práctica, porque gran parte de la investigación en realidad se realiza necesariamente con diseños muestrales complejos. Resulta económico seleccionar conglomerados que sean agrupaciones naturales de elementos, y éstos tienden a ser más bien homogéneos para la mayor parte de características. Los supuestos pueden fallar leve o gravemente; de ahí que el análisis estadístico estándar tienda a producir subestimaciones leves o graves en la amplitud de los intervalos de confianza obtenidos. Las sobreestimaciones son posibles, pero escasas y leves.

Kish estableció un conocimiento de referencia en el tema de la varianza muestral según el tipo de estimador y el diseño de muestra. Las distribuciones de la varianza muestral son bien conocidas para los estimadores univariantes y multivariantes en el caso de muestras aleatorias simples. El uso de variables de estratificación con una muestra aleatoria simple aún permite el cálculo matemático de las varianzas del muestreo, pero con un aumento considerable de complejidad. Como se muestra en la tabla 3.5, está disponible para algunos diseños el cálculo de varianzas muestrales para muestras en dos etapas, pero resultan bastante difíciles de calcular para índices multivariantes.

Tabla 3.4. Varianzas entre centros y dentro de los centros en la escala de matemáticas de PISA 2003[a]

	Varianza entre centros	Varianza dentro de los centros
AUS	1919,11	7169,09
AUT	5296,65	4299,71
BEL	7328,47	5738,33
BRA	4128,49	5173,60
CAN	1261,58	6250,12
CHE	3092,60	6198,65
CZE	4972,45	4557,50
DEU	6206,92	4498,70
DNK	1109,45	7357,14
ESP	1476,85	6081,74
FIN	336,24	6664,98
FRA	3822,62	4536,22
GBR	1881,09	6338,25
GRC	3387,52	5991,75
HKG	4675,30	5298,26
HUN	5688,56	4034,66
IDN	2769,48	3343,87
IRL	1246,70	6110,71
ISL	337,56	7849,99
ITA	4922,84	4426,67
JPN	5387,17	4668,82
KOR	3531,75	5011,56
LIE	3385,41	5154,08
LUX	2596,36	5806,97
LVA	1750,22	6156,52
MAC	1416,99	6449,96
MEX	2476,01	3916,46
NLD	552899	3326,09
NOR	599,49	7986,58
NZL	1740,61	7969,97
POL	1033,90	7151,46
PRT	2647,70	5151,93
RUS	2656,62	6021,44
SVK	3734,56	4873,69
SWE	986,03	8199,46
THA	2609,38	4387,08
TUN	2821,00	3825,36
TUR	6188,40	4891,13
URY	4457,08	5858,42
USA	2395,38	6731,45
YUG	2646,00	4661,59

[a] Los resultados se basan en el primer valor plausible de la escala de matemáticas, llamado PV1MATH en la base de datos de PISA 2003 (*www.pisa.oecd.org*).

Tabla 3.5. Estatus actual de los errores de muestreo

Métodos de selección	Medias y total de muestras completas	Medias y diferencias de subclases	Estadísticos analíticos complejos; por ejemplo, coeficientes de regresión
Selección aleatoria simple de elementos	Conocido	Conocido	Conocido
Selección estratificada de elementos	Conocido	Disponible	Conjeturado
Muestreo complejo por conglomerados	Conocido para algún diseño de muestras	Disponible	Difícil

Nota: la fila 1 se refiere a teoría estadística estándar (Kish y Frankel, 1974).

Los autores de manuales de muestreo suelen distinguir dos tipos de muestreo en dos etapas:

- muestreo en dos etapas con unidades de la primera etapa de mismo tamaño y
- muestreo en dos etapas con unidades de la primera etapa de tamaño desigual.

Más allá de esta distinción, para el cálculo de la varianza muestral es necesario tener en cuenta las diferentes características de la población y del diseño del muestreo, porque influyen en la varianza muestral. Algunos de los factores que deben ser considerados son:

- ¿es la población finita o infinita?;
- ¿era el tamaño un criterio determinante al seleccionar las unidades de la primera etapa?;
- ¿se utilizó un procedimiento sistemático para seleccionar las unidades de la primera etapa o de la segunda etapa?;
- ¿incluye el diseño muestral variables de estratificación?

El diseño muestral en dos etapas más simple se da con poblaciones de unidades infinitas en ambas etapas. Como las unidades de ambas etapas son poblaciones infinitas, se considera que las unidades muestrales primarias son de igual tamaño. Si se selecciona una muestra aleatoria simple de unidades muestrales primarias (UMP) y si, dentro de cada una de ellas, se selecciona una muestra aleatoria simple de unidades de la segunda etapa, entonces la varianza muestral de la media será igual a:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{entre\_UMP}}{n_{UMP}} + \frac{\sigma^2_{dentro\_de\_UMP}}{n_{UMP} n_{dentro}}$$

Apliquemos esta fórmula a una encuesta educativa y consideremos la población de los centros como infinita y la población de alumnos dentro de cada centro como infinita. El cálculo de la varianza muestral de la media será igual, por tanto, a:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{entre\_centro}}{n_{centro}} + \frac{\sigma^2_{dentro\_del\_centro}}{n_{alumnos}}$$

Tabla 3.6. Varianzas entre centros y dentro de los centros, y número de centros y alumnos participantes en Dinamarca y Alemania en PISA 2003

	Dinamarca	Alemania
Varianza entre centros	1109,45	6206,92
Varianza dentro de los centros	7357,14	4498,70
Número de centros participantes	206	216
Número de alumnos participantes	4218	4660

Según estos supuestos, en Dinamarca la varianza muestral de la media y su raíz cuadrada, es decir, el error típico, son iguales a:

$$\sigma^2_{(\hat{\mu})} = \frac{1109,45}{206} + \frac{7357,14}{4218} = 5,39 + 1,74 = 7,13$$

$$\sigma_{(\hat{\mu})} = \sqrt{7,13} = 2,67$$

En Alemania, la varianza muestral de la media y su raíz cuadrada, es decir, el error típico, son iguales a:

$$\sigma^2_{(\hat{\mu})} = \frac{6206,92}{216} + \frac{4498,79}{4660} = 28,74 + 0,97 = 29,71$$

$$\sigma_{(\hat{\mu})} = \sqrt{29,71} = 5,45$$

Si las dos muestras se considerasen como muestras aleatorias simples, el error típico de la media para Dinamarca y Alemania sería, respectivamente, igual a 1,42 y 1,51.

Basándonos en estos resultados, podemos realizar las siguientes observaciones:

- El error típico de la media es mayor para el muestreo en dos etapas que para el muestreo aleatorio simple. Por ejemplo, en el caso de Alemania, los errores típicos para el muestreo aleatorio simple y el muestreo en dos etapas son 1,51 y 5,45 respectivamente. Por tanto, considerar una muestra en dos etapas como una muestra aleatoria simple infraestimará considerablemente los errores típicos y, como consecuencia, los intervalos de confianza serán demasiado estrechos. El intervalo de confianza de la media de la escala de matemáticas, es decir, 503, sería igual a:
  [503 − (1,96 · 1,51) ; 503 + (1,96 · 1,51)] = [500,05 ; 505,96] en el caso de una muestra aleatoria simple, pero igual a [503 − (1,96 · 5,45) ; 503 + (1,96 · 5,45)] = [492,32 ; 513,68] en el caso de una muestra en dos etapas. Esto indica que cualquier valor medio estimado entre 492,32 y 500,05 y entre 505,96 y 513,68 puede ser tomado o no como estadísticamente distinto de la media alemana, según el error típico que se emplee.
- La varianza muestral de la media para muestras en dos etapas depende sobre todo de la varianza entre centros y del número de centros participantes. Así vemos que la varianza entre centros representa el 76% de la varianza muestral total en Dinamarca, es decir, 5,39/7,13 = 0,76. Para Alemania, la varianza entre centros supone un 97% de la varianza

muestral total (28,74/29,71 = 0,97). Por tanto, deberíamos esperar una varianza muestral mayor en países con mayor varianza entre centros, como Alemania y Austria, por ejemplo.

Sin embargo, la población de PISA no puede considerarse como una población infinita de centros con una población infinita de estudiantes. Además,

- los centros tienen tamaños desiguales;
- la muestra de PISA es una muestra sin reemplazamiento, es decir, un centro no puede ser seleccionado dos veces;
- los centros se seleccionan en proporción a su tamaño y según un procedimiento sistemático;
- en el diseño muestral se incluyen variables de estratificación.

Estas características del diseño del muestreo influyen en la varianza muestral, de modo que la fórmula utilizada más arriba también es inapropiada. Es más, el informe *Learning for Tomorrow's World – First Results from PISA 2003* (OCDE, 2004*a*) indica que los errores típicos de la media en la escala de matemáticas para Dinamarca y Alemania son 2,7 y 3,3, respectivamente.

Esto indica que el diseño muestral de PISA es bastante eficiente a la hora de reducir la varianza muestral. Sin embargo, el diseño se vuelve tan complejo que no hay una fórmula fácil para calcular la varianza muestral, o incluso los estimadores, como las medias.

Desde el estudio de competencia lectora de la IEA en 1990, se han utilizado métodos de replicación o remuestreo para calcular estimaciones de la varianza muestral en las encuestas internacionales sobre educación. Aún cuando estos métodos se conocían desde finales de la década de 1950, no se utilizaban a menudo ya que requieren numerosos cálculos. Con la disponibilidad de potentes ordenadores personales en la década de 1990 y el aumento del uso de bases de datos internacionales por parte de no-matemáticos, los centros internacionales de coordinación se decidieron a utilizar métodos de remuestreo para calcular las varianzas muestrales a partir de diseños de muestra complejos.

Según Rust y Rao (1996):

> El principio común que comparten estos métodos es la utilización de cálculos intensivos para superar las dificultades e incomodidades de utilizar una solución analítica para el problema. Dicho brevemente, el enfoque de replicación consiste en calcular la varianza de un parámetro de población de interés utilizando un gran número de submuestras algo distintas (o pesos muestrales algo distintos) para calcular el parámetro de interés. La variabilidad entre las estimaciones resultantes se utiliza para estimar el verdadero error muestral de la estimación inicial o de la muestra completa.

Estos métodos se describirán primero para las muestras aleatorias simples y para las muestras en dos etapas. Luego se presentará el método de replicación utilizado en PISA.

Métodos de replicación para las muestras aleatorias simples

Existen dos tipos principales de métodos de replicación para las muestras aleatorias simples, que se conocen como *jackknife* y *bootstrap*. Una de las diferencias más importantes entre el *jackknife* y el *bootstrap* se relaciona con el procedimiento empleado para producir las submuestras repetidas o muestras replicadas. A partir de una muestra de n unidades, el *jackknife* genera de forma sistemática n muestras replicadas de $n - 1$ unidades. El *bootstrap* genera aleatoriamente

un gran número de repeticiones de *n* unidades seleccionadas con reemplazamiento, donde cada unidad tiene más de una posibilidad de selección.

Puesto que PISA no emplea un método *bootstrap* de replicación adaptado a diseños muestrales polietápicos, esta sección sólo presentará el método *jackknife*.

Supongamos que se ha seleccionado una muestra de diez estudiantes mediante muestreo aleatorio simple. El método *jackknife* generará entonces diez submuestras o muestras replicadas, cada una de ellas de nueve estudiantes, como sigue:

Tabla 3.7. Muestras replicadas por el método *jackknife* y sus medias

Alumno	1	2	3	4	5	6	7	8	9	10	Media
Valor	10	11	12	13	14	15	16	17	18	19	14,50
Replicación 1	0	1	1	1	1	1	1	1	1	1	15,00
Replicación 2	1	0	1	1	1	1	1	1	1	1	14,88
Replicación 3	1	1	0	1	1	1	1	1	1	1	14,77
Replicación 4	1	1	1	0	1	1	1	1	1	1	14,66
Replicación 5	1	1	1	1	0	1	1	1	1	1	14,55
Replicación 6	1	1	1	1	1	0	1	1	1	1	14,44
Replicación 7	1	1	1	1	1	1	0	1	1	1	14,33
Replicación 8	1	1	1	1	1	1	1	0	1	1	14,22
Replicación 9	1	1	1	1	1	1	1	1	0	1	14,11
Replicación 10	1	1	1	1	1	1	1	1	1	0	14,00

Como se observa en la tabla 3.7, el método *jackknife* genera diez muestras replicadas de nueve alumnos. La media muestral basada en los diez estudiantes es igual a 14,5. Para la primera muestra replicada, el alumno 1 no se incluye en el cálculo de la media, y la media de los nueve alumnos incluidos en la muestra replicada 1 es 15,00. Para la segunda muestra replicada, el segundo alumno no está incluido y la media de los otros nueve es igual a 14,88, y así sucesivamente.

La estimación mediante *jackknife* de la varianza muestral de la media es igual a:

$$\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

donde $\hat{\theta}_{(i)}$ representa el estimador estadístico para la muestra replicada *i* y $\hat{\theta}$ representa el estimador estadístico basado en la muestra completa.

Basándose en los datos de la tabla 3.7, la varianza muestral de la media mediante *jackknife* es igual a:

$$\sigma^2_{(\hat{\mu})} = \frac{9}{10}\left[(15{,}00-14{,}50)^2 + (14{,}88-14{,}50)^2 + \cdots + (14{,}11-14{,}50)^2 + (14{,}00-14{,}50)^2\right]$$

$$\sigma^2_{(\hat{\mu})} = \frac{9}{10}(1{,}018519) = 0{,}9167$$

El estimador habitual de la varianza de la población es igual a:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{9}\left[(10-14{,}5)^2 + (11-14{,}5)^2 + \cdots + (18-14{,}5)^2 + (19-14{,}5)^2\right] = 9{,}17$$

Por tanto, la varianza muestral de la media, estimada mediante la fórmula matemática indicada, es igual a:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2}{n} = \frac{9{,}17}{10} = 0{,}917$$

Como se observa en este ejemplo, el método *jackknife* y la fórmula matemática proporcionan una estimación idéntica de la varianza muestral. Rust (1996) demuestra matemáticamente esta igualdad.

$$\hat{\mu}_{(i)} - \hat{\mu} = \frac{\left[\left(\sum_{i=1}^{n} x_i\right) - x_i\right]}{n-1} - \frac{\left[\sum_{i=1}^{n} x_i\right]}{n} = -\frac{x_i}{n-1} + \left[\sum_{i=1}^{n} x_i\right]\left[\frac{1}{n-1} - \frac{1}{n}\right] =$$

$$= -\frac{1}{(n-1)}\left[x_i - \left(\sum_{i=1}^{n} x_i\right)\left(1 - \frac{(n-1)}{n}\right)\right] = -\frac{1}{(n-1)}\left[x_i - \hat{\mu}(n-(n-1))\right] = -\frac{1}{(n-1)}(x_i - \hat{\mu})$$

Por tanto,

$$(\hat{\mu}_{(i)} - \hat{\mu})^2 = \frac{1}{(n-1)^2}(x_i - \hat{\mu})^2$$

$$\Rightarrow \sum_{i=1}^{n}(\hat{\mu}_{(i)} - \hat{\mu})^2 = \frac{1}{(n-1)^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{(n-1)}\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{(n-1)} = \frac{1}{(n-1)}\hat{\sigma}^2$$

$$\Rightarrow \sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\mu}_{(i)} - \hat{\mu})^2 = \frac{(n-1)}{n}\frac{1}{(n-1)}\hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

El método *jackknife* también puede utilizarse para calcular la varianza muestral de otros estadísticos, como los coeficientes de regresión. En este ejemplo concreto, el procedimiento consistirá en el cálculo de 11 coeficientes de regresión: uno basado en la muestra entera y otros diez, cada uno de ellos basado en la muestra replicada. La comparación entre el coeficiente de regresión de la muestra entera y cada uno de los diez coeficientes de regresión replicados proporcionará una estimación de la varianza muestral de ese estadístico.

Tabla 3.8. Valores de las variables X e Y para una muestra de 10 alumnos

Alumno	1	2	3	4	5	6	7	8	9	10
Valor Y	10	11	12	13	14	15	16	17	18	19
Valor X	10	13	14	19	11	12	16	17	18	15

El coeficiente de regresión para la muestra completa es igual a 0,53.

Tabla 3.9. Coeficientes de regresión para cada muestra replicada

	Coeficiente de regresión
Replicación 1	0,35
Replicación 2	0,55
Replicación 3	0,56
Replicación 4	0,64
Replicación 5	0,51
Replicación 6	0,55
Replicación 7	0,51
Replicación 8	0,48
Replicación 9	0,43
Replicación 10	0,68

La fórmula *jackknife*, es decir, $\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$, puede aplicarse para calcular la varianza muestral del coeficiente de regresión.

$$\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2 = \frac{9}{10}\left[(0{,}35 - 0{,}53)^2 + (0{,}55 - 0{,}53)^2 + \cdots + (0{,}68 - 0{,}53)^2\right] = 0{,}07$$

Este resultado es idéntico al que se obtendría mediante la fórmula habitual de varianza muestral para un coeficiente de regresión.

Métodos de remuestreo para muestras en dos etapas

Existen tres tipos de métodos de replicación para muestras en dos etapas:

- *jackknife*, con dos variantes: una para muestras sin estratificar y otra para muestras estratificadas;
- *balanced repeated replication* (BRR, «método de replicación repetido equilibrado») y su variante, la modificación de Fay;
- *bootstrap*.

PISA utiliza BRR con la modificación de Fay.[1]

El método *jackknife* para diseños muestrales en dos etapas sin estratificar

Si se selecciona una muestra aleatoria simple de unidades muestrales primarias sin usar ninguna variable de estratificación, puede demostrarse que la varianza muestral de la media obtenida mediante el método *jackknife* es matemáticamente igual a la fórmula descrita en la sección 2 de este capítulo, es decir:

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{entre\_UMP}}{n_{UMP}} + \frac{\sigma^2_{dentro\_de\_UMP}}{n_{UMP} n_{dentro}}$$

Consideremos una muestra de diez centros y, dentro de los centros seleccionados, una muestra aleatoria simple de alumnos. El método *jackknife* para una muestra en dos etapas sin estratificar consiste en generar diez replicaciones de nueve centros. Cada centro se retira una sola vez, de forma sistemática.

Tabla 3.10. Las muestras replicadas con *jackknife* para muestras en dos etapas sin estratificar

Replicación	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Centro 1	0,00	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11
Centro 2	1,11	0,00	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11
Centro 3	1,11	1,11	0,00	1,11	1,11	1,11	1,11	1,11	1,11	1,11
Centro 4	1,11	1,11	1,11	0,00	1,11	1,11	1,11	1,11	1,11	1,11
Centro 5	1,11	1,11	1,11	1,11	0,00	1,11	1,11	1,11	1,11	1,11
Centro 6	1,11	1,11	1,11	1,11	1,11	0,00	1,11	1,11	1,11	1,11
Centro 7	1,11	1,11	1,11	1,11	1,11	1,11	0,00	1,11	1,11	1,11
Centro 8	1,11	1,11	1,11	1,11	1,11	1,11	1,11	0,00	1,11	1,11
Centro 9	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11	0,00	1,11
Centro 10	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11	1,11	0,00

Para la primera replicación, llamada R1, se ha retirado el centro 1. Los pesos de los demás centros en la primera replicación se ajustan mediante un factor de 1,11, es decir, $\frac{10}{9}$ o, como regla general, un factor de $\frac{G}{G-1}$, donde *G* es el número de unidades muestrales primarias de la muestra. Este factor de ajuste se aplica luego cuando se combinan los pesos de los centros replicados y los pesos dentro de los centros replicados para obtener los pesos de alumnos replicados. Para la segunda replicación, se retira el centro 2 y los pesos de los centros restantes se ajustan según el mismo factor, y así sucesivamente.

El estadístico de interés se calcula para la muestra completa y, después, una vez más para cada replicación. Luego, se comparan las estimaciones replicadas con la estimación de la muestra completa para obtener la varianza muestral, como a continuación:

$$\sigma^2_{(\hat{\theta})} = \frac{(G-1)}{G} \sum_{i=1}^{G} \left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

Esta fórmula es idéntica a la utilizada para una muestra aleatoria simple, excepto que en lugar

de utilizar *n* replicaciones, donde *n* es el número de unidades de la muestra, esta fórmula usa *G* replicaciones, donde *G* es el número de unidades muestrales primarias.

El método *jackknife* para diseños muestrales en dos etapas estratificados

Como ya se mencionó al comienzo del capítulo 2, en todos los diseños muestrales subyacen dos principios importantes. El primero es el cuidado de evitar sesgos en el procedimiento de selección; el segundo, alcanzar la máxima precisión teniendo en cuenta los recursos financieros disponibles.

Para reducir la incertidumbre o para minimizar la varianza muestral sin modificar el tamaño de muestra, las encuestas de educación internacionales y nacionales suelen implementar los siguientes procedimientos en el diseño muestral:

- Las unidades muestrales primarias se seleccionan en proporción a su tamaño y según un procedimiento sistemático. Este procedimiento lleva a un procedimiento de muestreo de alumnos eficiente. Pueden seleccionarse en cada centro muestras de alumnos de igual tamaño. Al mismo tiempo, las probabilidades totales de selección (que combinan los componentes del muestreo de centros y el de alumnos) no varían gran cosa.
- Se estimula a los centros de investigación a que identifiquen variables de estratificación que estén asociadas estadísticamente con el rendimiento de los alumnos. Las características diferenciales tales como rural o urbano, académico o profesional, privado o público, resultan estar asociadas con el rendimiento de los alumnos. La reducción de la varianza muestral será proporcional al poder de explicación de estas variables de estratificación sobre el rendimiento estudiantil.

El método *jackknife* para las muestras estratificadas en dos etapas permite la reducción de la varianza muestral al tener en cuenta estos dos aspectos. No hacerlo así llevaría a una sobreestimación sistemática de las varianzas muestrales.

Supongamos que la lista de centros de la población se divide en dos partes, llamadas estratos: centros rurales y centros urbanos. Además, dentro de estos estratos, los centros se dividen según su tamaño. Dentro de cada estrato se seleccionan diez centros sistemáticamente y en proporción a su tamaño.

El método *jackknife* para los diseños muestrales en dos etapas estratificados consiste en emparejar sistemáticamente los centros que pertenecen a cada estrato en el orden en que se seleccionaron. Por tanto, los centros se emparejarán con otros centros similares.

Tabla 3.11. Las replicaciones con *jackknife* para diseños muestrales en dos etapas estratificados

Pseudoestrato	Centro	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1	2	1	1	1	1	1	1	1	1	1
1	2	0	1	1	1	1	1	1	1	1	1
2	3	1	0	1	1	1	1	1	1	1	1
2	4	1	2	1	1	1	1	1	1	1	1
3	5	1	1	2	1	1	1	1	1	1	1
3	6	1	1	0	1	1	1	1	1	1	1
4	7	1	1	1	0	1	1	1	1	1	1
4	8	1	1	1	2	1	1	1	1	1	1
5	9	1	1	1	1	2	1	1	1	1	1
5	10	1	1	1	1	0	1	1	1	1	1
6	11	1	1	1	1	1	0	1	1	1	1
6	12	1	1	1	1	1	2	1	1	1	1
7	13	1	1	1	1	1	1	2	1	1	1
7	14	1	1	1	1	1	1	0	1	1	1
8	15	1	1	1	1	1	1	1	0	1	1
8	16	1	1	1	1	1	1	1	2	1	1
9	17	1	1	1	1	1	1	1	1	2	1
9	18	1	1	1	1	1	1	1	1	0	1
10	19	1	1	1	1	1	1	1	1	1	0
10	20	1	1	1	1	1	1	1	1	1	2

La tabla 3.11 describe cómo se generan las replicaciones para este método. Los centros 1-10 son rurales y los centros 11-20, urbanos. Dentro de cada estrato, por tanto, existen cinco pares de centros o *pseudoestratos* (también llamados *estratos de varianza*).

El método *jackknife* para las muestras en dos etapas estratificadas generará tantas replicaciones como pares o pseudoestratos. En este ejemplo, se generarán por tanto diez replicaciones. Para cada muestra replicada, se retira al azar un centro de un pseudoestrato concreto y se dobla el peso del centro restante en el pseudoestrato. Para la replicación 1, llamada R1, se retira el centro 2 y se dobla el peso del centro 1 en el pseudoestrato 1. Para la replicación 2, se retira el centro 3 y se dobla el peso del centro 4 en el pseudoestrato, y así sucesivamente.

Como ya se mencionó antes, el estadístico de interés se calcula basándose en la muestra completa y, después, basándose de nuevo en cada muestra replicada. Las estimaciones replicadas se comparan entonces con la estimación de la muestra completa, para obtener la varianza muestral, así:

$$\sigma^2_{(\hat{\theta})} = \sum_{i=1}^{G} \left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

Este método de replicación se está utilizando actualmente en los estudios de la IEA.

El método BRR

Mientras que el método *jackknife* consiste en retirar sólo un centro para cada muestra replicada,

con el método BRR (*balanced repeated replication method,* «método de replicación repetido equilibrado») se selecciona al azar un centro dentro de cada pseudoestrato al que se le atribuye un peso de 0 y se doblan los pesos de los restantes centros.

Como este método proporciona un gran conjunto de replicaciones posibles, se genera un conjunto equilibrado de muestras replicadas según las matrices de Hadamard con objeto de evitar cálculos extensos. El número de replicaciones es el múltiplo más pequeño de cuatro, mayor o igual que el número de pseudoestratos. En este ejemplo, como hay 10 pseudoestratos, se generarán 12 replicaciones.

Tabla 3.12. Las replicaciones del método BRR

Pseudoestrato	Centro	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
1	1	2	0	0	2	0	0	0	2	2	2	0	2
1	2	0	2	2	0	2	2	2	0	0	0	2	0
2	3	2	2	0	0	2	0	0	0	2	2	2	0
2	4	0	0	2	2	0	2	2	2	0	0	0	2
3	5	2	0	2	0	0	2	0	0	0	2	2	2
3	6	0	2	0	2	2	0	2	2	2	0	0	0
4	7	2	2	0	2	0	0	2	0	0	0	2	2
4	8	0	0	2	0	2	2	0	2	2	2	0	0
5	9	2	2	2	0	2	0	0	2	0	0	0	2
5	10	0	0	0	2	0	2	2	0	2	2	2	0
6	11	2	2	2	2	0	2	0	0	2	0	0	0
6	12	0	0	0	0	2	0	2	2	0	2	2	2
7	13	2	0	2	2	2	0	2	0	0	2	0	0
7	14	0	2	0	0	0	2	0	2	2	0	2	2
8	15	2	0	0	2	2	2	0	2	0	0	2	0
8	16	0	2	2	0	0	0	2	0	2	2	0	2
9	17	2	0	0	0	2	2	2	0	2	0	0	2
9	18	0	2	2	2	0	0	0	2	0	2	2	0
10	19	2	2	0	0	0	2	2	2	0	2	0	0
10	20	0	0	2	2	2	0	0	0	2	0	2	2

De nuevo, el estadístico de interés se calcula basándose en la muestra completa y después otra vez por cada replicación. Luego, las estimaciones de las replicaciones se comparan con la estimación de la muestra completa para calcular la varianza muestral, como se hace a continuación:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

Con este método de replicación, cada muestra replicada sólo usa la mitad de las observaciones disponibles. Esta gran reducción de muestra, por tanto, podría ser problemática para estimar un estadístico en una subpoblación rara. Es más, el número de observaciones restantes podría ser tan pequeño, incluso igual a 0, que sea imposible estimar el parámetro poblacional para una muestra replicada en particular. Para superar esta desventaja, Fay desarrolló una variante del método BRR. En lugar de multiplicar los pesos de los centros por un factor de 0 o de 2, Fay sugirió multiplicar los pesos por un factor k de deflación entre 0 y 1, con un segundo factor de inflación igual a 2 menos k. Por ejemplo, si el factor de deflación del peso, llamado k, es igual a 0,6, el factor de inflación del peso será igual a $2 - k$, es decir, $1 - 0{,}6 = 1{,}4$ (Judkins, 1990).

PISA utiliza el método de Fay con un factor de 0,5. La tabla 3.13 describe cómo se generan las muestras y los pesos replicados mediante este método.

Tabla 3.13. Las replicaciones de Fay

Pseudoestrato	Centro	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
1	1	1,5	0,5	0,5	1,5	0,5	0,5	0,5	1,5	1,5	1,5	0,5	1,5
1	2	0,5	1,5	1,5	0,5	1,5	1,5	1,5	0,5	0,5	0,5	1,5	0,5
2	3	1,5	1,5	0,5	0,5	1,5	0,5	0,5	0,5	1,5	1,5	1,5	0,5
2	4	0,5	0,5	1,5	1,5	0,5	1,5	1,5	1,5	0,5	0,5	0,5	1,5
3	5	1,5	0,5	1,5	0,5	0,5	1,5	0,5	0,5	0,5	1,5	1,5	1,5
3	6	0,5	1,5	0,5	1,5	1,5	0,5	1,5	1,5	1,5	0,5	0,5	0,5
4	7	1,5	1,5	0,5	1,5	0,5	0,5	1,5	0,5	0,5	0,5	1,5	1,5
4	8	0,5	0,5	1,5	0,5	1,5	1,5	0,5	1,5	1,5	1,5	0,5	0,5
5	9	1,5	1,5	1,5	0,5	1,5	0,5	0,5	1,5	0,5	0,5	0,5	1,5
5	10	0,5	0,5	0,5	1,5	0,5	1,5	1,5	0,5	1,5	1,5	1,5	0,5
6	11	1,5	1,5	1,5	1,5	0,5	1,5	0,5	0,5	1,5	0,5	0,5	0,5
6	12	0,5	0,5	0,5	0,5	1,5	0,5	1,5	1,5	0,5	1,5	1,5	1,5
7	13	1,5	0,5	1,5	1,5	1,5	0,5	1,5	0,5	0,5	1,5	0,5	0,5
7	14	0,5	1,5	0,5	0,5	0,5	1,5	0,5	1,5	1,5	0,5	1,5	1,5
8	15	1,5	0,5	0,5	1,5	1,5	1,5	0,5	1,5	0,5	0,5	1,5	0,5
8	16	0,5	1,5	1,5	0,5	0,5	0,5	1,5	0,5	1,5	1,5	0,5	1,5
9	17	1,5	0,5	0,5	0,5	1,5	1,5	1,5	0,5	1,5	0,5	0,5	1,5
9	18	0,5	1,5	1,5	1,5	0,5	0,5	0,5	1,5	0,5	1,5	1,5	0,5
10	19	1,5	1,5	0,5	0,5	0,5	1,5	1,5	1,5	0,5	1,5	0,5	0,5
10	20	0,5	0,5	1,5	1,5	1,5	0,5	0,5	0,5	1,5	0,5	1,5	1,5

Como ocurre con todos los métodos de replicación, el estadístico de interés se calcula a partir de la muestra entera y, después, de nuevo a partir de cada replicación. Las estimaciones replicadas se comparan entonces con la estimación de la muestra entera para obtener la varianza muestral, como a continuación:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

En PISA, se decidió generar 80 muestras replicadas y, por tanto, 80 pesos replicados. De este modo, la fórmula se convierte en:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2 = \frac{1}{80(1-0{,}5)^2}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2 = \frac{1}{20}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

Otros procedimientos que tienen en cuenta el muestreo por conglomerados

Durante los dos últimos decenios, los modelos multinivel y los correspondientes paquetes de *software* para su estimación se han introducido en el campo de la investigación educativa. No hay duda de que estos modelos han permitido un gran avance a la hora de estudiar los fenómenos educativos. Efectivamente, los modelos de regresión multinivel ofrecen la posibilidad de

tener en cuenta el hecho de que los alumnos están anidados en clases y centros: cada factor puede evaluarse por separado cuando se establece la medida del resultado.

Los paquetes de *software* de regresión multinivel, como MLWin o HLM, del mismo modo que cualquier paquete profesional de estadística, proporcionan una estimación del error típico para cada uno de los parámetros poblacionales estimados. Mientras que SAS® y SPSS® consideran la muestra como una muestra aleatoria simple de elementos de la población, MLWin o HLM reconocen la estructura jerárquica de los datos, pero consideran que la muestra de centro es aleatoria simple. Por tanto, no tienen en cuenta la información complementaria para el diseño muestral que se utiliza en PISA con el fin de reducir la varianza muestral. Como consecuencia, en PISA las varianzas muestrales estimadas con modelos multinivel siempre serán mayores que las varianzas muestrales estimadas con muestras replicadas de Fay.

Puesto que estos paquetes de modelos multinivel no incorporan la información adicional del diseño muestral, sus estimaciones de error típico son similares a las del método *jackknife* para las muestras sin estratificar. Por ejemplo, los datos de PISA 2003 en Alemania se analizaron con el modelo multinivel propuesto por SAS® y llamado PROC MIXED. Los errores típicos de la media de los cinco valores plausibles para la escala combinada de competencia lectora fueron, respectivamente, 5,4565; 5,3900; 5,3911; 5,4692 y 5,3461. La media de estos cinco errores típicos es igual a 5,41. Recordemos que el uso de la fórmula en la sección 2 de este capítulo produce una estimación de la varianza muestral igual a 5,45.

Con los paquetes de *software* multinivel, es imposible evitar el uso de replicaciones si quieren obtenerse estimaciones no sesgadas de los errores típicos de los estimadores.

Conclusiones

Puesto que las encuestas internacionales sobre educación utilizan un diseño muestral en dos etapas la mayor parte de las veces, sería inapropiado aplicar las fórmulas de distribución muestral desarrolladas para el muestreo aleatorio simple. Hacerlo así llevaría a una subestimación de las varianzas muestrales.

Los diseños muestrales en las encuestas educativas pueden ser muy complejos. Como resultado, quizá no estén disponibles o sean demasiado complicadas las distribuciones muestrales incluso para estimadores simples como las medias. Desde el estudio sobre competencia lectora de la IEA en 1990, las varianzas muestrales se han estimado con ayuda de métodos de replicación. Estos métodos funcionan generando varias submuestras o muestras replicadas a partir de la muestra completa. A continuación, se estima el estadístico de interés para cada una de estas muestras replicadas y, después, se compara con la estimación de la muestra completa para proporcionar una estimación de la varianza muestral.

Una muestra replicada se forma simplemente mediante una transformación de los pesos muestrales totales según un algoritmo específico para el método de replicación. Por lo tanto, estos métodos pueden aplicarse a cualquier estimador[2] (medias, medianas, percentiles, correlaciones, coeficientes de regresión, etcétera), que puede calcularse con facilidad gracias a recursos informáticos avanzados. Además, usar estos pesos replicados no exige un extenso conocimiento de estadística, ya que los procedimientos pueden aplicarse sin que importe cuál sea el estadístico de interés.

[1] Véanse las razones para esta decisión en *PISA 2000 Technical Repor*t (OCDE, 2002*c*).

[2] Diversos estudios empíricos o teóricos han comparado los diferentes métodos de remuestreo para diseños de muestra complejos. Como advirtieron Rust y Krawchuk: «Una ventaja de los métodos BRR y BRR modificado sobre el *jackknife* es que aquellos tienen una base teórica sólida para usarlos con estadísticos no suavizados, por ejemplo cuantiles como la mediana. Desde hace mucho tiempo se sabe que el *jackknife* no es constante al estimar las varianzas de los cuantiles. Es decir, conforme el tamaño de muestra aumenta para un diseño de muestra dado, la estimación de las varianzas de los cuantiles no se vuelve necesariamente más precisa al usar el método *jackknife*». (Rust y Krawchuk, 2002)

Capítulo 4

El modelo de Rasch

Introducción

Las encuestas internacionales sobre educación, como PISA, están diseñadas para estimar el rendimiento en ciertas materias de varios subgrupos de alumnos, a cierta edad o en cierto curso.

Para que las encuestas se consideren válidas, es necesario desarrollar muchos ítems e incluirlos en los tests finales. Las publicaciones de la OCDE relacionadas con los marcos teóricos de las evaluaciones indican la amplitud y profundidad de las áreas de PISA y demuestran que hacen falta muchos ítems para evaluar un área con una definición tan amplia como, por ejemplo, *competencia matemática*.[1]

Al mismo tiempo, no es razonable ni deseable evaluar a cada alumno seleccionado con la batería completa de ítems, porque

- después de un tiempo de examen largo, la fatiga empieza a influir en los resultados de los alumnos, lo que sesgaría los resultados de la encuesta;
- los directores de los centros rehusarían dejar libres a sus alumnos para una duración de las pruebas tan largo como el que se necesitaría. Esto reduciría la tasa de participación de los centros, lo que, a su vez, podría sesgar considerablemente los resultados de las pruebas.

Para superar las exigencias opuestas de examinar a los alumnos durante un tiempo limitado y cubrir el dominio completo del área evaluada, se asigna a los alumnos un subconjunto de la batería de ítems y como resultado, sólo ciertas submuestras de alumnos responden a cada ítem.

Si el propósito de la encuesta fuera estimar el rendimiento detallando el porcentaje de respuestas correctas para cada ítem, no sería necesario informar sobre el rendimiento de cada alumno. Sin embargo, normalmente es necesario resumir información detallada a nivel de los ítems para comunicar los resultados de la encuesta a la comunidad investigadora, al público y, también, a los responsables de la política educativa. Además, las encuestas educativas pretenden explicar la diferencia de resultados entre países, centros y alumnos. Por ejemplo, un investigador podría estar interesado en la diferencia de rendimiento entre chicos y chicas.

¿Cómo puede resumirse la información?

Con relación a los países, el procedimiento más directo para resumir la información de los ítems sería calcular el porcentaje medio de respuestas correctas. Este sistema se ha usado con mucha frecuencia en encuestas anteriores, nacionales e internacionales, y todavía se usa en algunas encuestas internacionales actuales, incluso cuando se utilizan modelos más complejos. Estas encuestas pueden aportar el porcentaje total de respuestas correctas en matemáticas y ciencias, además de detallarlo según sub-áreas de contenido (por ejemplo, biología, física, química, ciencias de la Tierra, etcétera). Por ejemplo, en matemáticas, el porcentaje total de respuestas correctas para un país podría ser 54% y en otro, 65%.

La gran ventaja de este tipo de información es que todo el mundo puede entenderla. Cualquiera puede imaginar una prueba de matemáticas y visualizar qué representa un 54% y un 65% de respuestas correctas. Estas dos cifras también aportan un sentido a la diferencia de resultados entre los dos países.

Sin embargo, existen algunas debilidades en este enfoque, ya que el porcentaje de respuestas correctas depende de la dificultad de la prueba. El tamaño real de la diferencia de resultados entre dos países depende de esta dificultad, lo que puede llevar a interpretaciones equivocadas.

Las encuestas internacionales no pretenden sólo comunicar un nivel general de rendimiento. A lo largo de las últimas décadas, los responsables de políticas educativas se han interesado mucho por los indicadores de equidad. También pueden interesarse por el grado de dispersión de los resultados en su país. En algunos países, los resultados pueden estar agrupados alrededor de la media y en otros quizá haya gran número de alumnos con puntuaciones muy altas o muy bajas.

Sería imposible calcular índices de dispersión tan sólo con los índices de dificultad, basados en el porcentaje de respuestas correctas de todos los ítems. Para ello, la información recopilada en la prueba también debe resumirse a nivel del alumno.

Para comparar los resultados de dos alumnos evaluados con dos tests diferentes, éstos deben tener exactamente la misma dificultad media. En el caso de PISA, puesto que todos los ítems incluidos en el estudio principal suelen haberse probado en un estudio piloto, los responsables de las pruebas disponen de una cierta idea previa de las dificultades de los ítems y, por tanto, pueden asignarlos a las diferentes pruebas de modo que los ítems de cada una tengan más o menos la misma dificultad media. Sin embargo, las dos pruebas nunca tendrán exactamente la misma dificultad.

La distribución de las dificultades de los ítems influirá en la distribución del rendimiento de los alumnos expresado como puntuación directa. Por ejemplo, una prueba con tan sólo ítems de dificultad media generará una distribución de puntuaciones por alumno diferente de un test que contenga una gran cantidad de ítems difíciles.

Esto también se complica aún más en PISA, ya que este estudio evalúa tres, incluso cuatro áreas de conocimiento por ciclo. Esta evaluación múltiple reduce el número de ítems disponibles para cada área de conocimiento en la prueba, y es más fácil garantizar la comparabilidad de dos pruebas de 60 ítems que dos de, por ejemplo, 15 ítems.

Si las diferentes pruebas se asignan aleatoriamente a los alumnos, puede suponerse la igualdad de las subpoblaciones en cuanto a puntuación media y varianza en el rendimiento de los alumnos. Dicho de otro modo,

- la media de la puntuación directa debería ser idéntica para las distintas pruebas;
- la varianza de las puntuaciones directas de los alumnos debería también ser idéntica.

Si no ocurre así, eso significaría que las distintas pruebas no tienen exactamente las mismas propiedades psicométricas. Para superar este problema se pueden tipificar las puntuaciones directas de los alumnos en cada prueba. Como puede suponerse la igualdad de las subpoblaciones, las diferencias de los resultados se deberán a diferencias en las características de las pruebas. La tipificación neutralizaría el efecto de dichas diferencias en el rendimiento de los alumnos.

Sin embargo, normalmente sólo se realizan pruebas a una muestra de alumnos de las diferentes subpoblaciones. Como ya se ha explicado en los dos capítulos anteriores, el proceso de mues-

treo genera incertidumbre alrededor de cualquier estimación poblacional. Por tanto, incluso si unas pruebas distintas presentan exactamente las mismas propiedades psicométricas y se asignan de modo aleatorio, la media y la desviación típica del rendimiento de los alumnos en las distintas pruebas puede diferir ligeramente. Puesto que las características de las pruebas y la variabilidad del muestreo se conjuntan, no puede suponerse que las puntuaciones directas de los alumnos obtenidas con pruebas distintas sean totalmente comparables.

También pueden esgrimirse otros argumentos psicométricos contra el uso de puntuaciones directas, basadas en el porcentaje de respuestas correctas, para evaluar el rendimiento de los alumnos. Las puntuaciones directas se encuentran en una escala de razón en la medida en que la interpretación del resultado se limita al número de respuestas correctas. Un alumno que obtenga un 0 en esta escala no ha contestado correctamente a ninguna pregunta, pero no podría considerarse que carece completamente de capacidades, mientras que un alumno que obtiene un 10 ha contestado correctamente dos veces más que uno con 5, pero no posee necesariamente el doble de capacidad. Del mismo modo, no podría considerarse que un alumno con una puntuación máxima posea la capacidad también en grado máximo (Wright y Stone, 1979).

El modelo de Rasch para los ítemes dicotómicos

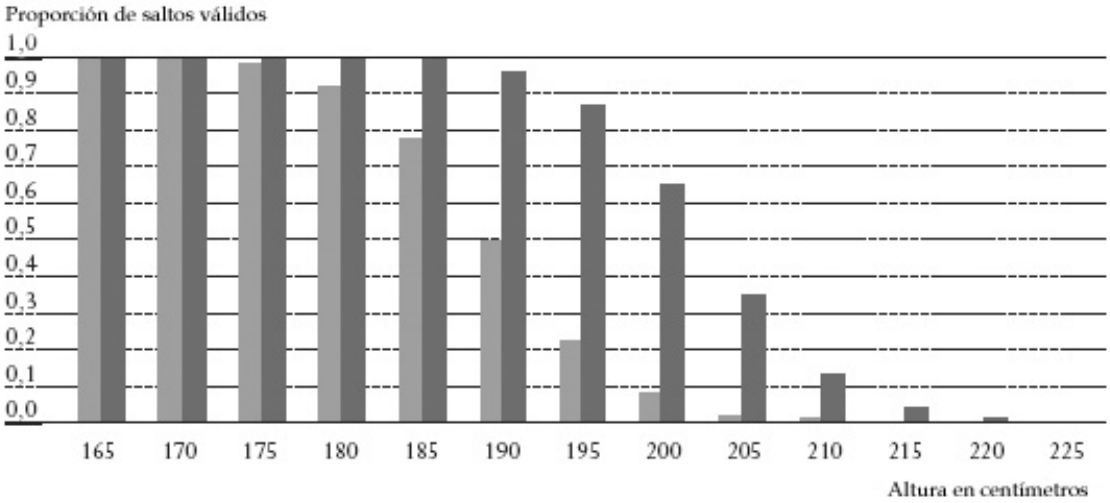
Introducción

Supongamos que alguien desea estimar la competencia de un saltador de altura. Podría medirla o expresarla como su

- mejor marca individual;
- mejor marca individual durante una competición oficial e internacional;
- rendimiento medio durante un período determinado de tiempo;
- rendimiento más frecuente durante un período determinado de tiempo.

La figura 4.1. presenta la proporción de saltos válidos para dos saltadores de altura durante el último año.

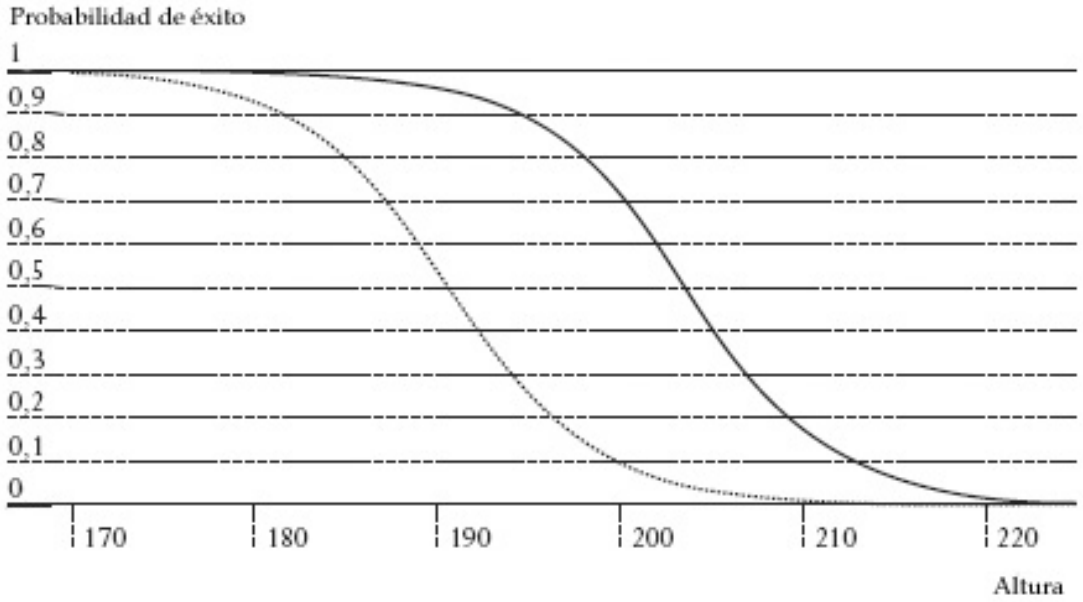
Figura 4.1. Proporción de saltos válidos por altura del salto



Los dos atletas han logrado siempre saltar 165 centímetros. A partir de ahí, la proporción de saltos válidos decae progresivamente hasta que alcanza 0 para ambos saltadores. Sin embargo, mientras que para el primer atleta dicha proporción comienza a descender a los 170 centímetros, para el segundo sólo ocurre a partir de 185 centímetros.

Estos datos pueden ilustrarse gracias a un modelo de regresión logística. Este análisis estadístico consiste en explicar una variable dicotómica mediante una variable continua. En este ejemplo, la variable continua explicará el éxito o el fracaso de un saltador determinado según la altura del salto. El resultado de este análisis permitirá la estimación de la probabilidad de éxito para cada altura. En la figura 4.2 se representa la probabilidad de saltos válidos para los dos atletas.

Figura 4.2. Probabilidad de saltos válidos según la altura del salto para los dos atletas



Estas dos funciones expresan la probabilidad de éxito para los dos saltadores de altura. La curva punteada representa la probabilidad de saltos válidos para el primero y la curva continua, la probabilidad para el segundo.

Por convención,[2] el nivel de rendimiento se definiría como la altura donde la probabilidad de éxito es igual a 0,5. Se trata de una convención razonable, ya que por debajo de ese nivel, la probabilidad de éxito es menor que la de fracaso, y por encima de ese nivel ocurre a la inversa.

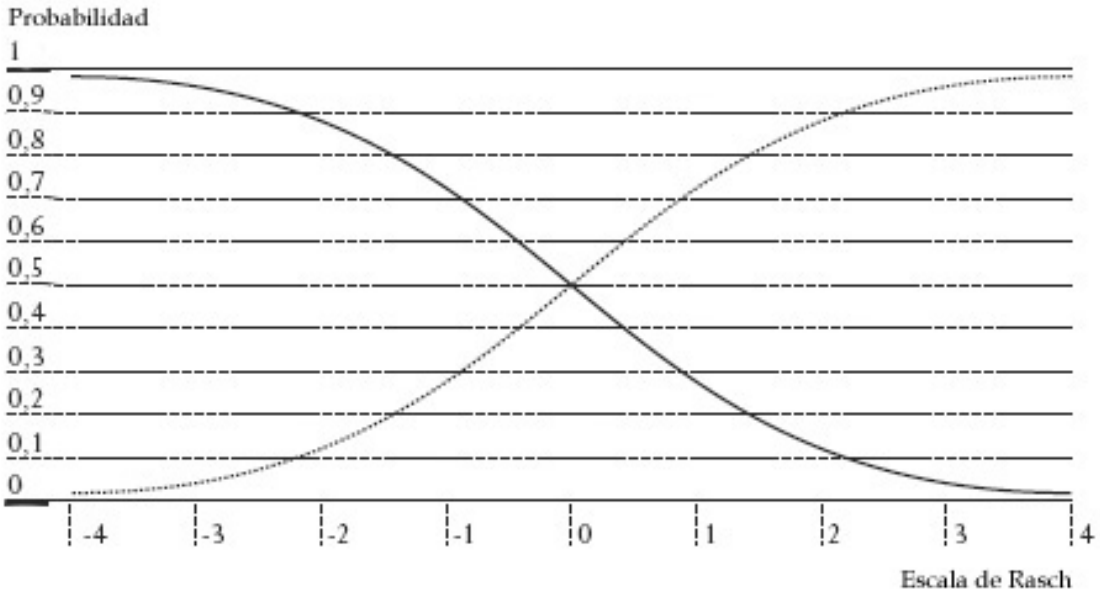
En este ejemplo concreto, el rendimiento de los dos saltadores es, respectivamente, 190 y 202,5. Adviértase, a partir de la figura 4.1, que el rendimiento del primer saltador se observa directamente, mientras que para el segundo saltador no es así y debe estimarse a partir del modelo. Una propiedad clave de este tipo de planteamiento es que la altura de la barra que hay que superar y el rendimiento de los saltadores se expresan en la misma métrica o escala.

El escalamiento de datos cognitivos según el modelo de Rasch sigue el mismo principio. La dificultad de los ítems es análoga a la dificultad del salto basada en la altura de la barra. Además, de la misma forma que un salto concreto tiene dos resultados posibles, es decir, éxito o fracaso, la respuesta de un alumno a una pregunta concreta es o bien correcta, o bien incorrecta.

Por último, así como el rendimiento de cada saltador se definía en el punto donde la probabilidad de éxito era 0,5, la capacidad o el rendimiento del alumno se mide de la misma forma, donde la probabilidad de éxito con un ítem es igual a 0,5.

Una característica del modelo de Rasch es que crea un continuo en el que se localizarán tanto el rendimiento del alumno como la dificultad del ítem, y una función probabilística relaciona estos dos componentes. Los estudiantes con bajo rendimiento y los ítems fáciles estarán situados a la izquierda del continuo o escala, mientras que los estudiantes con alto rendimiento y los ítems difíciles se situarán a la derecha. La figura 4.3 representa la probabilidad de éxito (curva punteada) y la probabilidad de fracaso (curva continua) para un ítem de dificultad cero.

Figura 4.3. Probabilidad de éxito de un ítem de dificultad cero como función de la capacidad del alumno



Como se muestra en la figura 4.3, un alumno con capacidad cero tiene una probabilidad de 0,5 de éxito en un ítem de dificultad cero y una probabilidad de 0,5 de fracaso. Un alumno con capacidad −2 tiene una probabilidad de éxito de algo más de 0,10 y una probabilidad de algo menos de 0,90 de fracaso en el mismo ítem de dificultad cero. Pero este estudiante tendrá una probabilidad de 0,5 de éxito en un ítem de dificultad −2.

Desde un punto de vista matemático, la probabilidad de que un estudiante *i*, con una capacidad β_i, conteste correctamente a un ítem *j* de dificultad δ_j es igual a:

$$P\left(X_{ij}=1 \mid \beta_i, \delta_j\right)=\frac{\exp\left(\beta_i-\delta_j\right)}{1+\exp\left(\beta_i-\delta_j\right)}$$

Del mismo modo, la probabilidad de fracaso es igual a:

$$P\left(X_{ij}=0 \mid \beta_i, \delta_j\right)=\frac{1}{1+\exp\left(\beta_i-\delta_j\right)}$$

Puede demostrarse fácilmente que:

$$P(X_{ij}=1 \mid \beta_i, \delta_j) + P(X_{ij}=0 \mid \beta_i, \delta_j) = 1$$

Dicho de otro modo, la probabilidad de éxito y la probabilidad de fracaso siempre suman uno. Las tablas 4.1 a 4.5 presentan la probabilidad de éxito para distintas capacidades de los alumnos y distintas dificultades de los ítems.

Tabla 4.1. Probabilidad de éxito cuando la capacidad del alumno es igual a la dificultad del ítem

Capacidad del alumno	Dificultad del ítem	Probabilidad de éxito
−2	−2	0,50
−1	−1	0,50
0	0	0,50
1	1	0,50
2	2	0,50

Tabla 4.2. Probabilidad de éxito cuando la capacidad del alumno es 1 unidad menor que la dificultad del ítem

Capacidad del alumno	Dificultad del ítem	Probabilidad de éxito
−2	−1	0,27
−1	0	0,27
0	1	0,27
1	2	0,27
2	3	0,27

Tabla 4.3. Probabilidad de éxito cuando la capacidad del alumno es 1 unidad mayor que la dificultad del ítem

Capacidad del alumno	Dificultad del ítem	Probabilidad de éxito
−2	−3	0,73
−1	−2	0,73
0	−1	0,73
1	0	0,73
2	3	0,73

Tabla 4.4. Probabilidad de éxito cuando la capacidad del alumno es 2 unidades menor que la dificultad del ítem

Capacidad del alumno	Dificultad del ítem	Probabilidad de éxito
−2	0	0,12
−1	1	0,12
0	2	0,12
1	3	0,12
2	4	0,12

Tabla 4.5. Probabilidad de éxito cuando la capacidad del alumno es 2 unidades mayor que la dificultad del ítem

Capacidad del alumno	Dificultad del ítem	Probabilidad de éxito
−2	−4	0,88
−1	−3	0,88
0	−2	0,88
1	−1	0,88
2	0	0,88

Debería advertirse que:

- Cuando la capacidad del alumno es igual a la dificultad del ítem, la probabilidad de éxito siempre será igual a 0,50, sin que importen las posiciones en el continuo ni de la capacidad del alumno ni de la dificultad del ítem.
- Si la dificultad del ítem supera a la capacidad del alumno en una unidad de Rasch, denominada *logit,* la probabilidad de éxito será siempre igual a 0,27, sin que importe la posición en el continuo de la capacidad del alumno.
- Si la capacidad del alumno supera a la dificultad del ítem en un *logit,* la probabilidad de éxito será siempre igual a 0,73, sin que importe la posición en el continuo de la capacidad del alumno.
- Si la capacidad del alumno y la dificultad del ítem están separadas por dos unidades, las probabilidades de éxito serán de 0,12 y 0,88 respectivamente.

A partir de estas observaciones, es evidente que el único factor que influye en la probabilidad de éxito es la distancia en el continuo de Rasch entre la capacidad del alumno y la dificultad del ítem.

Estos ejemplos también ilustran la simetría de la escala. Si la capacidad del alumno es un *logit* menor que la dificultad del ítem, la probabilidad de éxito será 0,27, que es 0,23 menor que la probabilidad de éxito cuando la capacidad y la dificultad son iguales. Si la capacidad del alumno es un *logit* mayor que la dificultad del ítem, la probabilidad de éxito será 0,73, que es 0,23 mayor que la probabilidad de éxito cuando la capacidad y la dificultad son iguales. Del mismo modo, una diferencia de dos *logit* genera un cambio de 0,38.

Calibración del ítem

Por supuesto, en situaciones reales la respuesta de un alumno será correcta o incorrecta; por tanto, ¿cuál es el significado de una probabilidad de éxito de 0,5 en términos de respuestas correctas o incorrectas? Dicho de manera simple, puede hacerse la siguiente interpretación:

- si 100 alumnos con una capacidad de 0 tienen que responder a un ítem de dificultad 0, el modelo predecirá 50 respuestas correctas y 50 incorrectas;
- si un alumno con capacidad de 0 tiene que responder a 100 ítems, todos ellos de dificultad

0, el modelo predecirá 50 respuestas correctas y 50 incorrectas.

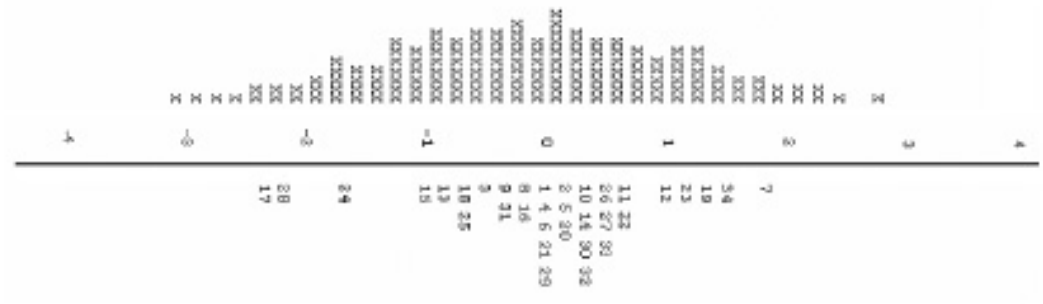
Como ya se ha descrito, el modelo de Rasch, aunque sea una función probabilística, elabora un continuo relativo en el que se localizan la dificultad del ítem y la capacidad del alumno. En el ejemplo de los saltadores de altura, el continuo ya existe, es decir, se trata del continuo físico de la altura en metros. Con datos cognitivos, es necesario construir el continuo. Por analogía, esto consiste en construir un continuo sobre el que se pueda ubicar la altura desconocida de las barras de salto, esto es, la dificultad de los ítems. Hay tres principios básicos subyacentes en la construcción del continuo de Rasch:

- La dificultad relativa de un ítem resulta de la comparación de ese ítem con todos los demás. Supongamos que una prueba consiste en sólo dos ítems. Intuitivamente, el patrón de respuesta (0,0) y (1,1) —donde 1 es una respuesta correcta y 0, incorrecta y donde los pares ordenados se refieren a las respuestas a los ítems 1 y 2— no aporta información para comparar los dos ítems. Las respuestas de estos modelos son idénticas. Por otra parte, las respuestas (1,0) y (0,1) son distintas y sí aportan información en esa comparación. Si 50 estudiantes tienen el patrón de respuesta (0,1) y sólo 10 tienen (1,0), el segundo ítem es considerablemente más fácil que el primero. De hecho, 50 alumnos acertaron en el segundo ítem, mientras que fallaban el primero, y sólo 10 acertaron en el primero, mientras que fallaban el segundo. Esto significa que si una persona acierta en uno de estos dos ítems, la probabilidad de éxito en el segundo ítem es cinco veces más alta que la probabilidad de acertar en el primero. Por lo tanto, es más fácil tener éxito en el segundo que en el primero. Adviértase que la dificultad relativa de los dos ítems es independiente de las capacidades de los alumnos.
- A medida que se determinan las dificultades mediante la comparación de ítems, se crea una escala relativa y, por tanto, hay un número infinito de puntos en esa escala. En términos generales, el proceso de superar esta condición es comparable a la necesidad de crear puntos de anclaje en las escalas de temperatura. Por ejemplo, Celsius fijó dos puntos de referencia: la temperatura a la que se congela el agua y la temperatura a la que hierve. Llamó 0 al primer punto de referencia, 100 al segundo y, como consecuencia, definió la unidad de medida como la centésima parte de la distancia entre los dos puntos de referencia. En el caso del modelo de Rasch, la unidad de medida se define según la función probabilística que relaciona los parámetros de dificultad del ítem y de capacidad del alumno. Por tanto, sólo es necesario definir un punto de referencia. El punto de referencia más común consiste en centrar las dificultades de los ítems en cero. Sin embargo, pueden usarse otros puntos de referencia arbitrarios, como centrar las capacidades de los alumnos en cero.
- Este continuo permite el cálculo de la dificultad relativa de los ítems aplicados parcialmente a distintas subpoblaciones. Supongamos que el primer ítem se aplicó a todos los alumnos y el segundo, sólo a los alumnos de baja capacidad. La comparación de ítems sólo se llevará a cabo en la subpoblación que recibió ambos ítems, es decir, la población de estudiantes de baja capacidad. La dificultad relativa de los dos ítems se basará en este subconjunto común de alumnos.

Una vez que las dificultades de los ítems se han situado en el continuo de Rasch, pueden calcularse las puntuaciones de los alumnos. La línea de la figura 4.4 representa un continuo de Rasch. Las dificultades de los ítems están situadas por encima de esa línea y los números de ítem, por

debajo. Por ejemplo, el ítem 7 representa un ítem difícil y el ítem 17, uno fácil. Esta prueba incluye algunos ítems fáciles, un gran número de ítems de dificultad media y algunos ítems difíciles. Los símbolos *X* por encima de la línea representan la distribución de las puntuaciones de los alumnos.

Figura 4.4. Distribuciones de la puntuación de los alumnos y de la dificultad de los ítems en un continuo de Rasch



Cálculo de la puntuación de un alumno

Una vez que se han situado las dificultades de los ítems en la escala de Rasch, pueden calcularse las puntuaciones de los alumnos. En una sección previa, se mencionó que la probabilidad de que un alumno *i*, con una capacidad β_i, conteste correctamente a un ítem *j* de dificultad δ_j es igual a:

$$P\left(X_{ij}=1 \mid \beta_i, \delta_j\right)=\frac{\exp\left(\beta_i-\delta_j\right)}{1+\exp\left(\beta_i-\delta_j\right)}$$

Del mismo modo, la probabilidad de fracaso es igual a:

$$P\left(X_{ij}=0 \mid \beta_i, \delta_j\right)=\frac{1}{1+\exp\left(\beta_i-\delta_j\right)}$$

El modelo de Rasch supone la independencia de los ítems; es decir, la probabilidad de una respuesta correcta no depende de las respuestas dadas a otros ítems. Como consecuencia, la probabilidad de acertar en dos ítems es igual al producto de las dos probabilidades individuales de acierto.

Consideremos una prueba de cuatro ítems con las siguientes dificultades: −1, −0,5, 0,5 y 1. Existen 16 patrones posibles de respuesta. Estos patrones se presentan en la tabla 4.6.

Tabla 4.6. Posibles patrones de respuesta para una prueba de cuatro ítems

Puntuación directa	Patrones de respuesta
0	(0,0,0,0)
1	(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)
2	(1,1,0,0), (1,0,1,0), (1,0,0,1), (0,1,1,0), (0,1,0,1), (0,0,1,1)
3	(1,1,1,0), (1,1,0,1), (1,0,1,1), (0,1,1,1)
4	(1,1,1,1)

Para cualquier capacidad del alumno llamada β_i, es posible calcular la probabilidad de cualquier patrón de respuesta. Calculemos la probabilidad del patrón de respuesta (1,1,0,0) para tres estudiantes con una capacidad de − 1, 0 y 1.

Tabla 4.7. Probabilidad del patrón de respuesta (1,1,0,0) para tres capacidades de alumnos

			$\beta_i = -1$	$\beta_i = 0$	$\beta_i = 1$
Ítem 1	$\delta_1 = -1$	Respuesta = 1	0,50	0,73	0,88
Ítem 2	$\delta_2 = -0,5$	Respuesta = 1	0,38	0,62	0,82
Ítem 3	$\delta_3 = 0,5$	Respuesta = 0	0,82	0,62	0,38
Ítem 4	$\delta_4 = 1$	Respuesta = 0	0,88	0,73	0,50
Probabilidad de obtener el patrón de respuesta			0,14	0,21	0,14

La probabilidad de acierto para el primer estudiante en el primer ítem es igual a:

$$P(X_{ij} = 1 \mid \beta_i, \delta_j) = P(X_{1,1} = 1 \mid -1, -1)\frac{\exp(-1-(-1))}{1+\exp(-1-(-1))} = 0,5$$

La probabilidad de acierto para el primer estudiante en el segundo ítem es igual a:

$$P(X_{ij} = 1 \mid \beta_i, \delta_j) = P(X_{1,2} = 1 \mid -1, -0,5)\frac{\exp(-1-(-0,5))}{1+\exp(-1-(-0,5))} = 0,38$$

La probabilidad de fallo para el primer estudiante en el tercer ítem es igual a:

$$P(X_{ij} = 0 \mid \beta_i, \delta_j) = P(X_{1,3} = 0 \mid -1, 0,5)\frac{1}{1+\exp(-1-0,5)} = 0,82$$

La probabilidad de fallo para el primer estudiante en el cuarto ítem es igual a:

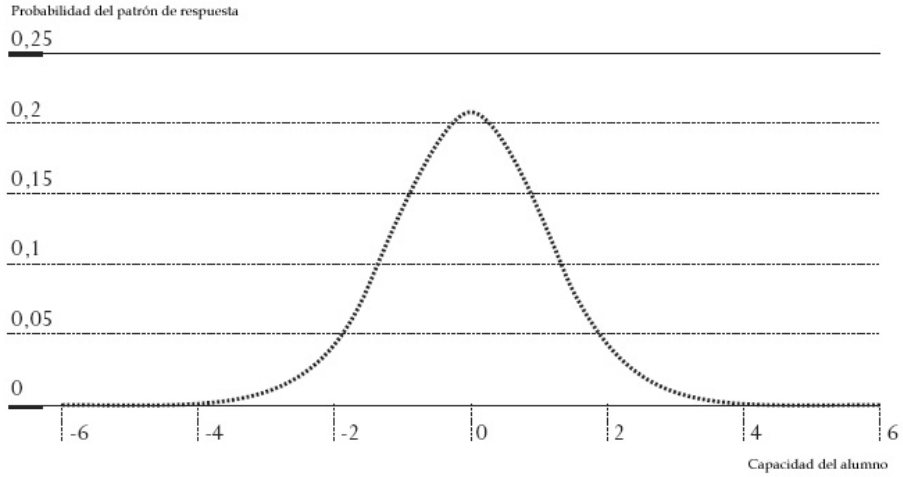
$$P(X_{ij} = 0 \mid \beta_i, \delta_j) = P(X_{1,4} = 0 \mid -1, 1)\frac{1}{1+\exp(-1-1)} = 0,88$$

Como estos cuatro ítems se consideran independientes, la probabilidad del patrón de respuesta (1,1,0,0) para un alumno de capacidad $\beta_i = -1$ es igual a:

$$0,50 \cdot 0,38 \cdot 0,82 \cdot 0,88 = 0,14.$$

Dadas las dificultades del ítem, un alumno con capacidad $\beta_i = -1$ tiene 14 probabilidades entre 100 de responder correctamente a los ítems 1 y 2 de dar una respuesta equivocada en los ítems 3 y 4. Del mismo modo, un alumno con capacidad $\beta_i = 0$ tiene una probabilidad de 0,21 de contestar según el mismo patrón de respuesta y un alumno con capacidad $\beta_i = 1$ tiene una probabilidad de 0,14.

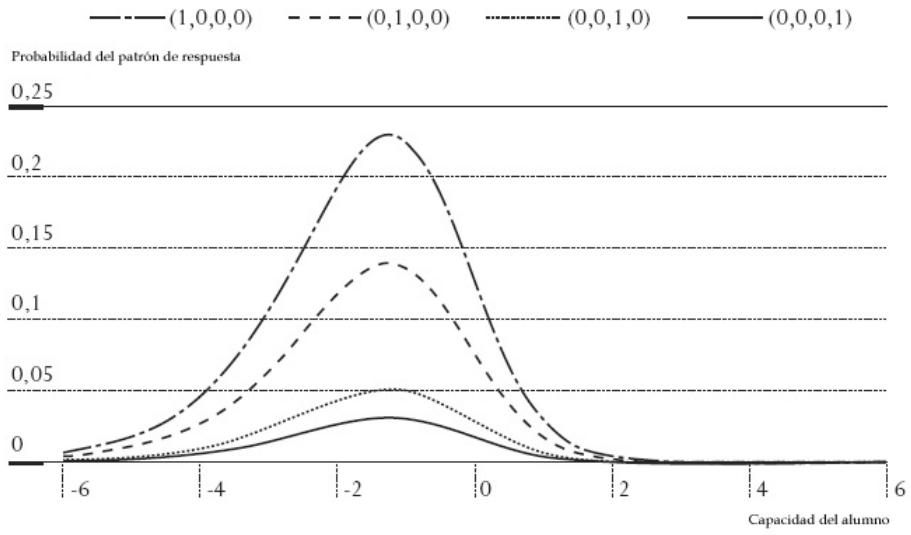
Figura 4.5. Probabilidades para el patrón de respuesta (1,1,0,0)



Este proceso puede aplicarse a una amplia gama de capacidades de alumnos y para todos los patrones de respuesta posibles. La figura 4.5 presenta la probabilidad de observar el patrón de respuesta (1,1,0,0) para todas las capacidades de alumnos entre −6 y +6. Como se ha demostrado, el valor más probable corresponde a una capacidad del alumno de 0. Por tanto, el modelo de Rasch estimará la capacidad de cualquier alumno con un patrón de respuesta (1,1,0,0) en 0.

La figura 4.6 presenta la distribución de las probabilidades para todos los patrones de respuesta con sólo un ítem correcto. Como se muestra en la tabla 4.6, hay cuatro patrones de respuesta con un solo ítem correcto: (1,0,0,0), (0,1,0,0), (0,0,1,0) y (0,0,0,1).

Figura 4.6. Probabilidades de los patrones de respuesta para una puntuación directa de 1



La figura 4.6 muestra claramente que:

- El patrón de respuesta más probable para todos los alumnos que sólo acierten un ítem es

(1,0,0,0) y el patrón menos probable, (0,0,0,1). Cuando un alumno sólo acierta una respuesta, se espera que la respuesta correcta sea al ítem más fácil, es decir, al número 1. También es inesperado que esta respuesta correcta sea al ítem más difícil, es decir, al número 4.

- Sea cual sea el patrón de respuesta, el valor más probable siempre corresponde al mismo valor para la capacidad del alumno. Por ejemplo, la capacidad del alumno más probable para el patrón de respuesta (1,0,0,0) está alrededor de −1,25. Esta también es la capacidad del alumno más probable para los demás patrones de respuesta.

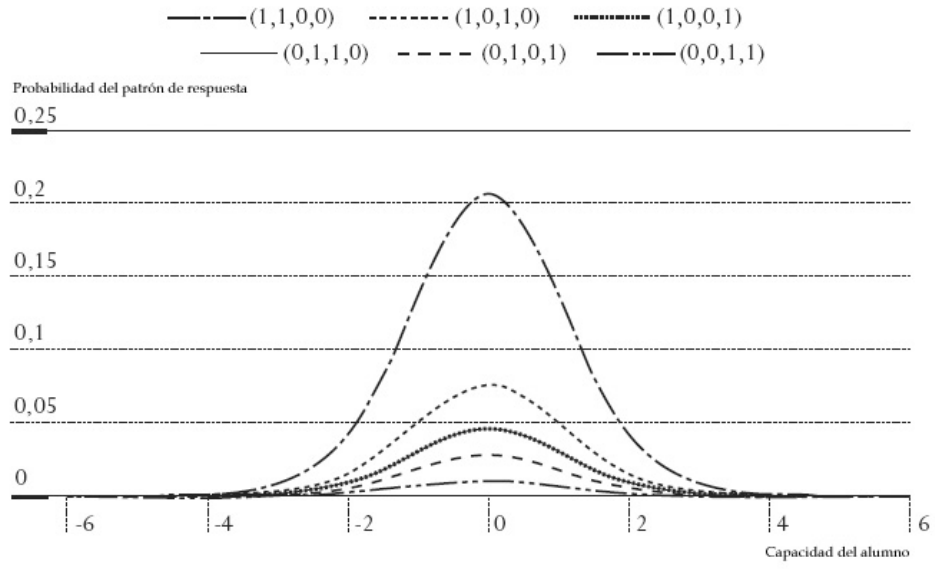
El modelo de Rasch, por tanto, proporcionará el valor −1,25 para todos los estudiantes que den una sola respuesta correcta, sea cual sea el ítem acertado.

De la misma forma, como se muestra en las figuras 4.7 y 4.8:

- el patrón de respuesta más probable con dos ítems correctos es (1,1,0,0);
- la capacidad del alumno más probable siempre es la misma para cualquier patrón de respuesta que incluya dos respuestas correctas (0 en este caso);
- el patrón de respuesta más probable con tres ítems correctos es (1,1,1,0);
- la capacidad del alumno más probable siempre es la misma para cualquier patrón de respuesta que incluya tres respuestas correctas (+1,25 en este caso).

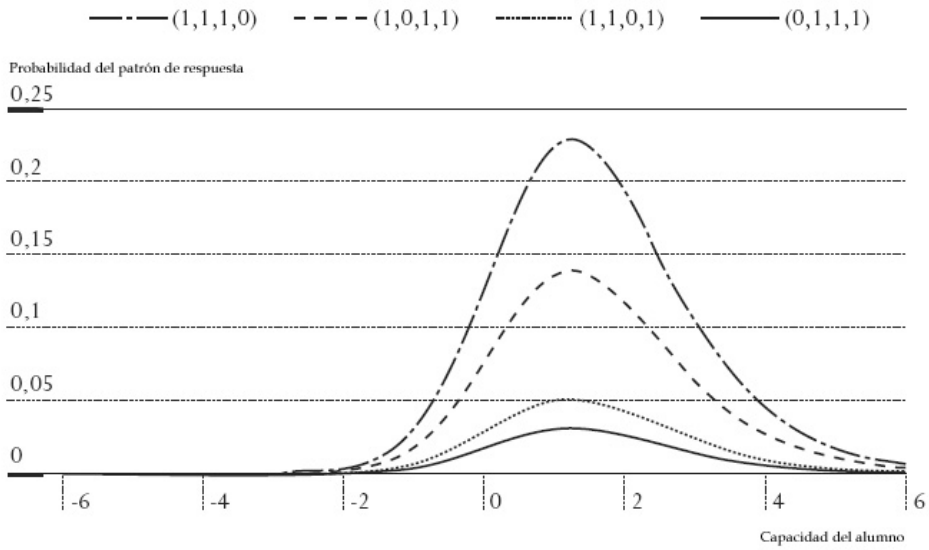
Este tipo de estimación de capacidades de Rasch suele llamarse *estimación de máxima verosimilitud.* Como se muestra en estas figuras, por puntuación directa, es decir, cero respuestas correctas, una respuesta correcta, dos respuestas correctas y así sucesivamente, el modelo de Rasch proporcionará una única estimación de máxima verosimilitud.

Figura 4.7. Probabilidades de los patrones de respuesta para una puntuación directa de 2[a]



[a] En este ejemplo, puesto que la función de verosimilitud para el modelo de respuesta (1,0,0,1) es totalmente similar al del modelo (0,1,1,0), estas dos líneas se superponen en la figura.

Figura 4.8. Probabilidades de los patrones de respuesta para una puntuación directa de 3



Se ha demostrado que esta estimación de máxima verosimilitud (MLE en sus siglas en inglés) está sesgada y se ha propuesto ponderar la contribución de cada ítem según la información que pueda proporcionar dicho ítem (Warm, 1989). Por ejemplo, un ítem difícil no proporciona mucha información para un estudiante de bajo rendimiento. Por otra parte, este ítem puede aportar más información para un alumno de alto rendimiento. Por tanto, para un alumno de bajo rendimiento, los ítems fáciles contribuirán más que los difíciles y, de la misma forma, para un alumno de alto rendimiento, los ítems difíciles contribuirán más que los fáciles. Por tanto, las estimaciones de Warm y las estimaciones MLE son tipos similares de estimación de la capacidad individual de los alumnos.

Puesto que la estimación de Warm corrige el pequeño sesgo de las estimaciones de máxima verosimilitud, suele preferirse como estimación de la capacidad de un individuo. Por tanto, en PISA, se calculan estimaciones de máxima verosimilitud ponderadas (WLE) aplicando pesos a las estimaciones de máxima verosimilitud (MLE), con objeto de corregir el sesgo inherente a ellas, siguiendo la propuesta de Warm.

Cálculo de la puntuación de un alumno para diseños incompletos

Como ya se ha dicho, PISA utiliza un diseño de cuadernillos rotado para superar las exigencias contrapuestas de una duración limitada de las pruebas para los alumnos y una amplia cobertura del área evaluada. Un diseño de prueba en el que los alumnos reciben un subconjunto de los ítems se llama un *diseño incompleto.* Los principios para calcular la estimación de la capacidad individual del alumno descritos en la sección anterior siguen siendo aplicables a los diseños

incompletos.

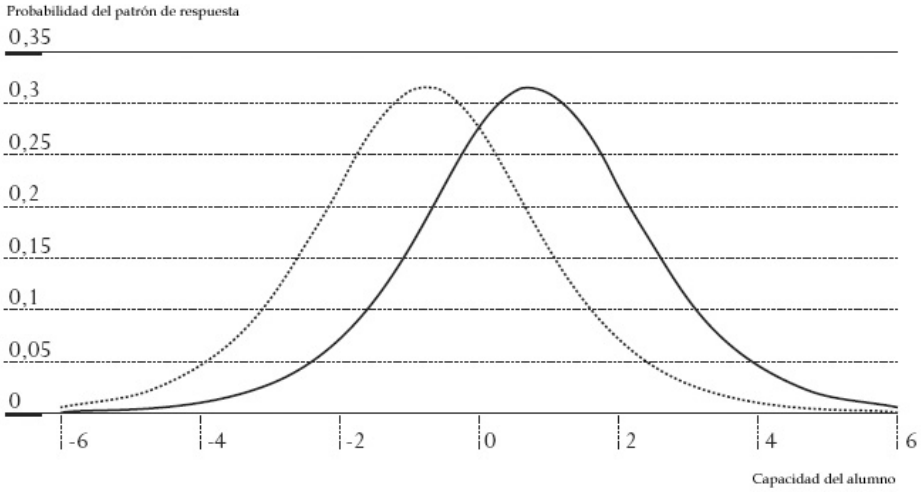
Supongamos que dos alumnos con capacidades de −1 y 1 tienen que responder dos de los cuatro ítems presentados en la tabla 4.8. El alumno con $\beta_i = -1$ debe responder a los primeros dos ítems, es decir, los más fáciles, y el alumno con $\beta_i = 1$ debe responder a los dos últimos, es decir, los más difíciles. Los dos alumnos aciertan en su primer ítem y fallan en el segundo.

Tabla 4.8. Probabilidad del patrón de respuesta (1,0) para dos alumnos de diferente capacidad en un diseño incompleto de tests

			$\beta_i = -1$	$\beta_i = 1$
Ítem 1	$\delta_1 = -1$	Respuesta = 1	0,50	
Ítem 2	$\delta_2 = -0,5$	Respuesta = 0	0,62	
Ítem 3	$\delta_3 = 0,5$	Respuesta = 1		0,62
Ítem 4	$\delta_4 = 1$	Respuesta = 0		0,50
Patrón de respuesta			0,31	0,31

Ambos patrones tienen una probabilidad de 0,31 para ambas capacidades de −1 y 1. Al igual que antes, estas probabilidades pueden calcularse para una amplia gama de capacidades del alumno. La figura 4.9 presenta las probabilidades del patrón de respuesta (1,0) para la prueba fácil (línea punteada) y para la prueba difícil (línea continua).

Figura 4.9. Probabilidad del patrón de respuesta para una prueba fácil y una difícil



Basándonos en la figura 4.9, podemos establecer que para cualquier alumno que haya acertado un ítem del test fácil, el modelo estimará la capacidad del alumno en −0,75, y que para cualquier alumno que haya acertado un ítem del test difícil, el modelo estimará la capacidad del alumno en 0,75. Si las puntuaciones directas se usaran como estimaciones de la capacidad del alumno, en ambos casos obtendríamos 1 de 2, o 0,5.

En resumen, la puntuación directa no tiene en cuenta la dificultad del ítem y, por tanto, la interpretación de la puntuación directa depende de las dificultades de los ítems. Por otra parte, el modelo de Rasch usa el número de respuestas correctas y las dificultades de los ítems aplicados

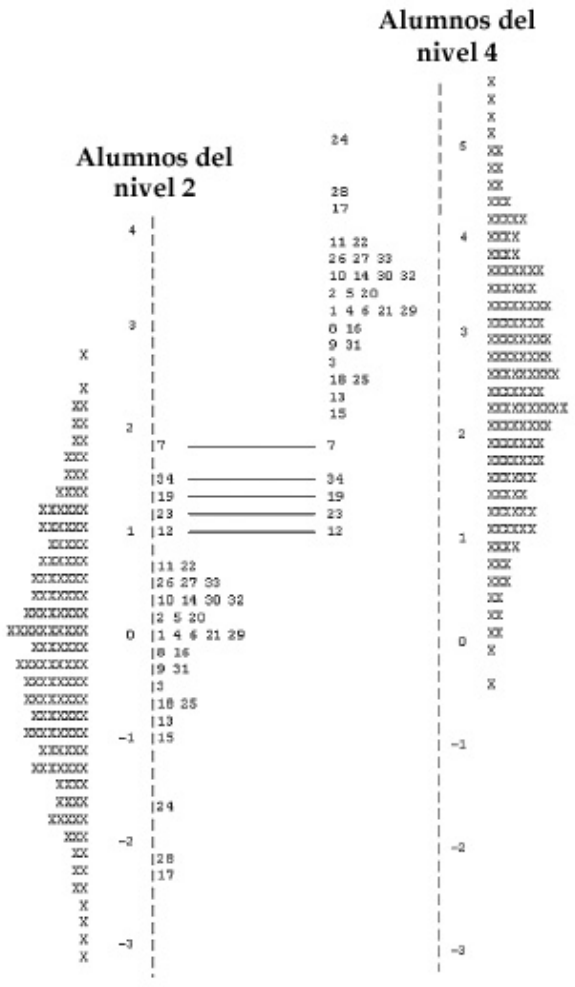
a un alumno concreto para estimar su capacidad. Así pues, una puntuación de Rasch puede interpretarse con independencia de las dificultades de los ítems. En la medida en que todos los ítems puedan ubicarse en el mismo continuo, el modelo de Rasch puede proporcionar estimaciones de la capacidad de los alumnos totalmente comparables, incluso si los alumnos fueron evaluados con distintos subconjuntos de ítems. Nótese, sin embargo, que una determinación válida de la puntuación Rasch del alumno depende de tener conocimiento exacto de las dificultades de los ítems.

Condiciones óptimas para relacionar ítemes

Cuando se usan pruebas diferentes, deben satisfacerse algunas condiciones. En primer lugar, los datos recogidos mediante esas pruebas deben enlanzarse. Sin ningún enlace, los datos recogidos mediante dos pruebas diferentes no pueden ubicarse en una única escala. Normalmente, las pruebas se relacionan haciendo que distintos alumnos contesten a ítems comunes o evaluando a los mismos alumnos con ambas pruebas.

Supongamos que un investigador desea estimar el crecimiento en competencia lectora entre una población de alumnos de 2º curso y otra de 4º curso. Se elaborarán dos pruebas y ambas estarán orientadas al grado de competencia esperado en ambas poblaciones. Para asegurar que ambas pruebas puedan escalarse en el mismo continuo, algunos ítems difíciles de la prueba de 2ºcurso se incluirán en la de 4º curso, digamos los ítems 7, 34, 19, 23 y 12.

Figura 4.10. Anclaje de ítems de Rasch



La figura 4.10 representa este proceso de anclaje entre ítems. La parte izquierda de la figura presenta los resultados del escalamiento de la prueba de 2º curso, con los ítemes centrados en cero. Para el escalamiento de los datos de la de 4º curso, el punto de referencia será la dificultad en 2º curso de los ítems de anclaje. Después, la dificultad de los otros ítemes de 4º curso se fijará según este punto de referencia, como se muestra en la parte derecha de la figura 4.10.

Con este proceso de anclaje, las dificultades de los ítems de los cursos 2º y 4º se situarán en un único continuo. Por tanto, las estimaciones de capacidad de los alumnos de 2º y 4º también se situarán en el mismo continuo.

Para estimar con exactitud el aumento entre los niveles 2 y 4, el investigador garantizará que la ubicación de los ítems de anclaje en ambas pruebas es similar.

Desde un punto de vista teórico, sólo es necesario un ítem para relacionar dos pruebas diferentes. Sin embargo, esta situación está lejos de ser óptima. Un diseño incompleto equilibrado presenta la mejor garantía para transmitir datos de diferentes pruebas en una única escala. PISA

2003 adoptó este sistema, en el que el conjunto de ítems se dividió en 13 bloques. La asignación de ítems a los bloques tiene en cuenta la dificultad esperada de los ítems y el tiempo calculado como necesario para responderlos. La tabla 4.9 presenta el diseño de de las pruebas de PISA 2003. Los 13 bloques de ítems se denominaron de C1 a C13, respectivamente. Se elaboraron 13 cuadernillos, cada uno de ellos con cuatro partes, llamadas grupo 1 a grupo 4. Cada cuadernillo comprendía cuatro bloques. Por ejemplo, el cuadernillo 1 se componía de los bloques 1, 2, 4 y 10.

TABLA 4.9. **Diseño de las pruebas de PISA 2003**

	Grupo 1	*Grupo 2*	*Grupo 3*	*Grupo 4*
Cuadernillo 1	C1	C2	C4	C10
Cuadernillo 2	C2	C3	C5	C11
Cuadernillo 3	C3	C4	C6	C12
Cuadernillo 4	C4	C5	C7	C13
Cuadernillo 5	C5	C6	C8	C1
Cuadernillo 6	C6	C7	C9	C2
Cuadernillo 7	C7	C8	C10	C3
Cuadernillo 8	C8	C9	C11	C4
Cuadernillo 9	C9	C10	C12	C5
Cuadernillo 10	C10	C11	C13	C6
Cuadernillo 11	C11	C12	C1	C7
Cuadernillo 12	C12	C13	C2	C8
Cuadernillo 13	C13	C1	C3	C9

Con este diseño, cada bloque aparece cuatro veces, una en cada posición. Además, cada par de bloques aparece sólo una vez.

Este diseño debería garantizar que el proceso de enlace no se vea influido por la ubicación respectiva de los ítems de anclaje en los distintos cuadernillos.

Extensión del modelo de Rasch

Wright y Masters han generalizado el modelo original de Rasch a ítems politómicos, con el normalmente llamado *modelo de crédito parcial* (Wright y Masters, 1982). Mediante este modelo, los ítems pueden puntuarse como incorrectos, parcialmente correctos y correctos. Los ítems cognitivos de PISA se han calibrado según este modelo.

Este modelo de ítems politómicos también puede aplicarse a datos de escala tipo Likert. Por supuesto, no hay respuestas correctas o incorrectas en tales escalas, pero los principios básicos son los mismos: las posibles respuestas pueden ser ordenadas. Los datos de los cuestionarios de PISA se escalan con el modelo logístico de un parámetro para los ítems politómicos.

Otros modelos de la teoría de la respuesta al ítem

Una distinción clásica entre modelos de la Teoría de Respuesta al Ítem se refiere al número de parámetros empleados para describir los ítems. El modelo de Rasch se designa como un modelo de parámetro único porque las curvas características de los ítems sólo dependen de la dificultad

de éstos. En el modelo logístico de tres parámetros, las curvas características de los ítems dependen de: 1) el parámetro *dificultad* del ítem; 2) el parámetro *discriminación* del ítem; y 3) lo que puede denominarse el parámetro «*acierto por adivinación*». Este último parámetro cubre el hecho de que, en una prueba de elección múltiple, todos los alumnos tienen alguna oportunidad por pura suerte de contestar al ítem correctamente, sin que importe la dificultad de éste.

Conclusiones

El modelo de Rasch fue diseñado para construir un continuo simétrico en el que puedan situarse tanto la dificultad del ítem como la capacidad del alumno. Ambas quedan enlazadas mediante una función logística. Con esta función es posible calcular la probabilidad de que un alumno acierte en un ítem.

Además, debido a esta equiparación probabilística, no es necesario aplicar la batería completa de ítems a cada alumno. Si se asegura la presencia de algunos ítems de anclaje, el modelo de Rasch será capaz de crear una escala en la que se situarán todos los ítems y todos los alumnos. Esta última característica del modelo de Rasch constituye una de las razones principales por las que este modelo se ha vuelto fundamental en las encuestas educativas.

---

[1] Véase *Measuring Student Knowledge and Skills – A New Framework for Assessment* (OCDE, 1999a) y *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (OCDE, 2003b) (existe traducción española: *Marcos teóricos de PISA 2003: Conocimientos y destrezas en Matemáticas, Lectura, Ciencias y Solución de problemas,* Madrid: Ministerio de Educación y Ciencia, Instituto Nacional de Evaluación y Calidad del Sistema Educativo, 2004).

[2] Las probabilidades de 0,5 fueron utilizadas pro primera vez por las teorías psicofísicas (Guilford, 1954).

Capítulo 5

Los valores plausibles

Estimaciones individuales frente a estimaciones de población

Las pruebas educativas pueden tener dos propósitos principales:

- Medir el conocimiento y las destrezas de determinados alumnos. Normalmente, el rendimiento de cada alumno influirá en su futuro (carrera académica, admisión en estudios postsecundarios, etcétera). Por tanto, es especialmente importante minimizar el error de medida asociado con la estimación de cada individuo.
- Evaluar el conocimiento y las destrezas de una población. El rendimiento de los individuos no tendrá impacto en su carrera académica ni en su vida profesional. En tal caso, el propósito de reducir el error al realizar inferencias sobre la población es más importante que el propósito de reducir el error al nivel del individuo.

Las encuestas educativas nacionales o internacionales pertenecen a la segunda categoría.

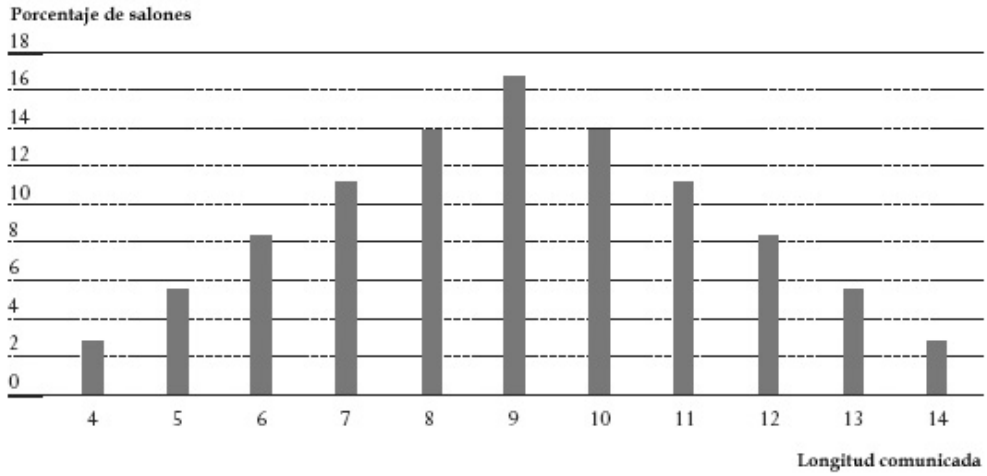
Las encuestas internacionales, como PISA, expresan el rendimiento de los alumnos mediante *valores plausibles* (PV, de sus siglas en inglés).[1] Este capítulo explicará el significado conceptual de los valores plausibles y la ventaja de utilizarlos al presentar los resultados. Se compararán los estimadores individuales (como las estimaciones de máxima verosimilitud ponderada [WLE] definidas en el capítulo 4) con los valores plausibles cuando el objetivo consiste en estimar una gama de estadísticos poblacionales.

El significado de los valores plausibles

Un ejemplo tomado de las ciencias físicas, la medida de un área, servirá para ilustrar este complejo concepto. Supongamos que el ayuntamiento de una ciudad decide imponer un nuevo impuesto sobre inmuebles para aumentar la recaudación. Este nuevo impuesto será proporcional a la longitud del salón de la vivienda familiar. Unos inspectores visitarán todos los hogares de la ciudad para medir la longitud de los salones, con cinta de medir e instrucciones de anotar la longitud en términos de números enteros tan sólo, es decir, 1 metro, 2 metros, 3 metros, 4 metros y así sucesivamente.

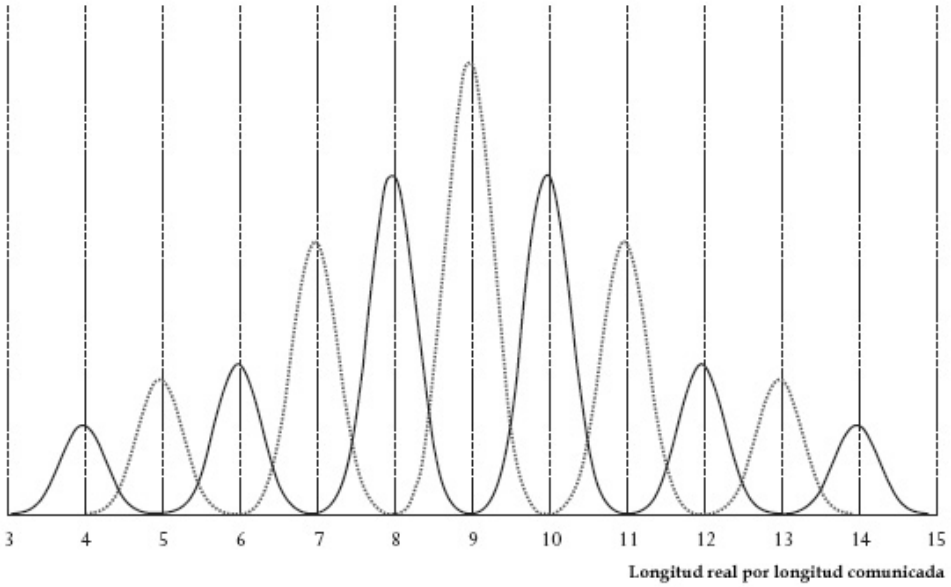
Los resultados de estas mediciones se muestran en la figura 5.1. Alrededor de un 3% de los salones tienen una longitud de 4 metros, algo más del 16% tienen una longitud de 9 metros, etcétera.

Figura 5.1. Longitud de los salones expresada en números enteros



Por supuesto, la realidad es bastante distinta, ya que la longitud es una variable continua. Con una variable continua, las observaciones pueden tener cualquier valor entre el mínimo y el máximo. Por otra parte, con una variable discontinua, las observaciones sólo pueden tener un número predefinido de valores. La figura 5.2 muestra la distribución de la longitud de los salones según la longitud anotada.

Figura 5.2. Longitud real según la longitud anotada



Todos los salones con una longitud anotada de 5 metros no tienen exactamente 5 metros. De media, tienen 5 metros de largo, pero su longitud varía alrededor de la media. La diferencia entre la longitud anotada y la longitud real se debe al redondeo y al error de la medida. Un inspector podría medir incorrectamente 5 metros en un determinado salón cuando en realidad

este mida 4,15 metros. Si el redondeo fuera la única fuente de error, la longitud anotada debería ser 4 metros. La segunda fuente de error, el error en la medida, explica el solapamiento en la distribución.

En este ejemplo concreto, las longitudes de los salones se distribuyen normalmente alrededor de la media, que también es la longitud anotada. Si la diferencia entre la longitud y el número entero más cercano es pequeña, la probabilidad de no anotar esta longitud con el número entero más cercano es muy baja. Por ejemplo, es improbable que una longitud de 4,15 metros se anote como 5 metros o 3 metros. Sin embargo, a medida que aumenta la distancia entre la longitud real y el número entero más cercano, la probabilidad de no registrar esta longitud mediante el entero más cercano aumentará también. Por ejemplo, es probable que una longitud de 4,95 metros se anote como de 5 metros, mientras que una longitud de 4,5 se anotará tanto como de 4 metros que como de 5 metros.

La metodología de los valores plausibles consiste en:

- calcular matemáticamente las distribuciones (llamadas *distribuciones "a posteriori" o posteriores*) alrededor de los valores anotados y la longitud comunicada en el ejemplo;
- asignar a cada observación un conjunto de valores aleatorios tomados a partir de las distribuciones posteriores.

Por lo tanto, los valores plausibles pueden definirse como valores aleatorios a partir de las distribuciones posteriores. En el ejemplo, a un salón de 7,154 metros que se registró como de 7 metros podría asignársele cualquier valor de la distribución normal alrededor de la longitud comunicada de 7. Podría ser 7,45, 6,55 o 6,95. Por tanto, los valores plausibles no deberían utilizarse para estimaciones individuales.

Este ejemplo ficticio del campo de las ciencias físicas puede trasladarse a las ciencias sociales. Por ejemplo, con una prueba de 6 ítems dicotómicos, una variable continua (por ejemplo, la capacidad mental) puede ser transformada en una variable discontinua. La variable discontinua será la puntuación directa del alumno o el número de respuestas correctas. Las únicas puntuaciones posibles son 0, 1, 2, 3, 4, 5 ó 6.

Al contrario que la mayoría de las medidas en las ciencias naturales, las medidas psicológicas o educativas incluyen errores considerables de medida, porque

- el concepto que se mide es más amplio;
- pueden verse afectadas por la disposición mental y física de los alumnos el día de la evaluación;
- las condiciones en que se realice la prueba también pueden influir en los resultados.

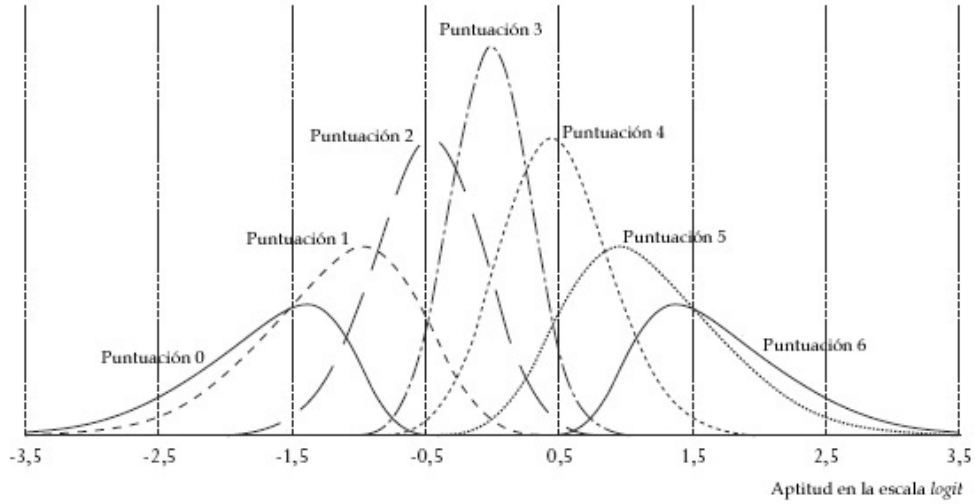
Esto significa que existen grandes solapamientos en las distribuciones posteriores, como se ilustra en la figura 5.3.

Además, con el ejemplo del salón, el error de medida de las distribuciones posteriores puede considerarse independiente del salón.[2] En educación, el error de medida no siempre es independiente del nivel de rendimiento de los alumnos. Puede ser más pequeño para los estudiantes de rendimiento medio y mayor para los que tienen rendimiento muy alto o muy bajo.

Además, en este ejemplo particular, las distribuciones posteriores para la puntuación 0 y la

puntuación 6 son considerablemente asimétricas, como lo serían las distribuciones posteriores de los salones con una longitud comunicada de 4 y 14 metros, si todos los salones menores de 4 metros se reseñaron como de 4 y todos los mayores de 14, como de 14. Esto significa que las distribuciones posteriores no se distribuyen normalmente, como se muestra en la figura 5.3.

Figura 5.3. Una distribución posterior de un test de 6 ítemes



Generar valores plausibles en una prueba educativa consiste en obtener números aleatorios a partir de las distribuciones posteriores. Este ejemplo muestra claramente que los valores plausibles no deberían utilizarse como medida del rendimiento individual. De hecho, un alumno que obtenga una puntuación 0 podría recibir −3, pero también −1. Un alumno que obtenga una puntuación de 6 podría recibir 3, pero también 1.

Se ha hecho notar que:

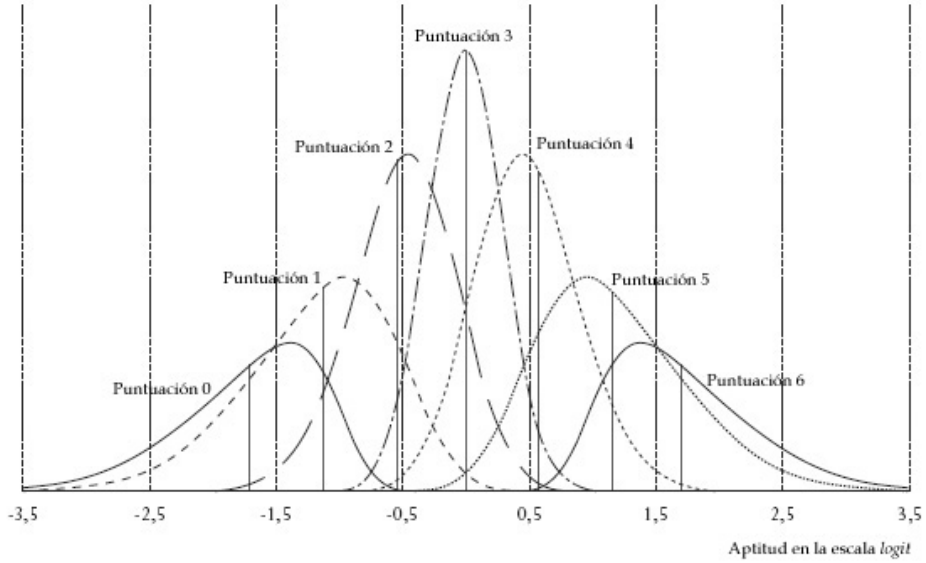
> El método más sencillo para describir valores plausibles es decir que los valores plausibles son una representación de la gama de capacidades que pueden suponerse razonablemente en un alumno. [...] En lugar de estimar directamente la capacidad θ de un alumno, se estima una distribución de probabilidad para θ. Es decir, en lugar de obtener una estimación puntual para θ de un alumno (como una estimación WLE [de máxima verosimilitud ponderada]), se estima un abanico de valores posibles para la magnitud θ de un alumno, con una probabilidad asociada para cada uno de estos valores. Los valores plausibles son selecciones aleatorias de esta distribución (estimada) de θ para un alumno (Wu y Adams, 2002).[3]

Toda esta metodología pretende elaborar un continuo a partir de una colección de variables discontinuas (es decir, de las puntuaciones de las pruebas). Está pensada para evitar que se produzcan inferencias sesgadas como resultado de medir una capacidad subyacente inobservable mediante una prueba que contiene un número relativamente pequeño de ítems.

Por último, a partir de las distribuciones posteriores también puede obtenerse una estimación individual de la capacidad del alumno. Esta estimación individual derivada se llama *estimador esperado a posteriori* (EAP). En vez de asignar un conjunto de valores aleatorios a partir de las

distribuciones posteriores, se asigna la media de estas últimas. Por lo tanto, el EAP puede considerarse como la media de un conjunto infinito de valores plausibles para un alumno determinado.

Figura 5.4. Estimadores EAP



Puesto que sólo se asigna un valor por cada distribución posterior, el estimador EAP es también una variable discontinua.[4] Sin embargo, las estimaciones EAP y las estimaciones de verosimilitud ponderada (WLE) difieren, ya que las primeras requieren un supuesto de distribución de la población, lo que no es el caso para las segundas. Además, mientras que cualquier puntuación directa de una prueba determinada se asociará siempre con una única WLE, a una determinada puntuación directa podrá asociarse distintos valores EAP, según las variables predictoras que se utilicen como variables de condicionamiento.

Los investigadores que no estén acostumbrados a trabajar con valores plausibles podrían considerar esta aparente aleatoriedad como una fuente de imprecisión. Pero esta impresión será neutralizada cuando se comparen distintos tipos de estimadores Rasch (WLE, EAP y PV) de la capacidad del alumno para la estimación de estadísticas poblacionales. Aunque la base de datos de PISA 2003 sólo incluye valores plausibles,[5] la comparación incorporará estimadores EAP para demostrar los sesgos que aparecen cuando los analistas de datos utilizan la media de los valores plausibles para obtener una única puntuación por alumno.

Comparación de la eficacia de las *Estimaciones de Máxima Verosimilitud de Warm*, de las *Estimaciones Esperadas A Posteriori* y de los *Valores Plausibles* para la estimación de algunos estadísticos de poblacionales[6]

Una comparación entre distintos estimadores de la capacidad de los alumnos puede realizarse sobre datos reales. Tal comparación identificará diferencias, pero no identificará los mejores estimadores para un estadístico poblacional concreto. Puede utilizarse una simulación para ilustrarlo.

La simulación consiste en tres pasos principales:

- Generación de un conjunto de datos, que comprende una variable continua que represente las capacidades del alumno (considerada como la *variable latente*), algunas variables de entorno (el género y un índice de entorno social llamado HISEI), y un patrón de respuestas a los ítems con código 0 para las respuestas incorrectas y 1 para las respuestas acertadas. Los resultados presentados a continuación se basan en una prueba ficticia de 15 ítems.[7]
- Cálculo de los estimadores de capacidad del alumno, esto es, WLE, EAP y los PV[8].
- Estimación de algunos parámetros poblacionales que involucran la capacidad del alumno (es decir, la variable latente) y los diferentes estimadores de esa capacidad. Se realizará una comparación para:
  - la media, la varianza y algunos percentiles;
  - la correlación;
  - la varianza entre centros y dentro de los centros.

El conjunto de datos contiene 5.250 alumnos distribuidos en 150 centros con 35 alumnos por centro. La tabla 5.1 presenta la estructura del conjunto de datos simulado antes de importar los estimadores Rasch de la capacidad del alumno.

Tabla 5.1. Estructura de los datos simulados

Centro	Alumno	Sexo	HISEI	Ítem 1	Ítem 2	...	Ítem 14	Ítem 15
001	01	1	32	1	1		0	0
001	02	0	45	1	0		1	0
...	...							
150	34	0	62	0	0		1	1
150	35	1	50	0	1		1	1

La tabla 5.2 presenta la media y la varianza de la variable latente, los estimadores WLE y EAP, y los cinco PV. También se incluye la media de los cinco PV.

Tabla 5.2. Medias y varianzas de la variable latente y de los distintos estimadores de la capacidad del alumno

	Media	Varianza
Variable latente	0,00	1,00
WLE	0,00	1,40
EAP	0,00	0,75
PV1	0,01	0,99
PV2	0,00	0,99
PV3	0,00	1,01
PV4	0,00	1,01
PV5	−0,01	0,00
Media de los 5 estadísticos PV	0,00	1,00

La tabla 5.2 muestra que una buena estimación de la media de la población (es decir, la estimación de la variable latente) se obtiene cualquiera que sea el tipo de variable latente usada (estimadores WLE o EAP, o valores plausibles). Puede demostrarse empíricamente que ninguna de

las estimaciones difiere significativamente de la media esperada, es decir, 0,00 en este caso concreto (Wu y Adams, 2002). Además, también se puede demostrar que la media de los WLE no será sesgada si la prueba está bien orientada, es decir, si la media de las dificultades de los ítems está alrededor de 0 en la escala de Rasch (Wu y Adams, 2002). Esto es, en una prueba bien orientada, los alumnos obtendrán una puntuación directa de aproximadamente un 50% de respuestas correctas. Si la prueba es demasiado fácil, la media de los WLE quedará subestimada (esto se llama *efecto techo*), mientras que, si es demasiado difícil, la media de los WLE quedará sobreestimada (esto se llama *efecto suelo*).

Estos últimos resultados explican por qué la media de los WLE proporcionada en la base de datos de PISA 2000 difiere de la media de los valores plausibles, especialmente para los países no pertenecientes a la OCDE. Para la subescala de reflexión en lectura, las medias obtenidas para Canadá utilizando WLE y PV son, respectivamente, 538,4 y 542,2 (es decir, muy cercanas). En contraste, las medias obtenidas para el Perú usando WLE y PV son, respectivamente, 352,2 y 322,7, que es una diferencia de aproximadamente 0,3 desviaciones típicas. Existe sesgo cuando se usan WLE para estimar la media, si la prueba no está bien orientada. Esta comparación no puede realizarse sobre la base de datos de PISA 2003, ya que sólo contiene el rendimiento de los alumnos con valores plausibles.

Para la varianza de la población, la tabla 5.2 muestra que los valores plausibles dan estimaciones más cercanas al valor esperado, mientras que los WLE lo sobreestiman y los EAP lo infraestiman. Los resultados coinciden con otros estudios de simulación.

La tabla 5.3 presenta algunos percentiles calculados sobre los diferentes estimadores de capacidad. Por ejemplo, puesto que la varianza calculada mediante valores plausibles no está sesgada, los percentiles basados en valores plausibles tampoco están sesgados. Sin embargo, como las estimaciones EAP y las varianzas de WLE están sesgadas, los percentiles, sobre todo los percentiles extremos, también estarán sesgados. Estos resultados coinciden con otros estudios de simulación ya citados.

La tabla 5.4 presenta la correlación entre el índice de entorno social HISEI y el género con las variables latentes y los distintos estimadores de la capacidad de los alumnos. Los coeficientes de correlación con los WLE están subestimados, mientras que los coeficientes de correlación con los estimadores EAP están sobreestimados. Sólo los coeficientes de correlación con los valores plausibles no presentan sesgo.[9]

Tabla 5.3. Percentiles para la variable latente y los distintos estimadores de la capacidad del alumno

	P5	P10	P25	P50	P75	P90	P95
Variable latente	−1,61	−1,26	−0,66	0,01	0,65	1,26	1,59
WLE	−2,15	−1,65	−0,82	−0,10	0,61	1,38	1,81
EAP	−1,48	−1,14	−0,62	−0,02	0,55	1,08	1,37
PV1	−1,68	−1,29	−0,71	−0,03	0,64	1,22	1,59
PV2	−1,67	−1,31	−0,69	−0,03	0,62	1,22	1,58
PV3	−1,67	−1,32	−0,70	−0,02	0,64	1,21	1,56
PV4	−1,69	−1,32	−0,69	−0,03	0,63	1,23	1,55
PV5	−1,65	−1,30	−0,71	−0,02	0,62	1,20	1,55
Media de los 5 estadísticos PV	−1,67	−1,31	−0,70	−0,03	0,63	1,22	1,57

Tabla 5.4. Correlación entre HISEI y género con la variable latente y los distintos estimadores de la capacidad del alumno

	HISEI	GÉNERO
Variable latente	0,40	0,16
WLE	0,33	0,13
EAP	0,46	0,17
PV1	0,41	0,15
PV2	0,42	0,15
PV3	0,42	0,13
PV4	0,40	0,15
PV5	0,40	0,14
Media de los 5 estadísticos PV	0,41	0,14

Debería hacerse notar que ninguno de los coeficientes de regresión está sesgado para los distintos tipos de estimadores. Sin embargo, como las varianzas están sesgadas para algunos estimadores, las varianzas residuales también presentarán sesgo. Por ello, el error típico de los coeficientes de regresión estará sesgado en los casos de los estimadores WLE y EAP.

Por último, la tabla 5.5 presenta las varianzas entre los centros y dentro de los centros. Las varianzas entre centros para los distintos estimadores no difieren del valor esperado de 0,33. Sin embargo, los WLE sobreestiman la varianza dentro del centro, mientras que los EAP la subestiman. Estos resultados coinciden con otros estudios de simulación (Monseur y Adams, 2002).

Como muestra este ejemplo, los valores plausibles proporcionan estimaciones sin sesgo.

Cómo llevar a cabo análisis con valores plausibles

De acuerdo con lo establecido en la sección anterior, un conjunto de valores plausibles, normalmente 5, se seleccionan para cada alumno dentro cada escala o subescala. Los estadísticos poblacionales deberían estimarse utilizando cada valor plausible por separado. El estadístico poblacional obtenido es, entonces, la media de los 5 estadísticos de valor plausible. Por ejemplo,

si estamos interesados en el coeficiente de correlación entre el índice social y el rendimiento de lectura en PISA, deberían calcularse y, después, promediarse cinco coeficientes de correlación.

Tabla 5.5. Varianzas entre los centros y dentro de los centros

	Varianza entre centros	Varianza dentro de los centros
Variable latente	0,33	0,62
WLE	0,34	1,02
EAP	0,35	0,38
PV1	0,35	0,61
PV2	0,36	0,60
PV3	0,36	0,61
PV4	0,35	0,61
PV5	0,35	0,61
Media de los 5 estadísticos PV	0,35	0,61

Los analistas de datos no deberían promediar nunca los valores plausibles para cada alumno, es decir, calcular en el conjunto de datos la media de los cinco valores plausibles a nivel del alumno y luego calcular el estadístico de interés a partir de ese valor promedio. Hacerlo así sería equivalente a una estimación EAP, que resulta sesgada, como se ha descrito en la sección anterior.

Matemáticamente, los análisis secundarios con valores plausibles pueden describirse como sigue. Si θ es el estadístico poblacional y θ_i es el estadístico de interés calculado sobre un valor plausible, entonces:

$$\theta = \frac{1}{M}\sum_{i=1}^{M}\theta_i \text{ , donde } M \text{ es el número de valores plausibles.}$$

Los valores plausibles también permiten calcular la incertidumbre en la estimación de θ debida a la falta de precisión en la medida por parte de la prueba. Si pudiera desarrollarse una prueba perfecta, el error de medida sería igual a cero y los cinco estadísticos a partir de los valores plausibles serían idénticos. Por desgracia, las pruebas perfectas no existen y no existirán nunca. Esta varianza de la medida, llamada normalmente *varianza de imputación*, es igual a:

$$B_M = \frac{1}{M-1}\sum_{i=1}^{M}(\theta_i - \theta)^2$$

Se corresponde con la varianza de los cinco estadísticos de interés obtenidos a partir de los valores plausibles. La etapa final supone combinar la varianza muestral y la varianza de imputación, como se muestra a continuación:

$$V = U + \left(1 + \frac{1}{M}\right)B_M \text{ , donde } U \text{ es la varianza muestral.}$$

En los próximos capítulos, mostraremos cómo calcular varianzas muestrales y varianzas de imputación y cómo combinarlas, utilizando la base de datos de PISA 2003.

Conclusiones

Este capítulo ha estado dedicado al significado de los valores plausibles y a los pasos necesarios cuando éstos se utilizan para analizar datos. Se ha presentado una comparación entre valores plausibles y estimaciones alternativas de la capacidad individual para convencer a los usuarios de los datos de PISA acerca de la superioridad de esta metodología para obtener estimaciones poblacionales.

---

[1] La metodología de los valores plausibles se implementó por primera vez en los estudios NAEP (Beaton, 1987).

[2] El error de medida será independiente de la longitud de los salones si los inspectores utilizan un instrumento de medida que tenga al menos 15 metros de longitud (como una cinta métrica). Si usan un metro ordinario, el error de medida total será proporcional a la longitud del salón.

[3] La distribución de probabilidad para la capacidad θ de un alumno puede basarse tan sólo en los datos cognitivos, es decir, en el patrón de respuesta al ítem, pero también puede incluir información adicional, como el género, el entorno social, etcétera, del alumno. La distribución de probabilidad queda, por tanto, condicionada por esta información adicional. En *PISA 2000 Technical Report* (OCDE, 2002*c*) puede encontrarse una explicación matemática del modelo utilizado para el escalamiento en PISA 2000.

[4] Si se utilizan varias variables predictoras como variables de condicionamiento, el estimador EAP tiende a ser una variable continua.

[5] Los archivos de datos de PISA incluyen tanto WLE como PV.

[6] Los estimadores EAP y los valores plausibles pueden calcularse con o sin variables predictoras. Puesto que los valores plausibles de PISA 2000 se generaron basándose en todas las variables recogidas mediante los cuestionarios de los alumnos, esta comparación sólo incluirá PV y estimadores EAP con el uso de variables predictoras.

[7] La generación de datos comienza con un análisis factorial en una matriz de correlación cuadrada de 3 por 3. La correlación entre la variable latente y el género se estableció en 0,20; la correlación entre la variable latente y el indicador de entorno social se estableció en 0,40 y la correlación entre el género y el indicador de entorno social se estableció en 0,00. Se toman tres variables aleatorias a partir de distribuciones normales y se combinan según los coeficientes de regresión factoriales para crear las tres variables de interés: lectura, sexo y entorno social. Basándose en la puntuación de los alumnos en la variable latente y un conjunto predefinido de 20 dificultades de ítems, las probabilidades de acierto se calculan según el modelo de Rasch. A continuación, estas probabilidades se comparan con la distribución uniforme y se recodifican en 0 y 1. Por último, el género se recodifica en una variable dicotómica.

[8] Los estimadores se calcularon con el programa Conquest, desarrollado por M. L. Wu, R. J. Adams y M. R. Wilson.

[9] Los resultados de los coeficientes de correlación de EAP y PV se observan cuando las distribuciones de probabilidad se generan con variables de condicionamiento. Sin el condicionamiento, la correlación con los valores plausibles quedaría subestimada.

Capítulo 6

El cálculo de errores típicos

Introducción

Como se ha mostrado en el capítulo 3, es necesario utilizar replicaciones para calcular el error típico de cualquier estimación poblacional. Este capítulo dará ejemplos de tales cálculos.

En PISA 2000 y PISA 2003, se utiliza la variante de Fay del método BRR. La fórmula general para calcular las varianzas muestrales con este método es:

$$\sigma_{\hat{\theta}}^2 = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

Puesto que las bases de datos de PISA incluyen 80 replicaciones y puesto que el coeficiente de Fay se estableció en 0,5 para ambas recopilaciones de datos, la fórmula anterior puede simplificarse así:

$$\sigma_{\hat{\theta}}^2 = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2 = \frac{1}{80(1-0{,}5)^2}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2 = \frac{1}{20}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

El error típico de estadísticos univariantes para variables numéricas

Para calcular la media y su respectivo error típico, es necesario calcular primero este estadístico ponderando los datos con el peso final del alumno, es decir, W_FSTUWT, y después calcular otras 80 medias, cada una de ellas ponderando los datos con una de las 80 replicaciones, es decir, de W_FSTR1 a W_FSTR80.

El cuadro 6.1 presenta la sintaxis de SPSS® para calcular estas 81 medias del índice de entorno social (llamado HISEI) con los datos de PISA 2003 para Alemania. La tabla 6.1 presenta el estimador final de HISEI, así como los 80 estimadores replicados.

Cuadro 6.1. Sintaxis de SPSS® para el cálculo de 81 medias

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if (cnt='DEU').
Weight by w_fstuwt.
means HISEI /CELL=mean.

Weight by w_fstr1.
means HISEI /CELL=mean.
Weight by w_fstr2.
means HISEI /CELL=mean.
.
.
.
Weight by w_fstr79.
means HISEI /CELL=mean.
Weight by w_fstr80.
means HISEI /CELL=mean.
```

La media que se publicará es igual a 49,33, es decir, la estimación obtenida con el peso final de los alumnos W_FSTUWT. Las 80 estimaciones replicadas se utilizan tan sólo para calcular el error típico de esa media de 49,33.

Tabla 6.1. Estimaciones de las medias de HISEI

Peso	Estimación de la media		Peso	Estimación de la media
Peso final	**49,33**			
Replicación 1	49,44		Replicación 41	49,17
Replicación 2	49,18		Replicación 42	49,66
Replicación 3	49,12		Replicación 43	49,18
Replicación 4	49,46		Replicación 44	49,04
Replicación 5	49,24		Replicación 45	49,42
Replicación 6	49,34		Replicación 46	49,72
Replicación 7	49,13		Replicación 47	49,48
Replicación 8	49,08		Replicación 48	49,14
Replicación 9	49,54		Replicación 49	49,57
Replicación 10	49,20		Replicación 50	49,36
Replicación 11	49,22		Replicación 51	48,78
Replicación 12	49,12		Replicación 52	49,53
Replicación 13	49,33		Replicación 53	49,27
Replicación 14	49,47		Replicación 54	49,23
Replicación 15	49,40		Replicación 55	49,62
Replicación 16	49,30		Replicación 56	48,96
Replicación 17	49,24		Replicación 57	49,54
Replicación 18	48,85		Replicación 58	49,14
Replicación 19	49,41		Replicación 59	49,27
Replicación 20	48,82		Replicación 60	49,42
Replicación 21	49,46		Replicación 61	49,56
Replicación 22	49,37		Replicación 62	49,75
Replicación 23	49,39		Replicación 63	48,98
Replicación 24	49,23		Replicación 64	49,00
Replicación 25	49,47		Replicación 65	49,35
Replicación 26	49,51		Replicación 66	49,27
Replicación 27	49,35		Replicación 67	49,44
Replicación 28	48,89		Replicación 68	49,08
Replicación 29	49,44		Replicación 69	49,09
Replicación 30	49,34		Replicación 70	49,15
Replicación 31	49,41		Replicación 71	49,29
Replicación 32	49,18		Replicación 72	49,29
Replicación 33	49,50		Replicación 73	49,08
Replicación 34	49,12		Replicación 74	49,25
Replicación 35	49,05		Replicación 75	48,93
Replicación 36	49,40		Replicación 76	49,45
Replicación 37	49,20		Replicación 77	49,13
Replicación 38	49,54		Replicación 78	49,45
Replicación 39	49,32		Replicación 79	49,14
Replicación 40	49,35		Replicación 80	49,27

Existen tres pasos principales para calcular el error típico:

1. Cada estimación replicada se comparará con la estimación final 49,33 y la diferencia se elevará al cuadrado. Matemáticamente, esto corresponde a $\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$ o, en este caso concreto,

$(\hat{\mu}_i - \hat{\mu})^2$. Para la primera replicación, será igual a: (49,44 − 49,33)² = 0,0140. Para la segunda replicación, corresponde a: (49,18 − 49,33)² = 0,0228. La tabla 6.2 presenta las diferencias al cuadrado.

2. Se calcula la suma de las diferencias al cuadrado, $\sum_{i=1}^{80}(\hat{\mu}_{(i)} - \hat{\mu})^2$, y se divide por 20. En el ejemplo, la suma es igual a: (0,0140 + 0,0228 + ... + 0,0354 + 0,0031) = 3,5195.

   La suma dividida por 20 es, por tanto, igual a 3,5195/20 = 0,1760. Este valor representa la varianza muestral de la estimación de la media para HISEI.

3. El error típico es igual a la raíz cuadrada de la varianza muestral, es decir:

$$\sigma_{(\hat{\mu})} = \sqrt{\sigma^2_{(\hat{\mu})}} = \sqrt{0{,}1760} = 0{,}4195$$

Esto significa que la distribución muestral de la media de HISEI para Alemania tiene una desviación típica de 0,4195. Este valor también permite formar un intervalo de confianza alrededor de esta media. Con una probabilidad de error de tipo I igual a 0,05, normalmente llamado α, el intervalo de confianza será igual a:

$$[49{,}33 - (1{,}96 \cdot 0{,}4195) ; 49{,}33 + (1{,}96 \cdot 0{,}4195)]$$

$$[48{,}51 ; 50{,}15]$$

Dicho de otra forma, hay 5 posibilidades entre 100 de que un intervalo formado de esta manera no logre capturar la media de la población. También significa que la media de la población alemana para HISEI es significativamente distinta de un valor de 51, por ejemplo, ya que este número no está incluido en el intervalo de confianza.

El capítulo 9 mostrará cómo este error típico puede utilizarse para realizar comparaciones entre o más países, o entre subpoblaciones dentro de un determinado país.

Tabla 6.2. Diferencias al cuadrado entre las estimaciones replicadas y la estimación final

Peso	Diferencia al cuadrado		Peso	Diferencia al cuadrado
Replicación 1	0,0140		Replicación 41	0,0239
Replicación 2	0,0228		Replicación 42	0,1090
Replicación 3	0,0421		Replicación 43	0,0203
Replicación 4	0,0189		Replicación 44	0,0818
Replicación 5	0,0075		Replicación 45	0,0082
Replicación 6	0,0002		Replicación 46	0,1514
Replicación 7	0,0387		Replicación 47	0,0231
Replicación 8	0,0583		Replicación 48	0,0349
Replicación 9	0,0472		Replicación 49	0,0590
Replicación 10	0,0167		Replicación 50	0,0014
Replicación 11	0,0124		Replicación 51	0,3003
Replicación 12	0,0441		Replicación 52	0,0431
Replicación 13	0,0000		Replicación 53	0,0032
Replicación 14	0,0205		Replicación 54	0,0086
Replicación 15	0,0048		Replicación 55	0,0868
Replicación 16	0,0009		Replicación 56	0,1317
Replicación 17	0,0074		Replicación 57	0,0438
Replicación 18	0,2264		Replicación 58	0,0354
Replicación 19	0,0077		Replicación 59	0,0034
Replicación 20	0,2604		Replicación 60	0,0081
Replicación 21	0,0182		Replicación 61	0,0563
Replicación 22	0,0016		Replicación 62	0,1761
Replicación 23	0,0041		Replicación 63	0,1173
Replicación 24	0,0093		Replicación 64	0,1035
Replicación 25	0,0199		Replicación 65	0,0008
Replicación 26	0,0344		Replicación 66	0,0030
Replicación 27	0,0007		Replicación 67	0,0139
Replicación 28	0,1919		Replicación 68	0,0618
Replicación 29	0,0139		Replicación 69	0,0557
Replicación 30	0,0001		Replicación 70	0,0324
Replicación 31	0,0071		Replicación 71	0,0016
Replicación 32	0,0215		Replicación 72	0,0011
Replicación 33	0,0302		Replicación 73	0,0603
Replicación 34	0,0411		Replicación 74	0,0052
Replicación 35	0,0778		Replicación 75	0,1575
Replicación 36	0,0052		Replicación 76	0,0157
Replicación 37	0,0150		Replicación 77	0,0378
Replicación 38	0,0445		Replicación 78	0,0155
Replicación 39	0,0000		Replicación 79	0,0354
Replicación 40	0,0004		Replicación 80	0,0031
		Suma de las diferencias al cuadrado		3,5195

La macro de SPSS® para calcular el error típico de una media

Escribir toda la sintaxis de SPSS® para calcular estas 81 medias y transferirlas a continuación a una hoja de cálculo de Microsoft® Excel® para obtener finalmente el error típico llevaría mucho tiempo. Por fortuna, las macros de SPSS® simplifican los cálculos iterativos. El programa ejecutará N veces las instrucciones contenidas entre la instrucción de comienzo (`!do !i=1 !to n`) y la instrucción final (`!doend`). Además, también guarda los resultados en un archivo temporal que puede utilizarse a continuación para calcular el error típico.

Se han escrito unas 12 macros de SPSS® para simplificar los principales cálculos de PISA. Estas macros están guardadas en sus correspondientes archivos (con la extensión .sps). El cuadro 6.2 muestra una sintaxis de SPSS® donde se invoca a una macro para calcular la media y el error típico de la variable HISEI.

Cuadro 6.2. Sintaxis de SPSS® para calcular la media de HISEI y su respectivo error típico

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if (cnt='DEU').
Save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
Include file 'C:\PISA\macros\mcr_SE_univ.sps'.

* EJECUTAR LA MACRO.
univar nrep = 80/
       stat = mean/
       dep = hisei/
       grp = cnt/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
```

Después de seleccionar los datos de Alemania a partir de la base de datos de alumnos de PISA 2003 y guardarlos en un archivo de datos temporal, la instrucción

```
Include file 'C:\PISA\macros\mcr_SE_univ.sps'.
```

importará y guardará en memoria, para usarlo más tarde, un nuevo procedimiento para calcular un estadístico univariante y su error típico. Este procedimiento se llama *univar* y se ejecutará cuando se invoque la macro.

Cuando se invoca la macro, el usuario debe suministrar los argumentos que aquella necesita. NREP es el número de replicaciones. STAT es el estadístico que se está calculando. El estadístico se calcula con la instrucción *aggregate*, lo que significa que los estadísticos mostrados en la tabla 6.4 están disponibles para esta macro. DEP es la variable para la que se calcula el estadístico y GRP es la variable de agrupación. En este ejemplo, la variable de agrupación «país» (CNT) es una constante (ya que el archivo de datos sólo contiene datos de Alemania, se calculará un solo estadístico y un solo error típico). WGT es el peso para alumnos completo y RWGT es la raíz de los pesos replicados (la macro concatenará esta raíz con los números de 1 a 80: de W_FSTR1 a W_FSTR80 en este caso). CONS es la constante que se usa al calcular la varianza muestral. Esta constante es $\frac{1}{G(1-k)^2}$, donde G es el número de replicaciones y k es el factor de Fay (PISA usa 0,5; véase el capítulo 3 y el comienzo de este capítulo). INFILE es el archivo de datos sobre el que se ejecutará el procedimiento.

Como estos cálculos iterativos podrían un tiempo de ejecución importante, el procedimiento sólo utilizará las variables necesarias.

A partir del archivo temporal de datos, esta macro calculará para cada país la media de HISEI y su error típico usando el peso para alumnos final (W_FSTUWT) y los 80 pesos replicados (de W_FSTR1 a W_FSTR80). Terminará entregando exactamente los mismos valores para la estimación de la media y su respectivo error típico que los obtenidos mediante las tablas 6.1 y 6.2.

La estructura del archivo de salida se presenta en la tabla 6.3.

Tabla 6.3. Estructura del archivo de salida del cuadro 6.2

CNT	STAT	SE
DEU	49,33	0,42

[donde CNT = país, STAT = valor del estadístico y SE = valor del error típico.]

Si el conjunto de datos no se hubiese reducido a los datos de Alemania, el número de filas en el archivo de salida sería igual al número de países en la base de datos.

Existen algunas restricciones, así como algunas opciones, con esta macro:

- sólo puede especificarse un único archivo de datos como entrada;
- pueden especificarse diversas variables de agrupación; por ejemplo, si se necesitan los resultados según el género, las variables de grupo serán CNT y ST03Q01, es decir `grp = cnt st03q01/;`
- sólo puede especificarse una variable numérica en el argumento STAT;
- la macro no guardará el archivo de salida.

Tabla 6.4. Estadísticos disponibles con la macro UNIVAR[a]

Estadísticos disponibles	Significado
SUM	(Suma)
MEAN	(Media)
SD	(Desv. típica)
PGT	(% casos > un valor)
PLT	(% casos < un valor)
PIN	(% dentro de valores)
POUT	(% fuera de valores)
FGT	(fracción > un valor)
FLT	(fracción < un valor)
FIN	(fracción dentro de valores)
FOUT	(fracción fuera de valores)

El cuadro 6.3 presenta la sintaxis para el cálculo de la desviación típica de HISEI según el género del alumnado, y la tabla 6.5, la estructura del archivo de salida.

[a] También se encuentran otros estadísticos disponibles a través de la función *aggregate* de SPSS®, como el mínimo, el máximo, el primero, el último, el número de observaciones, etcétera. Sin embargo, no se incluyen en la tabla, o bien porque no tiene sentido aplicar estos estadísticos a los datos de PISA, o bien porque el método de Fay no puede aplicarse a estos estadísticos. Por ejemplo, puesto que el conjunto de observaciones se utiliza en cada replicación, el valor mínimo o máximo para una determinada variable siempre será el mismo. Por tanto, la macro estimará un error típico de 0, lo que naturalmente carece de sentido.

Cuadro 6.3. Sintaxis de SPSS® para el cálculo de la desviación típica de HISEI y su respectivo error típico según el género del alumnado

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if cnt='DEU'.
* Seleccionar los datos que no están perdidos es opcional.
Select if (not missing(st03q01)).
Save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
Include file 'C:\PISA\macros\mcr_SE_univ.sps'.

* EJECUTAR LA MACRO.
univar nrep = 80/
       stat = sd/
       dep = hisei/
       grp = cnt st03q01/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
```

Tabla 6.5. Estructura del archivo de salida del cuadro 6.3

CNT	ST03Q01	STAT	SE
DEU	1	16,12	0,29
DEU	2	16,34	0,23

El error típico de los porcentajes

Para variables como el género, el estadístico de interés suele ser el porcentaje por categoría. El procedimiento para estimar el error típico es idéntico al procedimiento utilizado para estimar el error típico de una media o una desviación típica, es decir, por cada categoría de la variable es necesario calcular 81 porcentajes.

El cuadro 6.4 presenta la sintaxis de SPSS® para ejecutar la macro que calculará los porcentajes y sus respectivos errores típicos para cada categoría de la variable de género. La estructura del archivo de salida se presenta en la tabla 6.6.

Cuadro 6.4. Sintaxis de SPSS® para el cálculo de los porcentajes y su respectivo error típico para el género del alumnado

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if cnt='DEU'.
* Seleccionar los datos que no están perdidos es opcional.
Select if (not missing(st03q01)).
Save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file "C:\PISA\macros\mcr_SE_GrpPct.sps".

* EJECUTAR LA MACRO.
GRPPCT nrep = 80/
       within = cnt/
       grp = st03q01/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'/.
```

Tabla 6.6. Estructura del archivo de salida a partir del cuadro 6.4

CNT	ST03Q01	STAT	SE
DEU	1	49,66	1,04
DEU	2	50,34	1,04

La tabla 6.7 presenta las estimaciones del porcentaje de chicas para los 81 pesos y las diferencias al cuadrado. El porcentaje de chicas que se obtendrá es igual a 49,66, es decir, el porcentaje obtenido con el peso de alumnos final.

Como anteriormente, hay tres pasos principales para calcular el error típico:

1. Cada estimación replicada se comparará con la estimación final 49,66 y la diferencia se elevará al cuadrado. Matemáticamente, esto corresponde a $\left(\hat{\pi}_{(i)} - \hat{\pi}\right)^2$. Para la primera replicación, será igual a: $(49{,}82 - 49{,}66)^2 = 0{,}0256$.

2. Se calcula la suma de las diferencias al cuadrado, $\sum_{i=1}^{80}\left(\hat{\pi}_{(i)} - \hat{\pi}\right)^2$, y se divide por 20. En el ejemplo, la suma es igual a: (0,0252 + 0,1044 + ... + 0,3610 + 0,1313) = 21,4412.

   La suma dividida por 20 es, por tanto, igual a 21,4412 / 20 = 1,07206. Este valor representa la varianza muestral de la estimación del porcentaje para las chicas.

3. El error típico es igual a la raíz cuadrada de la varianza muestral, es decir:

$$\sigma_{(\hat{\pi})} = \sqrt{\sigma^2_{(\hat{\pi})}} = \sqrt{1{,}07206} = 1{,}035$$

El mismo proceso puede utilizarse para el porcentaje de los chicos. Debería advertirse que el error típico de los chicos es igual al de las chicas. Es más, puede demostrarse matemáticamente

que el error típico de π es igual al error típico de $1 - \pi$, es decir, $\sigma_{(p)} = \sigma_{(1-p)}$. Sin embargo, si se mantienen en el archivo de datos los datos perdidos para el género, el error típico del porcentaje de los chicos puede variar ligeramente del error típico del porcentaje de las chicas.

Al igual que con la macro para variables numéricas, puede utilizarse más de una variable de agrupación. En PISA 2003, la primera pregunta del cuestionario del alumnado (ST01Q01) informa del curso al que asisten los alumnos. Los quinceañeros alemanes se distribuyen entre los cursos del 7 al 11.

El cuadro 6.5 presenta la sintaxis de SPSS® y la tabla 6.8 presenta la distribución de alumnos según el curso y el género. Los porcentajes dentro de la variable de grupo WITHIN suman 100%. En este ejemplo, los porcentajes de los alumnos de los cursos 7 a 11 dentro del género y país suman un 100%. Si `within=cnt` y `grp=st03q01 st01q01`, la suma de los porcentajes de los diez grupos dentro del país será 100%.

Tabla 6.7. Porcentaje de chicas para los pesos final y replicados y las diferencias al cuadrado

Peso	Estimación del porcentaje	Diferencia al cuadrado		Peso	Estimación del porcentaje	Diferencia al cuadrado
Peso final	**49,66**					
Replicación 1	49,82	0,03		Replicación 41	50,00	0,11
Replicación 2	49,98	0,10		Replicación 42	49,95	0,09
Replicación 3	49,44	0,05		Replicación 43	49,70	0,00
Replicación 4	49,32	0,11		Replicación 44	50,59	0,87
Replicación 5	49,39	0,07		Replicación 45	49,07	0,35
Replicación 6	49,06	0,36		Replicación 46	48,82	0,71
Replicación 7	48,59	1,14		Replicación 47	49,88	0,05
Replicación 8	48,85	0,66		Replicación 48	49,14	0,27
Replicación 9	49,06	0,36		Replicación 49	49,53	0,02
Replicación 10	49,72	0,00		Replicación 50	49,81	0,02
Replicación 11	50,05	0,15		Replicación 51	49,87	0,04
Replicación 12	49,31	0,13		Replicación 52	49,82	0,02
Replicación 13	49,29	0,13		Replicación 53	49,42	0,06
Replicación 14	49,47	0,04		Replicación 54	48,99	0,45
Replicación 15	49,90	0,06		Replicación 55	50,07	0,17
Replicación 16	50,82	1,35		Replicación 56	50,68	1,04
Replicación 17	49,11	0,30		Replicación 57	50,34	0,46
Replicación 18	49,51	0,02		Replicación 58	49,54	0,02
Replicación 19	49,79	0,02		Replicación 59	48,75	0,83
Replicación 20	50,75	1,18		Replicación 60	50,14	0,23
Replicación 21	50,24	0,33		Replicación 61	49,45	0,05
Replicación 22	49,79	0,02		Replicación 62	49,46	0,04
Replicación 23	49,87	0,04		Replicación 63	50,11	0,20
Replicación 24	49,37	0,08		Replicación 64	49,64	0,00
Replicación 25	49,50	0,02		Replicación 65	49,72	0,00
Replicación 26	49,82	0,02		Replicación 66	50,79	1,27
Replicación 27	49,92	0,07		Replicación 67	49,73	0,00
Replicación 28	49,55	0,01		Replicación 68	49,96	0,09
Replicación 29	50,22	0,31		Replicación 69	50,31	0,42
Replicación 30	49,16	0,25		Replicación 70	49,17	0,24
Replicación 31	50,51	0,73		Replicación 71	50,10	0,19
Replicación 32	49,98	0,10		Replicación 72	49,93	0,07
Replicación 33	50,67	1,02		Replicación 73	49,55	0,01

Replicación 34	49,29	0,13		Replicación 74	49,42	0,06
Replicación 35	48,96	0,49		Replicación 75	49,60	0,00
Replicación 36	49,98	0,10		Replicación 76	49,45	0,05
Replicación 37	50,23	0,33		Replicación 77	49,80	0,02
Replicación 38	48,25	1,99		Replicación 78	49,91	0,07
Replicación 39	49,56	0,01		Replicación 79	49,06	0,36
Replicación 40	49,66	0,00		Replicación 80	50,02	0,13
				Suma de las diferencias al cuadrado		21,44

Cuadro 6.5. Sintaxis de SPSS® para el cálculo de porcentajes de curso según el sexo

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if cnt='DEU'.
Select if (not missing(st03q01) & not missing(st01q01)).
Save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file "C:\PISA\macros\mcr_SE_GrpPct.sps".

* EJECUTAR LA MACRO.
GRPPCT nrep = 80/
       within = cnt st03q01/
       grp = st01q01/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'C:\PISA\Data2003\DEU.sav'/.
```

Como se muestra en la tabla 6.8, en los cursos inferiores de Alemania tiende a haber más chicos que chicas, y en los superiores, más chicas que chicos.

Tabla 6.8. Estructura del archivo de salida a partir del cuadro 6.5

CNT	ST03Q01	ST01Q01	STAT	SE
DEU	1	7	1,15	0,26
DEU	1	8	13,09	0,83
DEU	1	9	59,33	1,00
DEU	1	10	26,28	1,08
DEU	1	11	0,17	0,08
DEU	2	7	2,28	0,45
DEU	2	8	16,92	1,04
DEU	2	9	60,32	1,06
DEU	2	10	20,41	0,79
DEU	2	11	0,08	0,05

El error típico de los coeficientes de regresión

Para cualquier estadístico buscado, el cálculo de la estimación y su error típico seguirá exactamente el mismo procedimiento que los que se han descrito para la media de HISEI y para el porcentaje de chicas. El resto de este capítulo explicará el uso de otras dos macros de SPSS® desarrolladas para analizar los datos de PISA.

La primera macro es para los análisis de regresión lineal simple. Además de los argumentos comunes a todas las macros de SPSS® descritas en este manual, es decir, 1) NREP=, 2) GRP=, 3) W_FSTUWT=, 4) W_FSTR=, 5) CONS= y 6) INFILE=, es preciso especificar dos argumentos: la variable dependiente y las variables independientes. Sólo puede especificarse una única variable dependiente, aunque sí diversas variables independientes.

El cuadro 6.6 proporciona la sintaxis para ejecutar la macro de regresión lineal simple. En este ejemplo, la variable dependiente es la profesión esperada del alumno cuando tenga 30 años (BSMJ) y las variables independientes son el índice socioeconómico familiar (HISEI) y el género del alumno después de la recodificación (GENDER). Después de recodificar la variable de género en una variable dicotómica 0-1, la macro se importa mediante la instrucción «`include`».

Cuadro 6.6. Sintaxis de SPSS® para los análisis de regresión (1)

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU' & not missing(st03q01)).
compute gender=0.
if (st03q01=1) gender=1.
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file='C:\PISA\macros\mcr_SE_reg.sps'.

* EJECUTAR LA MACRO.
REGnoPV nrep = 80/
        ind = hisei gender/
        dep = bsmj/
        grp = cnt/
        wgt = w_fstuwt/
        rwgt = w_fstr/
        cons = 0.05/
        infile = 'c:\PISA\Data2003\DEU.sav'/.
```

La tabla 6.9 presenta la estructura del archivo de salida del análisis de regresión.[1]

Tabla 6.9. Estructura del archivo de resultados a partir del cuadro 6.6

CNT	IND	STAT	SE
DEU	b_0	32,90	1,29
DEU	HISEI	0,37	0,03
DEU	gender	2,07	0,62

Donde b_0 es la constante (o intercepto) e HISEI y GENDER, los coeficientes de regresión (o pendientes) de las correspondientes variables. Para reestructurar los datos, de modo que cada país (grupo) ocupe un solo registro o fila en el archivo de salida, pueden ejecutarse las instrucciones del cuadro 6.7 (una de las instrucciones no está disponible en las versiones de SPSS® anteriores a la versión 11). La estructura del archivo de salida se muestra en la tabla 6.10.

Cuadro 6.7. Sintaxis de SPSS® para reestructurar los datos después del análisis de regresión

```
recode ind ('HISEI'='b1')('gender'='b2').
sort cases by cnt ind.
Casestovars /id=cnt /index=ind /groupby=index.
```

Tabla 6.10. Reestructuración del archivo de salida

CNT	STAT.B0	SE.B0	STAT.B1	SE.B1	STAT.B2	SE.B2
DEU	32,90	1,29	0,37	0,03	2,07	0,62

Otras macros más complejas, como esta para calcular el error típico para coeficientes de regresión, a veces provocan errores que no siempre son visibles. Estos errores ocurren especialmente cuando se ejecuta la misma macro más de una vez dentro de la misma sesión de SPSS®. Merece la pena comprobar siempre los resultados de la macro ejecutando el análisis fuera de la macro con el peso total de alumnos y comprobar si los coeficientes de regresión (o cualquier otro estadístico) son correctos. Si no son los mismos, lo mejor es cerrar SPSS® y, a menudo, incluso salir y volver a entrar en el sistema operativo (Windows). Asimismo, proceder a borrar todos los archivos situados en la carpeta 'C:\Temp\' soluciona a veces el problema.

Existen dos maneras de determinar si los coeficientes de regresión son significativamente distintos de 0. El primer método consiste en construir un intervalo de confianza alrededor del coeficiente de regresión estimado. El intervalo de confianza para el coeficiente de regresión GENDER de BSMJ puede calcularse para un valor de α igual a 0,05 como:

$$[2{,}07 - (1{,}96 \cdot 0{,}62)\ ;\ 2{,}07 + (1{,}96 \cdot 0{,}62)] = [0{,}85\ ;\ 3{,}29]$$

Como el valor 0 no se incluye en este intervalo de confianza, el coeficiente de regresión es significativamente distinto de 0. Puesto que el valor 0 se asignó a los chicos y el valor 1 a las chicas, esto quiere decir que, de media, las chicas tienen expectativas profesionales significativamente más altas.

Otra forma de comprobar la hipótesis nula del coeficiente de regresión consiste en dividir el coeficiente de regresión por su error típico. Este procedimiento tipifica el coeficiente de regresión. También significa que la distribución muestral del coeficiente de regresión tipificado, bajo la hipótesis nula, tiene una media esperada de 0 y una desviación típica de 1. Por lo tanto, si la razón entre el coeficiente de regresión y su error típico es menor que −1,96 o mayor que 1,96, se considerará significativamente distinta de 0.

Es preciso mencionar que la sintaxis del cuadro 6.8 proporcionará distintos resultados que la del cuadro 6.6. En este último, GENDER se considera una variable explicativa, mientras que en el cuadro 6.8 GENDER se utiliza como variable de agrupación. En el segundo modelo, sólo hay una variable explicativa, esto es, HISEI.

Cuadro 6.8. Sintaxis de SPSS® para los análisis de regresión (2)

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU' & not missing(st03q01)).
compute gender=0.
if (st03q01=1) gender=1.
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file='C:\PISA\macros\mcr_SE_reg.sps'.

* EJECUTAR LA MACRO.
REGnoPV nrep = 80/
        ind = hisei/
        dep = bsmj/
        grp = cnt gender/
        wgt = w_fstuwt/
        rwgt = w_fstr/
        cons = 0.05/
        infile = 'c:\PISA\Data2003\DEU.sav'/.
```

La tabla 6.11 presenta la estructura del archivo de salida para el segundo modelo.

Tabla 6.11. Estructura del archivo de salida a partir del cuadro 6.8

CNT	GENDER	IND	STAT	SE
DEU	0	b_0	32,54	1,44
DEU	0	HISEI	0,37	0,03
DEU	1	b_0	35,33	1,66
DEU	1	HISEI	0,36	0,03

El error típico de los coeficientes de correlación

La tabla 6.12 y el cuadro 6.9 presentan, respectivamente, la sintaxis de SPSS® y la estructura del archivo de salida para la macro dedicada al cálculo de una correlación entre dos variables. Adviértase que la macro borra del archivo de datos los casos con valores perdidos para alguna de las dos variables, (NOTA: la versión española del SPSS lo denomina exclusión de casos según lista).

Cuadro 6.9. Sintaxis de SPSS® para la macro de correlación

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU').
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file "c:\pisa\macros\mcr_SE_cor.sps".

* EJECUTAR LA MACRO.
CORnoPV  nrep = 80/
         var1 = hisei/
         var2 = bsmj/
         grp = cnt/
         wgt = w_fstuwt/
         rwgt = w_fstr/
         cons = 0.05/
         infile = 'c:\PISA\Data2003\DEU.sav'/.
```

Tabla 6.12. Estructura del archivo de salida a partir del cuadro 6.9

CNT	STAT	SE
DEU	0,34	0,02

Conclusiones

Este capítulo ha descrito el cálculo del error típico usando los 80 pesos replicados. El procedimiento es el mismo para cualquier otro estadístico.

Además, al mismo tiempo que se ha utilizado ejemplos, se han proporcionado las sintaxis para ejecutar las macros de SPSS®, desarrolladas con objeto de facilitar el cálculo de los errores típicos.

Sin embargo, ninguna de las macros descritas en este capítulo puede ser utilizada si se incluyen valores plausibles en los análisis. El capítulo 7 describirá cómo proceder con dichas variables.

[1] SPSS® produce muchas tablas en el visor de salida cuando se replica 80 veces la regresión. Esto ralentiza el ordenador e impide que esté libre para otras tareas mientras se ejecuta la macro. SPSS® 12 ha introducido un sistema de gestión de salida *(output management system, OMS)* que proporciona al usuario, entre otros controles de la salida, la oportunidad de elegir lo que imprimirá. Añadiendo las instrucciones «`OMS/select tables warnings headings/destination viewer=no.`» antes de ejecutar la macro se evitará que SPSS® imprima encabezamientos, tablas, anotaciones y advertencias en el visor de salida. Esto quedará activo hasta que se desactive el OMS mediante la instrucción OMSEND.

Capítulo 7

Los análisis con valores plausibles

Introducción

El área de evaluación principal en PISA 2003 fue la competencia matemática, mientras que la lectura, las ciencias y la solución de problemas fueron áreas secundarias. Se creó una sola escala para cada área secundaria, mientras que se generaron cinco para la evaluación de las matemáticas: una escala de matemáticas y cuatro subescalas (espacio y forma, cambio y relaciones, cantidad, incertidumbre).

Como se describió en el capítulo 5, estos datos cognitivos se escalaron con el modelo de Rasch y el rendimiento de los alumnos se expresó mediante valores plausibles. Para cada escala y subescala, se incluyeron cinco valores plausibles por alumno en las bases de datos internacionales. Este capítulo describe cómo llevar a cabo análisis con valores plausibles (PV).

Puesto que los valores plausibles se utilizaron principalmente para comunicar el rendimiento de los alumnos en la prueba cognitiva, este capítulo sólo será útil cuando se realicen análisis sobre datos de rendimiento y sus relaciones con características de alumnos o centros.

Estadísticos univariantes a partir de valores plausibles

El cálculo de un estadístico a partir de valores plausibles consistirá siempre en seis pasos, sea cual sea el estadístico considerado.

1. El estadístico buscado y su error típico respectivo deben calcularse para cada valor plausible. En el capítulo 6, se mencionó que fueron necesarias 81 replicaciones para obtener la estimación final y su error típico. Por lo tanto, cualquier análisis que involucre cinco valores plausibles exigirá 405 estimaciones. Si es preciso calcular una media, entonces se calcularán 405 medias. Las medias estimadas con el peso final se llaman $\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4$ y $\hat{\mu}_5$. A partir de las 80 replicaciones aplicadas a cada uno de los cinco valores plausibles, se estiman cinco varianzas muestrales, llamadas respectivamente $\sigma^2_{(\hat{\mu}_1)}, \sigma^2_{(\hat{\mu}_2)}, \sigma^2_{(\hat{\mu}_3)}, \sigma^2_{(\hat{\mu}_4)}$ y $\sigma^2_{(\hat{\mu}_5)}$. Estas cinco estimaciones de la media y sus respectivas varianzas muestrales se dan en la tabla 7.1.
2. La estimación de la media final es igual al promedio de las cinco estimaciones de la media, es decir,

$$\hat{\mu} = \frac{1}{5}\left(\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3 + \hat{\mu}_4 + \hat{\mu}_5\right)$$

3. La varianza muestral final es igual al promedio de las cinco varianzas muestrales, es decir,

$$\sigma^2_{(\hat{\mu})} = \frac{1}{5}\left(\sigma^2_{(\hat{\mu}_1)} + \sigma^2_{(\hat{\mu}_2)} + \sigma^2_{(\hat{\mu}_3)} + \sigma^2_{(\hat{\mu}_4)} + \sigma^2_{(\hat{\mu}_5)}\right)$$

4. La varianza de imputación, también llamada *varianza del error de medida*, se calcula como $\sigma^2_{(prueba)} = \frac{1}{4}\sum_{i=1}^{5}\left(\hat{\mu}_i - \hat{\mu}\right)^2$. De hecho, como PISA devuelve cinco valores plausibles por es-

cala, $\sigma^2_{(prueba)} = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\mu}_i - \hat{\mu})^2 = \frac{1}{4}\sum_{i\_1}^{5}(\hat{\mu}_i - \hat{\mu})^2$. Esta fórmula es similar a la utilizada para estimar la varianza poblacional, sólo que en este caso concreto, las observaciones no se comparan con la media de la población, sino que cada media PV se compara con la estimación de la media final.

5. La varianza muestral y la varianza de imputación se combinan para obtener la varianza del error final como $\sigma^2_{(error)} = \sigma^2_{(\hat{\mu})} + \left(1{,}2\sigma^2_{(prueba)}\right)$.

   De hecho,

$$\sigma^2_{(error)} = \sigma^2_{(\hat{\mu})} + \left(\left(1+\frac{1}{M}\right)\sigma^2_{(prueba)}\right) = \sigma^2_{(\hat{\mu})} + \left(\left(1+\frac{1}{5}\right)\sigma^2_{(prueba)}\right) = \sigma^2_{(\hat{\mu})} + \left(1{,}2\sigma^2_{(prueba)}\right).$$

6. El error típico es igual a la raíz cuadrada de la varianza del error.

Tabla 7.1. Las 405 estimaciones de la media

Peso	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
Replicación 1	$\hat{\mu}_{1\_1}$	$\hat{\mu}_{2\_1}$	$\hat{\mu}_{3\_1}$	$\hat{\mu}_{4\_1}$	$\hat{\mu}_{5\_1}$
Replicación 2	$\hat{\mu}_{1\_2}$	$\hat{\mu}_{2\_2}$	$\hat{\mu}_{3\_2}$	$\hat{\mu}_{4\_2}$	$\hat{\mu}_{5\_2}$
Replicación 3	$\hat{\mu}_{1\_3}$	$\hat{\mu}_{2\_3}$	$\hat{\mu}_{3\_3}$	$\hat{\mu}_{4\_3}$	$\hat{\mu}_{5\_3}$
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
Replicación 80	$\hat{\mu}_{1\_80}$	$\hat{\mu}_{2\_80}$	$\hat{\mu}_{3\_80}$	$\hat{\mu}_{4\_80}$	$\hat{\mu}_{5\_80}$
Varianza muestral	$\sigma^2_{(\hat{\mu}_1)}$	$\sigma^2_{(\hat{\mu}_2)}$	$\sigma^2_{(\hat{\mu}_3)}$	$\sigma^2_{(\hat{\mu}_4)}$	$\sigma^2_{(\hat{\mu}_5)}$

Puede calcularse la estimación de la media en la escala de matemáticas y su correspondiente error típico relativos a los datos de Alemania para PISA 2003. La macro descrita en el capítulo 6 y guardada en el archivo `mcr_se_univ.sps` puede ser utilizada sucesivamente cinco veces y los resultados ser combinados en una hoja de cálculo de Microsoft® Excel®. La tabla 7.2 presenta las distintas medias de PV y sus respectivas varianzas muestrales, así como las estimaciones de las medias en las replicaciones primera y última.

Tabla 7.2. Estimaciones de las medias y sus respectivas varianzas muestrales en la escala de matemáticas para Alemania

Peso	PV1	PV2	PV3	PV4	PV5
Final	503,08	503,10	502,72	503,03	503,00
Replicación 1	503,58	504,16	503,43	503,96	503,94
.....	.....	.....	.....	.....	.....
Replicación 80	503,18	503,62	503,46	503,30	503,83
Varianza muestral	$(3{,}34)^2$	$(3{,}27)^2$	$(3{,}36)^2$	$(3{,}28)^2$	$(3{,}32)^2$

El cuadro 7.1 presenta la sintaxis SPSS® para invocar y ejecutar sucesivamente la macro `mcr_se_univ.sps` descrita en el capítulo 6.

La estimación de la media final para Alemania en la escala de matemáticas de PISA 2003 es igual a (μ_1 = stat1, etcétera, en el archivo de datos):

$$\hat{\mu} = \frac{1}{5}\left(\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3 + \hat{\mu}_4 + \hat{\mu}_5\right), \text{ es decir:}$$

$$\hat{\mu} = \frac{\left(503{,}08 + 503{,}10 + 502{,}72 + 503{,}03 + 503{,}00\right)}{5} = 502{,}99$$

La varianza muestral final de la estimación de la media para la escala de competencia matemática es igual a ($\sigma_{(\mu_1)}$ = se1, etcétera, en el archivo de datos):

$$\sigma^2_{(\hat{\mu})} = \frac{1}{5}\left(\sigma^2_{(\hat{\mu}_1)} + \sigma^2_{(\hat{\mu}_2)} + \sigma^2_{(\hat{\mu}_3)} + \sigma^2_{(\hat{\mu}_4)} + \sigma^2_{(\hat{\mu}_5)}\right), \text{ es decir:}$$

$$\sigma^2_{(\hat{\mu})} = \frac{\left(3{,}34\right)^2 + \left(3{,}27\right)^2 + \left(3{,}36\right)^2 + \left(3{,}28\right)^2 + \left(3{,}32\right)^2}{5} = 10{,}98$$

La varianza de imputación es igual a:

$$\sigma^2_{(prueba)} = \frac{1}{4}\sum_{i=1}^{5}\left(\hat{\mu}_i - \hat{\mu}\right)^2 \text{ , es decir:}$$

$$\sigma^2_{(prueba)} = \frac{\left[\left(503{,}08 - 502{,}99\right)^2 + \left(503{,}10 - 502{,}99\right)^2 + \cdots + \left(503{,}00 - 502{,}99\right)^2\right]}{4} = \frac{0{,}09}{4} = 0{,}02$$

La varianza de error final es igual a:

$$\sigma^2_{(error)} = \sigma^2_{(\hat{\mu})} + \left(1{,}2\sigma^2_{(prueba)}\right), \text{ es decir:}$$

$$\sigma^2_{(error)} = 10{,}98 + \left(1{,}2 \cdot 0{,}02\right) = 11{,}00$$

El error típico final es, por tanto, igual a:

$$SE = \sqrt{\sigma^2_{(error)}} = \sqrt{11{,}00} = 3{,}32$$

Cuadro 7.1. Sintaxis de SPSS® para calcular la media en la escala de matemáticas

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if (cnt='DEU').
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file "c:\pisa\macros\mcr_SE_univ.sps".

* EJECUTAR LA MACRO 5 VECES.
univar nrep = 80/ stat = mean/ dep = pv1math/ grp = cnt/
       wgt = w_fstuwt/ rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
rename vars (stat se=stat1 se1).
save outfile='c:\temp\ex1.sav' /drop=var.
univar nrep = 80/ stat = mean/ dep = pv2math/ grp = cnt/
       wgt = w_fstuwt/ rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
rename vars (stat se=stat2 se2).
save outfile='c:\temp\ex2.sav' /drop=var.
univar nrep = 80/ stat = mean/ dep = pv3math/ grp = cnt/
       wgt = w_fstuwt/ rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
rename vars (stat se=stat3 se3).
save outfile='c:\temp\ex3.sav' /drop=var.
univar nrep = 80/ stat = mean/ dep = pv4math/ grp = cnt/
       wgt = w_fstuwt/ rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
rename vars (stat se=stat4 se4).
save outfile='c:\temp\ex4.sav' /drop=var.
univar nrep = 80/ stat = mean/ dep = pv5math/ grp = cnt/
       wgt = w_fstuwt/ rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
rename vars (stat se=stat5 se5).
save outfile='c:\temp\ex5.sav' /drop=var.

match files file='c:\temp\ex1.sav' /file='c:\temp\ex2.sav'
 /file='c:\temp\ex3.sav' /file='c:\temp\ex4.sav'
 /file='c:\temp\ex5.sav' /by cnt.
exe.
```

Puede evitarse ejecutar sucesivamente cinco veces de la macro UNIVAR y la posterior combinación de resultados: se ha desarrollado una macro de SPSS® para tratar con los valores plausibles. Esta macro también calcula:

- las cinco estimaciones de la media (STAT 1 a STAT5);
- la estimación final (STAT);
- las cinco varianzas muestrales (VAR 1 a VAR 5);
- la media de las cinco varianzas muestrales (PV_VAR);
- la varianza de imputación (PVMERR);
- el error típico final (SE) combinando la varianza muestral final y la varianza de imputación.

Cuadro 7.2. Sintaxis de SPSS® para calcular simultáneamente la media de los valores plausibles y su error típico

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if (cnt='DEU').
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file 'c:\pisa\macros\mcr_SE_pv.sps'.

* EJECUTAR MACRO.
PV  nrep = 80/
    stat = mean/
    dep = math/
    grp = cnt/
    wgt = w_fstuwt/
    rwgt = w_fstr/
    cons = 0.05/
    infile = 'c:\PISA\Data2003\DEU.sav'/.
```

Los argumentos son idénticos a los argumentos de la macro para estadísticos univariantes sin valores plausibles, descritos en el capítulo 6. La diferencia es la descripción del argumento «DEP». Sólo se necesita la raíz de la variable (la dimensión); la macro antepondrá a la raíz los prefijos «pv1» a «pv5». Es decir, cuando «DEP=READ», la macro utilizará las variables pv1read a pv5read. Cuando «DEP=MATH1», la macro utilizará pv1math1 a pv5math1 y, por tanto, calculará los estadísticos para la primera subescala de matemáticas.

La estructura del archivo de salida se presenta en la tabla 7.3.

Tabla 7.3. Estructura del archivo de salida a partir del cuadro 7.2

CNT	STAT	SE
DEU	502,99	3,32

De modo similar a las macros de SPSS® descritas en el capítulo anterior, puede usarse más de una variable de agrupación. Por ejemplo, si se quiere determinar si la dispersión del rendimiento en matemáticas es mayor para las chicas que para los chicos, puede usarse la macro PV como a continuación:

Cuadro 7.3. Sintaxis de SPSS® para calcular la desviación típica y su error típico en valores plausibles según el género

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if (cnt='DEU' & not missing(st03q01)).
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file 'c:\pisa\macros\mcr_SE_pv.sps'.

* EJECUTAR LA MACRO.
PV  nrep = 80/
    stat = sd/
    dep = math /
    grp = cnt st03q01/
    wgt = w_fstuwt/
    rwgt = w_fstr/
    cons = 0.05/
    infile = 'c:\PISA\Data2003\DEU.sav'/.
```

La estructura del archivo de salida se presenta en la tabla 7.4.

Tabla 7.4. Estructura del archivo de salida a partir del cuadro 7.3

CNT	ST03Q01	STAT	SE
DEU	1	99,29	2,05
DEU	2	105,05	2,54

Según la tabla 7.4, la desviación típica («STAT», estadístico) es mayor para los chicos que para las chicas. Por desgracia, como se explicará en el capítulo 10, estos dos errores típicos («SE», *standard error*) no pueden utilizarse para contrastar la igualdad de los dos valores de la desviación típica, ya que las estimaciones de la desviación típica para los chicos y para las chicas pueden estar correlacionadas.

El error típico de los porcentajes con valores plausibles

La segunda macro, presentada por primera vez en el capítulo 6, se desarrolló para el cálculo de porcentajes y sus respectivos errores típicos. El capítulo 8 tratará de la aplicación de esta macro a valores plausibles; es necesario dedicar un capítulo entero a este tipo de análisis debido a las cuestiones implicadas.

El error típico de los coeficientes de regresión con valores plausibles

Supongamos que es necesario estimar el efecto estadístico del género y el contexto socioeconómico del alumno sobre el rendimiento en matemáticas. Igual que para estimar una media, esta cuestión puede solucionarse aplicando sucesivamente cinco veces la macro REGNOPV descrita en el capítulo 6.

El cuadro 7.4. presenta una sintaxis de SPSS® para un enfoque así.

Cuadro 7.4. Sintaxis de SPSS® para calcular los coeficientes de regresión de los valores plausibles y sus errores típicos

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU' & not missing(st03q01) & not missing(HISEI)).
compute gender=0.
if (st03q01=1) gender=1.
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file='C:\PISA\macros\mcr_SE_reg.sps'.

* EJECUTAR LA MACRO 5 VECES.
REGnoPV nrep = 80/ ind = hisei gender/ dep = pv1math/
        grp = cnt/ wgt = w_fstuwt/ rwgt = w_fstr/
        cons = 0.05/ infile = 'c:\PISA\Data2003\DEU.sav'/.
RENAME VARS (STAT SE=STAT1 SE1).
save outfile='c:\temp\temp1.sav' /drop=var.
REGnoPV nrep = 80/ ind = hisei gender/ dep = pv2math/
        grp = cnt/ wgt = w_fstuwt/ rwgt = w_fstr/
        cons = 0.05/ infile = 'c:\PISA\Data2003\DEU.sav'/.
RENAME VARS (STAT SE=STAT2 SE2).
save outfile='c:\temp\temp2.sav' /drop=var.
REGnoPV nrep = 80/ ind = hisei gender/ dep = pv3math/
        grp = cnt/ wgt = w_fstuwt/ rwgt = w_fstr/
        cons = 0.05/ infile = 'c:\PISA\Data2003\DEU.sav'/.
RENAME VARS (STAT SE=STAT3 SE3).
save outfile='c:\temp\temp3.sav' /drop=var.
REGnoPV nrep = 80/ ind = hisei gender/ dep = pv4math/
        grp = cnt/ wgt = w_fstuwt/ rwgt = w_fstr/
        cons = 0.05/ infile = 'c:\PISA\Data2003\DEU.sav'/.
RENAME VARS (STAT SE=STAT4 SE4).
save outfile='c:\temp\temp4.sav' /drop=var.
REGnoPV nrep = 80/ ind = hisei gender/ dep = pv5math/
        grp = cnt/ wgt = w_fstuwt/ rwgt = w_fstr/
        cons = 0.05/ infile = 'c:\PISA\Data2003\DEU.sav'/.
RENAME VARS (STAT SE=STAT5 SE5).
save outfile='c:\temp\temp5.sav' /drop=var.

MATCH FILES file='c:\temp\temp1.sav' /file='c:\temp\temp2.sav'
 file='c:\temp\temp3.sav' file='c:\temp\temp4.sav' file='c:\temp\temp5.sav'
 by CNT ind.
exe.
```

Al igual que en el cálculo de una media y su error típico, el cálculo de los coeficientes de regresión y sus respectivos errores típicos consistirá en seis pasos:

1. Para cada valor plausible y para cada variable explicativa, se calculan los coeficientes de regresión con el peso final y los 80 pesos replicados. Se calcularán 405 coeficientes de regresión por cada variable explicativa. La macro de SPSS® REGNOPV, aplicada secuencialmente cinco veces, producirá por cada variable explicativa cinco estimaciones, llamadas $\hat{\beta}_1,\ldots,\hat{\beta}_5$, y cinco errores típicos, llamados $\sigma_{(\hat{\beta}_1)},\ldots,\sigma_{(\hat{\beta}_5)}$. La tabla 7.5 muestra la expresión matemática para estas 405 estimaciones y la tabla 7.6 da algunos de los valores de los 405 coeficientes de regresión obtenidos de los datos de Alemania para la variable HISEI.

2. La estimación del coeficiente de regresión final es igual a $\hat{\beta} = \frac{\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5}{5}$, es decir, para HISEI, $\hat{\beta} = \frac{2{,}30 + 2{,}27 + 2{,}26 + 2{,}31 + 2{,}34}{5} = 2{,}30$.

3. La estimación de la varianza muestral final es igual a

$$\sigma^2_{(\hat{\beta})} = \frac{1}{5}\left(\sigma^2_{(\hat{\beta}_1)} + \sigma^2_{(\hat{\beta}_2)} + \sigma^2_{(\hat{\beta}_3)} + \sigma^2_{(\hat{\beta}_4)} + \sigma^2_{(\hat{\beta}_5)}\right),$$

es decir, para HISEI,

$$\sigma^2_{(\hat{\beta})} = \frac{(0{,}11)^2 + (0{,}11)^2 + (0{,}11)^2 + (0{,}11)^2 + (0{,}11)^2}{5} = 0{,}012$$

4. La varianza de la imputación es igual a $\sigma^2_{(prueba)} = \frac{1}{4}\sum_{i=1}^{5}\left(\hat{\beta}_i - \hat{\beta}\right)^2$, es decir, para HISEI,

$$\sigma^2_{(prueba)} = \frac{(2{,}30 - 2{,}30)^2 + (2{,}27 - 2{,}30)^2 + \cdots + (2{,}34 - 2{,}30)^2}{4} = \frac{0{,}0041}{4} = 0{,}001$$

5. La varianza del error es igual a $\sigma^2_{(error)} = \sigma^2_{(\hat{\beta})} + \left(1{,}2\sigma^2_{(prueba)}\right)$, es decir, para HISEI,

$$\sigma^2_{(error)} = 0{,}01248 + (1{,}2 \cdot 0{,}0001) = 0{,}01368$$

6. El error típico es igual a $SE = \sqrt{\sigma^2_{(error)}} = \sqrt{0{,}01368} = 0{,}117$

Como 2,30 dividido por 0,117 es 19,66, el coeficiente de regresión para HISEI es significativamente distinto de 0.

Tabla 7.5. Las estimaciones de los 405 coeficientes de regresión

Peso	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Replicación 1	$\hat{\beta}_{1\_1}$	$\hat{\beta}_{2\_1}$	$\hat{\beta}_{3\_1}$	$\hat{\beta}_{4\_1}$	$\hat{\beta}_{5\_1}$
Replicación 2	$\hat{\beta}_{1\_2}$	$\hat{\beta}_{2\_2}$	$\hat{\beta}_{3\_2}$	$\hat{\beta}_{4\_2}$	$\hat{\beta}_{5\_2}$
Replicación 3	$\hat{\beta}_{1\_3}$	$\hat{\beta}_{2\_3}$	$\hat{\beta}_{3\_3}$	$\hat{\beta}_{4\_3}$	$\hat{\beta}_{5\_3}$
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
Replicación 80	$\hat{\beta}_{1\_80}$	$\hat{\beta}_{2\_80}$	$\hat{\beta}_{3\_80}$	$\hat{\beta}_{4\_80}$	$\hat{\beta}_{5\_80}$
Varianza muestral	$\sigma^2_{(\hat{\beta}_1)}$	$\sigma^2_{(\hat{\beta}_2)}$	$\sigma^2_{(\hat{\beta}_3)}$	$\sigma^2_{(\hat{\beta}_4)}$	$\sigma^2_{(\hat{\beta}_5)}$

Tabla 7.6. Estimaciones de los coeficientes de regresión de HISEI y su respectiva varianza muestral en la escala de competencia matemática para Alemania después de eliminar el efecto del género

Peso	PV1	PV2	PV3	PV4	PV5
Final	2,30	2,27	2,26	2,31	2,34
Replicación 1	2,31	2,30	2,31	2,33	2,35
.....	.....	.....	.....	.....	.....
Replicación 80	2,24	2,21	2,21	2,23	2,27
Varianza muestral	$(0,11)^2$	$(0,11)^2$	$(0,11)^2$	$(0,11)^2$	$(0,11)^2$

También se ha desarrollado una macro de SPSS® para los análisis de regresión con valores plausibles como variables dependientes. La sintaxis de SPSS® se presenta en el cuadro 7.5.

Cuadro 7.5. Sintaxis de SPSS® para ejecutar la macro de regresión lineal simple con valores plausibles

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU' & not missing(st03q01) & not missing(HISEI)).
compute gender=0.
if (st03q01=1) gender=1.
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file 'c:\pisa\macros\mcr_SE_reg_PV.sps'.

* EJECUTAR LA MACRO.
REG_PV  nrep = 80/
        ind = hisei gender/
        dep = math/
        grp = cnt/
        wgt = w_fstuwt/
        rwgt = w_fstr/
        cons = 0.05/
        infile = 'c:\PISA\Data2003\DEU.sav'/.

* REESTRUCTURAR LOS DATOS (OPCIONAL).
Recode IND ('HISEI'='b1')('gender'='b2').
sort cases by cnt ind.
CASESTOVARS /ID = cnt /INDEX = ind /GROUPBY = index
 /drop= stat1 stat2 stat3 stat4 stat5 var1 var2 var3 var4 var5 pv_var pvar1
pvar2 pvar3 pvar4 pvar5 pvmerr.
```

Además de los argumentos comunes a todas las macros, la raíz de los nombres de variables de valores plausibles también debe especificarse, así como la lista de variables independientes. La estructura del archivo de salida se presenta en la tabla 7.7.

Tabla 7.7. Estructura del archivo de salida a partir del cuadro 7.5

CNT	CLASS	STAT	SE
DEU	b_0	409,20	7,22
DEU	HISEI	2,30	0,117
DEU	gender	−13,83	3,56

Una visión rápida de estos resultados muestra que todos los parámetros de regresión son significativamente distintos de 0.

El error típico de los coeficientes de correlación con valores plausibles

También se ha desarrollado una macro de SPSS® para calcular la correlación entre un conjunto de valores plausibles y otra variable. La sintaxis de SPSS® para ejecutar esta macro se presenta en el cuadro 7.6 y la estructura del archivo de salida, en la tabla 7.8.

Cuadro 7.6. Sintaxis de SPSS® para ejecutar la macro de correlación con valores plausibles

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU').
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file "C:\PISA\macros\mcr_SE_cor_1PV.sps".

* EJECUTAR LA MACRO.
COR_1PV nrep = 80/
       nopv = hisei/
       pv = math/
       grp = cnt/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'/.
```

Tabla 7.8. Estructura del archivo de salida a partir del cuadro 7.6

CNT	STAT	SE
DEU	0,39	0,02

Correlación entre dos conjuntos de valores plausibles

Quizá haya investigadores interesados en la correlación entre las distintas áreas y subáreas. Por ejemplo, algunos podrían desear calcular la correlación entre las subáreas de lectura o entre las subáreas de matemáticas, o entre lectura y matemáticas, usando las bases de datos de PISA 2000 y PISA 2003.

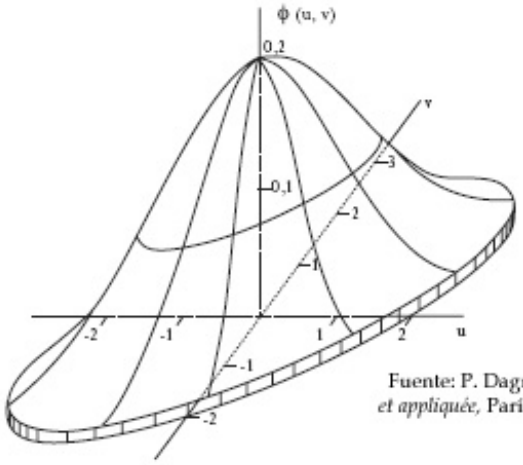
Como se describe en *PISA 2003 Technical Report* (OCDE, 2005), la evaluación PISA utilizó diseños de evaluación incompletos, es decir, los alumnos debían responder a un subconjunto de la batería de ítems. Además, mientras que todos los alumnos fueron evaluados en el área principal, sólo un subconjunto de ellos fue evaluado en las áreas secundarias.

PISA 2000 sólo incluían valores plausibles de alumnos para un área secundaria si respondían a las preguntas de esta. Por lo tanto, calcular la correlación entre la lectura y las matemáticas, por ejemplo, utilizando la base de datos de PISA 2000, exigiría trabajar con un subconjunto de alumnos.[1]

Para facilitar análisis secundarios, PISA 2003 aporta valores plausibles para todas las áreas y todos los alumnos, sin considerar si éstos han sido evaluados en realidad o no. Pasar por alto el estatus de evaluación es posible porque los datos cognitivos de PISA están escalados según modelos multidimensionales.

Puesto que esto es más fácil de ilustrar de modo gráfico, supongamos que sólo se evaluaron dos subáreas, concretamente espacio y forma y cantidad, ambas de matemáticas. Si los materiales de espacio y forma y cantidad se escalaran independientemente, la correlación entre las dos subáreas quedaría muy subestimada. Con objeto de evitar este problema, ambos materiales se escalan juntos. El modelo elaborará una distribución posterior en dos dimensiones, en lugar de dos dimensiones posteriores unidimensionales como se describió en el capítulo 5. La figura 7.1 presenta gráficamente una distribución normal bidimensional.

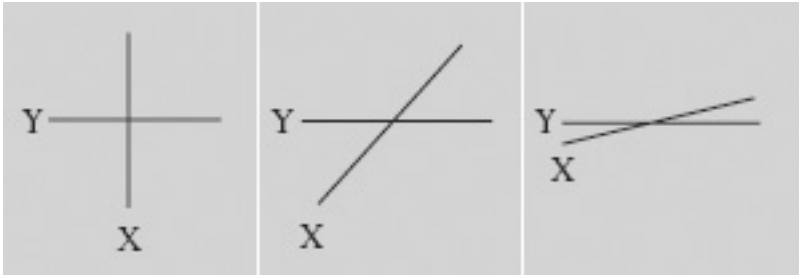
Figura 7.1. Una distribución bidimensional



Fuente: P. Dagnelie (1988): *Statistique theorique et appliquée,* París y Bruselas: De Boek et Lavcier.

Para describir correctamente tales distribuciones, hacen falta dos medias, dos varianzas y una correlación. Si la correlación es igual a 0, los dos ejes serán ortogonales. Conforme el valor absoluto de la correlación empieza a aumentar, el ángulo formado por los dos ejes baja de 90 grados[2]. Dos ejes que se superpongan totalmente representarían una correlación de 1,0 (o −1,0). Estos distintos casos se ilustran en la figura 7.2.

Figura 7.2. Ejes para distribuciones normales bidimensionales



Con un modelo bidimensional, el primer valor plausible para la subárea cantidad se obtendrá al mismo tiempo que el primer valor plausible para la subárea espacio y forma. Por cada alumno, esto consistirá en dibujar aleatoriamente un punto en el diagrama de dispersión. Las magnitudes de los dos valores plausibles serán las coordenadas del punto en los dos ejes. El mismo procedimiento se aplica a los valores plausibles segundo, tercero, cuarto y quinto.

Puesto que las áreas y subáreas de PISA tienen gran correlación, como muestra el gráfico de la derecha en la figura 7.2, es muy improbable que un alumno obtenga una puntuación alta para el primer valor plausible en la subárea cantidad (PV1MATH4) y una puntuación baja para el primer valor plausible en la subárea espacio y forma (PV1MATH1). Si se obtuvieran valores plausibles independientemente para estas dos subáreas de matemáticas, un caso así sería posible y, por tanto, la correlación quedaría subestimada.

Puesto que cada selección es independiente, para calcular la correlación entre las dos subáreas es necesario calcular la correlación entre cada conjunto de valores plausibles:

- PV1MATH1 Y PV1MATH4;
- PV2MATH1 Y PV2MATH4;
- PV3MATH1 Y PV3MATH4;
- PV4MATH1 Y PV4MATH4;
- PV5MATH1 Y PV5MATH4.

La tabla 7.9 presenta los 25 coeficientes de correlación entre los cinco valores plausibles en las subáreas de cantidad y espacio y forma, respectivamente, para Alemania en PISA 2003.

Tabla 7.9. Correlación entre los cinco valores plausibles para las subáreas cantidad y espacio y forma

	PV1MATH1	PV2MATH1	PV3MATH1	PV4MATH1	PV5MATH1
PV1MATH4	0,90	0,83	0,84	0,84	0,83
PV2MATH4	0,83	0,90	0,84	0,84	0,83
PV3MATH4	0,84	0,83	0,90	0,84	0,83
PV4MATH4	0,83	0,83	0,84	0,90	0,83
PV5MATH4	0,83	0,83	0,84	0,84	0,90

Como se muestra en la tabla 7.9, los coeficientes de correlación en la diagonal de la matriz cuadrada son considerablemente más altos que los demás coeficientes. Por lo tanto, la estimación de la correlación final entre estas dos subáreas será la media de los cinco coeficientes de correlación de la diagonal.

La sintaxis de SPSS® se da en el cuadro 7.7.

Cuadro 7.7. Sintaxis de SPSS® para el cálculo de la correlación entre cantidad y espacio y forma

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU').
save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file "c:\pisa\macros\mcr_SE_cor_2PV.sps".

* EJECUTAR LA MACRO.
COR_2PV nrep = 80/
       pv1 = math1/
       pv2 = math4/
       grp = cnt/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'/.
```

Las cinco estimaciones de correlación y sus respectivos errores típicos se dan en la tabla 7.10.

Tabla 7.10. Las cinco estimaciones de correlación entre cantidad y forma y espacio y sus respectivas varianzas muestrales

	PV1	PV2	PV3	PV4	PV5
Correlación	0,8953	0,8964	0,8996	0,8978	0,8958
Varianza muestral	$(0,0040)^2$	$(0,0033)^2$	$(0,0034)^2$	$(0,0037)^2$	$(0,0038)^2$

La estimación de la correlación final es igual a:

$$\hat{\rho} = \frac{\hat{\rho}_1 + \hat{\rho}_2 + \hat{\rho}_3 + \hat{\rho}_4 + \hat{\rho}_5}{5}$$, es decir:

$$\hat{\rho} = \frac{0,8953 + 0,8964 + \cdots + 0,8958}{5} = 0,8970$$

La varianza muestral final es igual a:

$$\sigma^2_{(\hat{\rho})} = \frac{\sum_{i=1}^{5} \sigma^2_{(\hat{\rho}_i)}}{5}$$, es decir:

$$\sigma^2_{(\hat{\rho})} = \frac{(0,0040)^2 + (0,0033)^2 + \cdots + (0,0038)^2}{5} = 0,000013$$

La varianza de imputación (o de medida) puede calcularse como:

$$\sigma^2_{(prueba)} = \frac{1}{4}\sum_{i=1}^{5} (\hat{\rho}_1 - \hat{\rho})^2 = 0,000003$$

La varianza de error es igual a:

$$\sigma^2_{(error)} = \sigma^2_{(\hat{\rho})} + \left(1{,}2\sigma^2_{(prueba)}\right) = 0{,}000017$$

El error típico es igual a:

$$SE = \sqrt{\sigma^2_{(error)}} = \sqrt{0{,}000017} = 0{,}0041$$

El cálculo de la correlación entre dos áreas o entre una subárea y un área podría resultar problemático en algunos casos en las bases de datos de PISA. PISA 2000 utilizó dos modelos de escala:

- un modelo tridimensional con matemáticas, lectura y ciencias;
- un modelo pentadimensional con matemáticas, lectura (recuperación de la información, interpretación de los textos y reflexión) y ciencias.

PISA 2003 también usó dos modelos de escala:

- un modelo tetradimensional con matemáticas, solución de problemas, lectura y ciencias;
- un modelo heptadimensional con matemáticas (espacio y forma, cambio y relaciones, incertidumbre, cantidad), solución de problemas, lectura y ciencias.

Las bases de datos de PISA deberían contener dos conjuntos de valores plausibles para cada una de las áreas secundarias. Como esto resultaría demasiado confuso, sólo se proporcionó un conjunto. Por tanto, los coeficientes de correlación están subestimados.

Esto puede confirmarse al examinar los datos. En el caso de un área secundaria y una subescala del área principal, los coeficientes de correlación en la diagonal no difieren de las demás correlaciones, puesto que estos dos conjuntos de valores plausibles se generaron mediante dos modelos distintos.

En PISA 2003, así como en PISA 2000, los valores plausibles para las áreas secundarias incluidas en las bases de datos se generaron con el área principal como escala combinada. Esto significa que:

- puede calcularse la correlación entre un área secundaria y la escala combinada del área principal;
- puede calcularse la correlación entre dos áreas secundarias;
- puede calcularse la correlación entre las subáreas;
- no es posible calcular la correlación entre áreas secundarias y una de las subescalas del área principal.

Un método abreviado incorrecto

Un método abreviado que contiene un grave error, común cuando se realizan análisis mediante valores plausibles, implica el cálculo de la media de los cinco valores plausibles, antes de seguir analizando.

En el capítulo 5 se describió el estimador EAP del rendimiento del alumno. Recordemos que el estimador EAP es igual a la media de la distribución posterior. Por lo tanto, calcular para cada alumno la media de los cinco valores plausibles es más o menos igual a la estimación EAP.

En el capítulo 5, se comparó la eficiencia del estimador EAP con el WLE y los PV para las estimaciones de algunos estadísticos. Se indicó que el estimador EAP

- subestima la desviación típica;
- sobreestima la correlación entre el rendimiento de los alumnos y algunas variables del entorno;
- subestima la varianza entre centros.

Por tanto, calcular la media de los cinco valores plausibles y luego calcular estadísticos a partir de esta nueva puntuación sesgaría los resultados, como ocurre con los estimadores EAP. La tabla 7.11 proporciona, según el país, la desviación típica en la escala de matemáticas, usando el método correcto, como se describe en este capítulo, y también el método incorrecto de obtener las medias de los cinco valores plausibles para cada alumno y, después, calcular la desviación típica según esta nueva puntuación. El resultado del último proceso se denomina *pseudo EAP*. Como se muestra en la tabla 7.11, el pseudo EAP subestima la desviación típica.

Tabla 7.1. Desviaciones típicas en la escala de matemáticas mediante el método correcto (valores plausibles) y promediando los valores plausibles de cada alumno (*pseudo EAP*)

	Valores plausibles	Pseudo EAP			Valores plausibles	Pseudo EAP
AUS	95,42	91,90		KOR	92,38	89,07
AUT	93,09	89,91		LIE	99,06	95,42
BEL	109,88	106,65		LUX	91,86	88,28
BRA	99,72	94,79		LVA	87,90	83,92
CAN	87,11	83,37		MAC	86,95	82,72
CHE	98,38	94,97		MEX	85,44	80,52
CZE	95,94	92,50		NLD	92,52	89,89
DEU	102,59	99,54		NOR	92,04	88,31
DNK	91,32	87,52		NZL	98,29	95,07
ESP	88,47	84,52		POL	90,24	86,49
FIN	83,68	79,77		PRT	87,63	83,91
FRA	91,70	88,07		RUS	92,25	87,81
GBR	92,26	89,18		SVK	93,31	89,86
GRC	93,83	89,49		SWE	94,75	91,07
HKG	100,19	96,99		THA	81,95	77,15
HUN	93,51	89,71		TUN	81,97	76,86
IDN	80,51	74,86		TUR	104,74	100,79
IRL	85,26	82,03		URY	99,68	95,21
ISL	90,36	86,55		USA	95,25	92,12
ITA	95,69	92,00		YUG	84,65	80,43
JPN	100,54	96,96				

Un método abreviado sin sesgo

Las tablas 7.1 y 7.5 dan, respectivamente, los 405 estimadores de las medias y de la regresión necesarios para calcular una estimación final de una media o de un coeficiente de regresión y

sus respectivos errores típicos.

En promedio, analizar un valor plausible en lugar de cinco proporciona estimaciones de población no sesgadas, así como varianzas muestrales no sesgadas de estas estimaciones. Sin embargo, no será posible estimar la varianza de imputación mediante este método.

Por tanto, un método abreviado sin sesgo debería consistir en:

- calcular, mediante uno de los cinco valores plausibles, el estimador estadístico y su varianza muestral utilizando el peso final de los alumnos, así como los 80 pesos replicados;
- calcular el estimador estadístico usando sólo el peso final de los alumnos en los otros cuatro valores plausibles;
- calcular el estimador estadístico final promediando los estimadores estadísticos de los valores plausibles;
- calcular la varianza de imputación, como ya se ha descrito;
- combinar la varianza de imputación y la varianza muestral, también como ya se ha descrito.

Este método abreviado sin sesgo se presenta en la tabla 7.12 para la estimación de una media y su error típico. Sólo necesita el cálculo de 85 estimaciones, en lugar de 405. La estimación final según este método será igual a la obtenida gracias al procedimiento largo, pero el error típico quizá difiera ligeramente.

Tabla 7.12. Método abreviado sin sesgo para una estimación poblacional y su error típico

Peso	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
Replicación 1	$\hat{\mu}_{1\_1}$				
Replicación 2	$\hat{\mu}_{1\_2}$				
Replicación 3	$\hat{\mu}_{1\_3}$				
.....	.....				
.....	.....				
Replicación 80	$\hat{\mu}_{1\_80}$				
Varianza muestral	$\sigma^2_{(\hat{\mu}_1)}$				

Conclusiones

Este capítulo describe los distintos pasos para analizar datos con valores plausibles. También proporciona algunas macros de SPSS® para facilitar los cálculos.

Se ha llamado la atención así mismo sobre un error frecuente que consiste en calcular la media de los valores plausibles alumno a alumno y añadir este valor a la base de datos para utilizarlo como la puntuación de los alumnos en los análisis. A diferencia de este sistema, el método correcto supone que el proceso de promediar debe ocurrir siempre en la etapa final, es decir, sobre el estadístico que se obtendrá.

Así mismo se presentó la cuestión concreta de analizar dos conjuntos de valores plausibles en el caso de una correlación. El procedimiento aplicado puede extenderse a un análisis de regresión lineal.

Por último, se describió un método abreviado sin sesgo que resulta útil para procedimientos que llevan mucho tiempo, como los procedimientos multinivel.

---

[1] Para más información, véase *Manual for the PISA 2000 Database* (OCDE, 2002*b*).

[2] Un coeficiente de correlación puede expresarse mediante los cosenos del ángulo formado por las dos variables.

Capítulo 8

El uso de niveles de rendimiento

Introducción

Los valores para el rendimiento de los alumnos en cuanto a su competencia en lectura, matemáticas y ciencias suelen considerarse como variables latentes continuas. Con objeto de facilitar la interpretación de las puntuaciones asignadas a los alumnos, en PISA 2000 la escala combinada de competencia lectora y las escalas de matemáticas y ciencias se diseñaron para que tuvieran una puntuación media de 500 puntos y una desviación típica de 100 en los países de la OCDE que participaron. Esto significa que aproximadamente dos tercios de los alumnos de la OCDE obtuvieron entre 400 y 600 puntos.

En PISA 2003, se elaboraron por primera vez cinco escalas de matemáticas: la escala de matemáticas, la escala de espacio y forma, la escala de cambio y relaciones, la escala de cantidad y la escala de incertidumbre, con el objetivo de alcanzar una puntuación media de 500 puntos entre los países de la OCDE. Sin embargo, a diferencia de la escala de matemáticas, las escalas de lectura y ciencias de PISA 2003 se anclaron a los resultados de PISA 2000.

Con el propósito de mejorar la accesibilidad de los resultados a los responsables políticos y educadores, se desarrollaron las escalas descritas de rendimento para las áreas de evaluación. Puesto que estas escalas se dividen en niveles de dificultad y rendimiento, puede obtenerse una clasificación del rendimiento de los alumnos, así como una descripción de las destrezas asociadas con cada nivel de rendimiento. Los niveles sucesivos se asocian con tareas de dificultad cada vez mayor.

En PISA 2000, se definieron y publicaron cinco niveles de rendimiento en lectura en el informe inicial *Knowledge and Skills for Life: First Results for PISA 2000* (OCDE, 2001). En PISA 2003, también se definieron y publicaron seis niveles de rendimiento en matemáticas en el informe inicial *Learning for Tomorrow's World – First Results from PISA 2003* (OCDE, 2004*a*).

Este capítulo mostrará cómo se obtienen los niveles de rendimiento a partir de las bases de datos de PISA y cómo se utilizan.

Generación de los niveles de rendimiento

Los niveles de rendimiento no se incluyen en las bases de datos, pero pueden obtenerse a partir de los valores plausibles.

En PISA 2003, los puntos de corte que enmarcan los niveles de rendimiento en matemáticas son exactamente: 357,77; 420,07; 482,38; 544,68; 606,99 y 669,3[1]. Mientras que algunos investigadores quizá comprendan que pueden asignarse distintas puntuaciones posibles a cada alumno, comprender que pueden asignarse niveles distintos es más difícil. Por lo tanto, quizá sientan la tentación de calcular la media de los cinco valores plausibles y asignen después a cada alumno un nivel de rendimiento basado en esta media.

Como ya se explicó en los capítulos 5 y 7, tal procedimiento es similar a la asignación de una puntuación EAP a cada alumno, y ahora conocemos bien los sesgos que aquejan a un estimador semejante. Puesto que usar puntuaciones EAP subestima la desviación típica, la estimación por este medio de los porcentajes de alumnos en cada nivel de rendimiento subestimará los porcentajes en los niveles más bajos y más altos y sobreestimará los porcentajes de los niveles medios.

Como ya se ha indicado, las encuestas educativas internacionales no pretenden estimar con exactitud el rendimiento de los alumnos individuales; pretenden describir características de la población. Por ello, es perfectamente posible asignar a ciertos alumnos un nivel de rendimiento distinto para diferentes valores plausibles. Así, se asignarán cinco niveles plausibles de rendimiento a cada alumno según sus cinco valores plausibles. La sintaxis de SPSS® para la generación de niveles plausibles de rendimiento en matemáticas se proporciona en el cuadro 8.1.

PISA 2000 sólo proporcionó puntos de corte para niveles de rendimiento en lectura. Por tanto, sólo pueden generarse niveles de rendimiento en la escala combinada de competencia lectora y en las tres subescalas.

PISA 2003 proporcionó otros puntos de corte para los niveles de rendimiento en matemáticas. Por tanto, pueden generarse niveles de rendimiento en la escala de matemáticas y en sus cuatro subescalas, así como en la escala combinada de competencia lectora.

El proceso iterativo recodificará cada una de las 25 variables de valores plausibles en una nueva variable, con siete categorías etiquetadas del 0 al 6.

Cuadro 8.1. Sintaxis de SPSS® para la generación de niveles de rendimiento

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU').
do repeat pv=pv1math pv2math pv3math pv4math pv5math
             pv1math1 pv2math1 pv3math1 pv4math1 pv5math1
             pv1math2 pv2math2 pv3math2 pv4math2 pv5math2
             pv1math3 pv2math3 pv3math3 pv4math3 pv5math3
             pv1math4 pv2math4 pv3math4 pv4math4 pv5math4
 /lev=mlev1 to mlev5 m1lev1 to m1lev5 m2lev1 to m2lev5 m3lev1
  to m3lev5 m4lev1 to m4lev5.

if (pv<=357.77) lev=0.
if (pv>357.77 and pv<=420.07) lev=1.
if (pv>420.07 and pv<=482.38) lev=2.
if (pv>482.38 and pv<=544.68) lev=3.
if (pv>544.68 and pv<=606.99) lev=4.
if (pv>606.99 and pv<=669.3) lev=5.
if (pv>669.3) lev=6.

End repeat.
Formats mlev1 to mlev5 m1lev1 to m1lev5 m2lev1 to m2lev5 m3lev1 to
   m3lev5 m4lev1 to m4lev5 (F1.0).
save outfile='c:\PISA\Data2003\DEU_LEV.sav'.
```

El cálculo del porcentaje de alumnos en cada nivel de rendimiento y su respectivo error típico es exactamente igual al cálculo de la estimación de una media y su error típico que se describió en el capítulo 7, es decir:

- Para cada valor plausible, es preciso calcular el porcentaje de alumnos en cada nivel de rendimiento y su respectivo error típico. Por cada nivel de rendimiento, se obtendrán 5 estimaciones de porcentajes, llamadas $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4$ y $\hat{\pi}_5$. A partir de las 80 replicaciones aplicadas a cada una de las 5 variables de los niveles de rendimiento, por cada uno se estimarán 5 varianzas muestrales, llamadas respectivamente $\sigma^2_{(\hat{\pi}_1)}, \sigma^2_{(\hat{\pi}_2)}, \sigma^2_{(\hat{\pi}_3)}, \sigma^2_{(\hat{\pi}_4)}$ y $\sigma^2_{(\hat{\pi}_5)}$. Estas cinco

estimaciones de porcentajes y de sus respectivas varianzas muestrales se presentan en la tabla 8.1.

- La estimación de la media final es igual al promedio de las cinco estimaciones de la media, es decir,

$$\hat{\pi} = \frac{1}{5}\left(\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 + \hat{\pi}_4 + \hat{\pi}_5\right).$$

- La varianza muestral final es igual al promedio de las cinco varianzas muestrales, es decir,

$$\sigma^2_{(\hat{\pi})} = \frac{1}{5}\left(\sigma^2_{(\hat{\pi}_1)} + \sigma^2_{(\hat{\pi}_2)} + \sigma^2_{(\hat{\pi}_3)} + \sigma^2_{(\hat{\pi}_4)} + \sigma^2_{(\hat{\pi}_5)}\right).$$

- La varianza de imputación, también llamada *varianza del error de medida*, se calcula así[2]:

$$\sigma^2_{(prueba)} = \frac{1}{4}\sum_{i=1}^{5}\left(\hat{\pi}_i - \hat{\pi}\right)^2 .$$

- La varianza muestral y la varianza de la imputación se combinan para obtener la varianza de error final así: $\sigma^2_{(error)} = \sigma^2_{(\hat{\pi})} + \left(1{,}2\sigma^2_{(prueba)}\right)$.

- El error típico es igual a la raíz cuadrada de la varianza de error.

Este proceso se repite para cada nivel de aptitud.

Tabla 8.1. Las 405 estimaciones de porcentajes para un determinado nivel de aptitud

Peso	MLEV PV1	MLEV PV2	MLEV PV3	MLEV PV4	MLEV PV5
Final	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$
Replicación 1	$\hat{\pi}_{1\_1}$	$\hat{\pi}_{2\_1}$	$\hat{\pi}_{3\_1}$	$\hat{\pi}_{4\_1}$	$\hat{\pi}_{5\_1}$
Replicación 2	$\hat{\pi}_{1\_2}$	$\hat{\pi}_{2\_2}$	$\hat{\pi}_{3\_2}$	$\hat{\pi}_{4\_2}$	$\hat{\pi}_{5\_2}$
Replicación 3	$\hat{\pi}_{1\_3}$	$\hat{\pi}_{2\_3}$	$\hat{\pi}_{3\_3}$	$\hat{\pi}_{4\_3}$	$\hat{\pi}_{5\_3}$
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
Replicación 80	$\hat{\pi}_{1\_80}$	$\hat{\pi}_{2\_80}$	$\hat{\pi}_{3\_80}$	$\hat{\pi}_{4\_80}$	$\hat{\pi}_{5\_80}$
Varianza muestral	$\sigma^2_{(\hat{\pi}_1)}$	$\sigma^2_{(\hat{\pi}_2)}$	$\sigma^2_{(\hat{\pi}_3)}$	$\sigma^2_{(\hat{\pi}_4)}$	$\sigma^2_{(\hat{\pi}_5)}$

De esta forma, se estimarán 405 porcentajes por cada nivel de aptitud. Como existen siete niveles en matemáticas, esto significa que se estimarán 2.835 porcentajes.

Los siete niveles de aptitud en matemáticas son:
1. por debajo del nivel 1;
2. nivel 1;
3. nivel 2;
4. nivel 3;
5. nivel 4;

6. nivel 5;
7. nivel 6.

Al aplicar sucesivamente cinco veces la macro GRPPCT descrita en el capítulo 5, se obtendrán, según el nivel de rendimiento, cinco estimaciones de porcentajes y cinco estimaciones de errores típicos que pueden combinarse para obtener la estimación final y su error típico.

El cuadro 8.2 presenta la sintaxis de SPSS® para ejecutar sucesivamente cinco veces la macro GRPPCT. La tabla 8.2 presenta, por cada nivel de rendimiento, las cinco estimaciones y sus respectivas varianzas muestrales.

Cuadro 8.2. Sintaxis de SPSS® para calcular los porcentajes de alumnos por nivel de rendimiento en matemáticas

```
* IMPORTAR LA MACRO.
include file "c:\pisa\macros\mcr_SE_GrpPct.sps".

* EJECUTAR LA MACRO 5 VECES.
GRPPCT nrep = 80/ within = cnt/ grp = mlev1/
       wgt = w_fstuwt/ rwgt = w_fstr/
       cons = 0.05/ infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
rename vars (stat se mlev1=stat1 se1 mlev).
save outfile='c:\temp\grppct1.sav' /drop=var.
GRPPCT nrep = 80/ within = cnt/ grp = mlev2/
       wgt = w_fstuwt/ rwgt = w_fstr/
       cons = 0.05/ infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
rename vars (stat se mlev2=stat2 se2 mlev).
save outfile='c:\temp\grppct2.sav' /drop=var.
GRPPCT nrep = 80/ within = cnt/ grp = mlev3/
       wgt = w_fstuwt/ rwgt = w_fstr/
       cons = 0.05/ infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
rename vars (stat se mlev3=stat3 se3 mlev).
save outfile='c:\temp\grppct3.sav' /drop=var.
GRPPCT nrep = 80/ within = cnt/ grp = mlev4/
       wgt = w_fstuwt/ rwgt = w_fstr/
       cons = 0.05/ infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
rename vars (stat se mlev4=stat4 se4 mlev).
save outfile='c:\temp\grppct4.sav' /drop=var.
GRPPCT nrep = 80/ within = cnt/ grp = mlev5/
       wgt = w_fstuwt/ rwgt = w_fstr/
       cons = 0.05/ infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
rename vars (stat se mlev5=stat5 se5 mlev).
save outfile='c:\temp\grppct5.sav' /drop=var.

match files file='c:\temp\grppct1.sav' /file='c:\temp\grppct2.sav'
 /file='c:\temp\grppct3.sav' /file='c:\temp\grppct4.sav'
 /file='c:\temp\grppct5.sav' /by mlev.
exe.
```

Para combinar los resultados:

- se promedian las cinco estimaciones de porcentajes por nivel de rendimiento;
- se promedian las cinco varianzas muestrales por nivel de rendimiento;
- se calcula la varianza de imputación comparando la estimación final y las cinco estimaciones de valores plausibles;

- la varianza muestral final y la varianza de imputación se combinan como de costumbre para obtener la varianza de error final;
- se obtiene el error típico extrayendo la raíz cuadrada de la varianza de error.

Tabla 8.2. Estimaciones y varianzas muestrales, en Alemania, por nivel de rendimiento en matemáticas

Nivel		PV1	PV2	PV3	PV4	PV5
Por debajo delnivel 1	$\hat{\pi}_i$	9,69	9,02	9,12	9,36	8,75
	$\sigma^2_{(\hat{\pi}_i)}$	(0,79)²	(0,73)²	(0,75)²	(0,74)²	(0,71)²
Nivel 1	$\hat{\pi}_i$	11,87	12,68	12,67	12,33	12,52
	$\sigma^2_{(\hat{\pi}_i)}$	(0,74)²	(0,74)²	(0,72)²	(0,71)²	(0,72)²
Nivel 2	$\hat{\pi}_i$	18,20	18,83	19,53	18,87	19,56
	$\sigma^2_{(\hat{\pi}_i)}$	(0,80)²	(0,80)²	(0,86)²	(0,89)²	(0,88)²
Nivel 3	$\hat{\pi}_i$	23,11	22,69	22,14	22,23	22,66
	$\sigma^2_{(\hat{\pi}_i)}$	(0,72)²	(0,67)²	(0,68)²	(0,62)²	(0,81)²
Nivel 4	$\hat{\pi}_i$	21,05	20,95	20,30	20,85	19,91
	$\sigma^2_{(\hat{\pi}_i)}$	(0,89)²	(0,93)²	(0,85)²	(0,82)²	(0,85)²
Nivel 5	$\hat{\pi}_i$	11,65	11,74	12,50	12,13	12,82
	$\sigma^2_{(\hat{\pi}_i)}$	(0,65)²	(0,66)²	(0,70)²	(0,65)²	(0,73)²
Nivel 6	$\hat{\pi}_i$	4,42	4,09	3,74	4,23	3,78
	$\sigma^2_{(\hat{\pi}_i)}$	(0,35)²	(0,38)²	(0,33)²	(0,37)²	(0,36)²

Los resultados finales se presentan en la tabla 8.3.

Tabla 8.3. Estimaciones finales del porcentaje de alumnos, en Alemania, por nivel de rendimiento en matemáticas, y sus respectivos errores típicos

Nivel de aptitud	%	ET
Por debajo del nivel 1	9,19	0,84
Nivel 1	12,42	0,81
Nivel 2	19,00	1,05
Nivel 3	22,57	0,82
Nivel 4	20,61	1,02
Nivel 5	12,17	0,87
Nivel 6	4,05	0,48

Se ha elaborado una macro de SPSS® para calcular el porcentaje de alumnos en cada nivel de rendimiento, así como su respectivo error típico, con una sola ejecución. El cuadro 8.3 presenta la sintaxis de SPSS® para ejecutar la macro y la tabla 8.4 presenta la estructura del archivo de salida. Para la escala de matemáticas, el argumento GRP se establecerá como MLEV.

Cuadro 8.3. Sintaxis de SPSS® para calcular el porcentaje de alumnos según el nivel de rendimiento

```
* IMPORTAR LA MACRO.
include file "c:\pisa\macros\mcr_SE_PctLev.sps".

* EJECUTAR LA MACRO.
PCTLEV nrep = 80/
       within = cnt/
       grp = mlev/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
```

Tabla 8.4. Estructura del archivo de salida del cuadro 8.3

CNT	MLEV	STAT	SE
DEU	0	9,19	0,84
DEU	1	12,42	0,81
DEU	2	19,00	1,05
DEU	3	22,57	0,82
DEU	4	20,61	1,02
DEU	5	12,17	0,87
DEU	6	4,05	0,48

Una vez más, pueden utilizarse diversas variables de agrupación. Por ejemplo, la distribución de alumnos entre los niveles de rendimiento según el género puede obtenerse como en el cuadro 8.4.

Cuadro 8.4. Sintaxis de SPSS® para calcular el porcentaje de alumnos según nivel de rendimiento y género

```
* IMPORTAR LA MACRO.
include file "c:\pisa\macros\mcr_SE_PctLev.sps".

* EJECUTAR LA MACRO.
PCTLEV nrep = 80/
       within = cnt st03q01/
       grp = mlev/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'/.
```

En este caso, la suma de los porcentajes será igual a 100 por país y por género, como muestra la tabla 8.5.

Tabla 8.5. Estructura del archivo de salida a partir del cuadro 8.4

CNT	ST03Q01	MLEV	STAT	SE
DEU	1	0	9,24	1,05
DEU	1	1	12,15	1,02
DEU	1	2	19,92	1,42
DEU	1	3	23,92	1,37
DEU	1	4	20,65	1,18
DEU	1	5	11,25	0,97
DEU	1	6	2,87	0,57
DEU	2	0	8,88	1,04
DEU	2	1	12,53	0,99
DEU	2	2	18,14	1,21
DEU	2	3	21,43	0,98
DEU	2	4	20,72	1,32
DEU	2	5	13,03	1,14
DEU	2	6	5,27	0,65

Como se muestra en la tabla 8.5, el porcentaje de chicas en el nivel 6 es superior al de chicos en el mismo nivel.

La significación estadística de estas diferencias no puede evaluarse con este procedimiento. En el capítulo 10 se proporcionarán más detalles sobre esta cuestión.

Otros análisis con niveles de rendimiento

Otro de los índices elaborados en PISA 2003 es un índice de autoeficacia en matemáticas, denominado MATHEFF.

Para PISA 2003, es relevante analizar la relación entre los niveles de rendimiento y la autoeficacia en matemáticas, ya que probablemente existe una relación recíproca entre estos dos conceptos. Se cree que una mejor autopercepción en matemáticas aumenta el rendimiento del alumno en esta área, pero un aumento de este último podría, a su vez, influir en la primera.

Supongamos que el estadístico de interés es la autoeficacia media por nivel de rendimiento. En términos estadísticos, la autoeficacia en matemáticas se considera la variable dependiente y el nivel de rendimiento, la variable independiente. No existe una macro que pueda calcular directamente la media de una variable continua por cada nivel de rendimiento. Por otra parte, la macro UNIV descrita en el capítulo 6 puede aplicarse cinco veces secuencialmente y los resultados podrían combinarse en una hoja de cálculo de Microsoft® Excel®, por ejemplo. Así será el caso cuando los niveles de rendimiento se utilicen como variables independientes o como variables de clasificación.

El cuadro 8.5 presenta la sintaxis de SPSS® para calcular la media de la autoeficacia de los alumnos por cada nivel de rendimiento. Las estimaciones de las medias y sus respectivos errores típicos se presentan en la tabla 8.6.

Cuadro 8.5. Macro de SPSS® para calcular la media de la autoeficacia en matemáticas por nivel de rendimiento

```
* IMPORTAR LA MACRO.
Include file 'C:\PISA\macros\mcr_SE_univ.sps'.

* EJECUTAR LA MACRO 5 VECES.
univar nrep = 80/ stat = mean/ dep = matheff/
       grp = cnt mlev1/ wgt = w_fstuwt/
       rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'.
rename vars (mlev1 stat var se=mlev stat1 var1 se1).
save outfile='c:\temp\mlev1.sav'.
univar nrep = 80/ stat = mean/ dep = matheff/
       grp = cnt mlev2/ wgt = w_fstuwt/
       rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'.
rename vars (mlev2 stat var se=mlev stat2 var2 se2).
save outfile='c:\temp\mlev2.sav'.
univar nrep = 80/ stat = mean/ dep = matheff/
       grp = cnt mlev3/ wgt = w_fstuwt/
       rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'.
rename vars (mlev3 stat var se=mlev stat3 var3 se3).
save outfile='c:\temp\mlev3.sav'.
univar nrep = 80/ stat = mean/ dep = matheff/
       grp = cnt mlev4/ wgt = w_fstuwt/
       rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'.
rename vars (mlev4 stat var se=mlev stat4 var4 se4).
save outfile='c:\temp\mlev4.sav'.
univar nrep = 80/ stat = mean/ dep = matheff/
       grp = cnt mlev5/ wgt = w_fstuwt/
       rwgt = w_fstr/ cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU_LEV.sav'.
rename vars (mlev5 stat var se=mlev stat5 var5 se5).
save outfile='c:\temp\mlev5.sav'.

match files file'c:\temp\mlev1.sav' /file='c:\temp\mlev2.sav'
 /file='c:\temp\mlev3.sav'  /file='c:\temp\mlev4.sav'
 /file='c:\temp\mlev5.sav' /by MLEV.
exe.
```

Para combinar los resultados:

- se promedian las cinco estimaciones de las medias por nivel de rendimiento;
- se promedian las cinco varianzas muestrales por nivel de rendimiento;
- se calcula la varianza de imputación comparando la estimación final y las cinco estimaciones de valores plausibles;
- la varianza muestral final y la varianza de imputación se combinan como de costumbre para obtener la varianza de error final;
- se obtiene el error típico calculando la raíz cuadrada de la varianza de error.

Tabla 8.6. Estimaciones de las medias y sus errores típicos para la autoeficacia en matemáticas por nivel de rendimiento

Nivel		PV1	PV2	PV3	PV4	PV5
Por debajo del nivel 1	$\hat{\mu}_i$	–0.68	–0.70	–0.74	–0.72	–0.77
	$\sigma^2_{(\hat{\mu}_i)}$	(0.06)²	(0.06)²	(0.06)²	(0.05)²	(0.06)²
Nivel 1	$\hat{\mu}_i$	–0.44	–0.45	–0.42	–0.43	–0.40
	$\sigma^2_{(\hat{\mu}_i)}$	(0.06)²	(0.05)²	(0.06)²	(0.04)²	(0.05)²
Nivel 2	$\hat{\mu}_i$	–0.18	–0.16	–0.17	–0.18	–0.18
	$\sigma^2_{(\hat{\mu}_i)}$	(0.03)²	(0.03)²	(0.03)²	(0.03)²	(0.03)²
Nivel 3	$\hat{\mu}_i$	0.09	0.09	0.12	0.11	0.10
	$\sigma^2_{(\hat{\mu}_i)}$	(0.03)²	(0.03)²	(0.03)²	(0.03)²	(0.03)²
Nivel 4	$\hat{\mu}_i$	0.43	0.45	0.41	0.45	0.44
	$\sigma^2_{(\hat{\mu}_i)}$	(0.03)²	(0.03)²	(0.03)²	(0.03)²	(0.03)²
Nivel 5	$\hat{\mu}_i$	0.85	0.84	0.86	0.79	0.82
	$\sigma^2_{(\hat{\mu}_i)}$	(0.04)²	(0.04)²	(0.03)²	(0.04)²	(0.04)²
Nivel 6	$\hat{\mu}_i$	1.22	1.23	1.27	1.28	1.29
	$\sigma^2_{(\hat{\mu}_i)}$	(0.05)²	(0.05)²	(0.06)²	(0.05)²	(0.07)²

Los resultados finales se presentan en la tabla 8.7.

La sintaxis del cuadro 8.5 puede mejorarse añadiendo instrucciones que calculen la estimación final y su error típico, como en el cuadro 8.6. Puesto que los resultados se han guardado en cinco archivos de datos de salida con exactamente los mismos nombres de variables, es necesario renombrarlas antes.

Cuadro 8.6. Sintaxis de SPSS® para calcular la media de la autoeficacia en matemáticas por nivel de rendimiento

```
*** CÁLCULO DE LA MEDIA FINAL Y DE LA VARIANZA MUESTRAL FINAL (Bm) ***.

compute stat=mean(stat1,stat2,stat3,stat4,stat5).
compute PV_var=mean(var1,var2,var3,var4,var5).
EXE.

*** CÁLCULO DE LA VARIANZA DE IMPUTACIÓN O DE MEDIDA (Bm) ***.

do repeat a=stat1 stat2 stat3 stat4 stat5 /b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** CÁLCULO DEL ERROR TÍPICO: SQRT[V = U + (1+1/M)Bm] ***.

compute SE=sqrt(PV_var+1.2*pvmerr).
Exe.
formats stat (F8.3) se (F10.5).
list cnt mlev stat se.
```

Una vez que los archivos de datos de salida se fusionan según las variables de agrupación (en este caso concreto, según CNT) y según los niveles de aptitud (MLEV):

- la estimación final se calcula promediando las cinco estimaciones;
- la varianza muestral final se calcula promediando las cinco varianzas muestrales,
- se calcula la varianza de imputación;
- se calcula el error típico, combinando la varianza muestral y la varianza de imputación, y extrayendo la raíz cuadrada.

La estructura del archivo de salida se presenta en la tabla 8.7.

Tabla 8.7. Estructura del archivo de salida a partir del cuadro 8.6

CNT	MLEV	STAT	SE
DEU	0	–0,72	0,07
DEU	1	–0,43	0,06
DEU	2	–0,17	0,03
DEU	3	0,10	0,03
DEU	4	0,44	0,03
DEU	5	0,83	0,05
DEU	6	1,26	0,07

La tabla 8.7 muestra que una gran autoeficacia en matemáticas (STAT, «estadístico») se asocia con un nivel de rendimiento más alto (MLEV).

Conclusiones

Este capítulo muestra cómo calcular el porcentaje de alumnos por cada nivel de rendimiento. Como se ha visto, el algoritmo es similar al utilizado para otros estadísticos.

También se comentó la dificultad de efectuar análisis utilizando los niveles de rendimiento como variables independientes o explicativas.

---

[1] En PISA 2000, los puntos de corte que enmarcan los niveles de rendimiento en lectura son exactamente 334,7526; 407,4667; 480,1807; 552,8948 y 625,6088.

[2] Esta fórmula es una simplificación de la fórmula general proporcionada en el capítulo 4. *M,* que representaba el número de valores plausibles, se ha sustituido por 5.

Capítulo 9

Los análisis con las variables del centro

Introducción

La población objetivo en PISA son los alumnos de 15 años. Se eligió esta población porque, a esta edad, en la mayoría de países de la OCDE los alumnos se acercan al final de la escolarización obligatoria, de modo que PISA debería poder aportar una indicación del efecto acumulativo que sobre ellos ha tenido la educación a lo largo de los años. En PISA se utiliza un procedimiento de muestreo en dos etapas. Después de definir la población, se seleccionan las muestras de centros con una probabilidad proporcional a su tamaño. A continuación, se seleccionan aleatoriamente 35 alumnos en cada centro. Como la población objetivo se define por la edad, es posible, por tanto, que los alumnos provengan de cursos diferentes.

La tabla 9.1 presenta la distribución de los alumnos de 15 años según el país y el curso en PISA 2003.

Tabla 9.1. Estimaciones del porcentaje de alumnos de 15 años por curso y país en PISA 2003[a]

	7	8	9	10	11	12
AUS	0,01	0,14	8,34	72,26	19,21	0,05
AUT	0,30	5,07	43,18	51,45		
BEL	0,33	3,69	29,64	65,49	0,85	
BRA	13,70	24,82	42,89	18,08	0,51	
CAN	0,57	2,47	13,74	82,04	1,17	0,00
CHE	0,75	16,90	62,77	19,40	0,18	
CZE	0,15	2,82	44,67	52,36		
DEU	1,70	14,99	59,94	23,25	0,12	
DNK	0,07	9,10	86,96	3,83	0,05	
ESP	0,03	3,18	27,03	69,73	0,02	
FIN	0,26	12,43	87,31			
FRA	0,20	5,37	34,86	57,29	2,23	0,05
GBR			0,02	33,81	63,56	2,61
GRC	0,22	2,09	6,55	76,13	15,01	
HKG	5,12	10,75	25,70	58,36	0,08	
HUN	1,08	5,00	65,13	28,76	0,02	
IDN	2,40	12,68	48,78	34,51	1,57	0,07
IRL	0,02	2,78	60,87	16,68	19,65	
ISL				100,00		
ITA	0,18	1,38	14,20	79,95	4,28	
JPN				100,00		
KOR			1,57	98,33	0,10	
LIE	0,61	20,37	71,26	7,75		
LUX		14,85	55,79	29,25	0,10	
LVA	1,09	16,76	75,96	6,08	0,13	
MAC	12,30	25,88	36,82	24,66	0,34	
MEX	3,62	10,95	40,76	43,69	0,93	0,04
NLD	0,14	4,44	45,61	49,32	0,47	0,02
NOR			0,62	98,68	0,69	
NZL		0,06	6,79	89,38	3,74	0,02
POL	0,72	3,07	95,70	0,51		
PRT	4,25	10,58	20,26	64,32	0,58	
RUS	0,35	2,58	28,74	67,23	1,10	
SVK	0,58	0,92	37,10	60,93	0,46	

[a] Los resultados se basan en la información proporcionada en los cuestionarios de los alumnos; por lo tanto, no están sesgados debido a una tasa de participación diferencial según los cursos.

SWE	0,03	2,36	93,00	4,61		
THA	0,18	1,09	44,06	53,26	1,41	
TUN	15,39	21,99	25,15	34,52	2,94	
TUR	0,84	4,39	3,20	52,12	39,19	0,25
URY	5,67	9,67	18,22	59,36	7,09	
USA	0,28	2,40	29,71	60,63	6,98	
YUG			97,60	2,40		

Especialmente en algunos países, la mayoría de la población de alumnos de 15 años tiende a estar en un mismo curso, mientras que en otros se reparte entre varios cursos.

La población objetivo de PISA puede estar repartida entre varios cursos por diversas razones:

- Si el alumno no pasa un determinado examen en el curso, debe repetir este. Por ejemplo, en algunos países, quizá haya hasta un 30% de alumnos que ya han repetido al menos un curso.
- Incluso si no se recurre a la repetición de curso, la población de alumnos de 15 años podría estar repartida en dos cursos en el momento de la evaluación. Por razones logísticas, la evaluación de PISA tiene lugar en un único año natural. Puesto que el período recomendado de pruebas es en abril (en el hemisferio norte), la población objetivo de PISA queda definida como todos los alumnos cuya edad esté entre 15 años y tres meses y 16 años y dos meses a comienzos del período de pruebas. Si las normas de acceso a la escolarización obligatoria están basadas en años naturales completos, la población objetivo de PISA habría de asistir a un único curso.

Puesto que la población de alumnos de 15 años se reparte entre varios cursos en la mayoría de los países de la OCDE, las muestras dentro de los centros sólo pueden consistir en una muestra aleatoria de alumnos. Por tanto, los alumnos que participan en PISA estudian en diferentes cursos y, dentro de un curso determinado, en diferentes clases o grupos, dependiendo del tamaño del centro. En buena parte porque la muestra de PISA no se basa en las clases, PISA 2000 y PISA 2003 no recopilaron datos de los profesores. Sin embargo, PISA sí recopila datos de los centros. Este capítulo describe cómo deberían analizarse los datos de los centros y por qué.

Ya que la población objetivo de PISA asiste a distintos cursos en la mayoría de los países, sería interesante calcular la mejoría media del rendimiento entre dos cursos adyacentes, de modo que las diferencias de rendimiento entre países puedan traducirse al efecto de un año escolar. Sin embargo, esto llevaría sin duda a la sobreestimación de la mejoría del rendimiento: los alumnos de 15 años que estudian en los cursos inferiores o bien tienen peor rendimiento o bien son más jóvenes, y los que estudian en cursos superiores o bien tienen mejor rendimiento o bien son mayores. Por lo tanto, no pueden hacerse con seguridad comparaciones entre subpoblaciones de distinto curso. Puede intentarse equiparar igualar a estas subpoblaciones eliminando el efecto del rendimiento mediante un conjunto de características del entorno, pero el éxito sería dudoso ya que los elementos comparados no llegarían a ser realmente iguales.

Limitaciones de las muestras de centros en PISA

Como ya se ha mencionado antes, la siguiente afirmación es válida tanto para los estudios de PISA como para los de la IEA:

> Aunque las muestras de alumnos se seleccionaron dentro de una muestra de centros, la muestra de centros se diseñó para optimizar la muestra resultante de alumnos más que para proporcionar una muestra óptima de centros. Por esta razón, siempre es preferible analizar las variables de los centros como atributos de los alumnos, más que como elementos en sí mismos (Gonzalez y Kennedy, 2003).

Estas indicaciones son especialmente importantes en PISA, ya que la población objetivo no se define como un curso, sino como todos los alumnos de una edad determinada.

En algunos países, la educación secundaria de nivel inferior y la de nivel superior se imparten en el mismo centro, mientras que en otros no es así: la educación secundaria inferior y la superior se imparten en centros diferentes. En estos países, normalmente, la transición entre los niveles inferior y superior de la educación secundaria tiene lugar alrededor de los 15 años, esto es, en la mayoría de los casos, al final de la escolarización obligatoria. Como PISA se centra en la población de 15 años, esto significa que una parte de la población objetivo está recibiendo el nivel superior de la educación secundaria, mientras que la otra acude al nivel inferior. Por consiguiente, en algunos países los jóvenes de 15 años pueden estar escolarizados en distintas instituciones educativas.

Como ya se comentó en el capítulo 2, los centros se seleccionan a partir del marco muestral de centros mediante el método de muestreo con probabilidades proporcionales a su tamaño, es decir, en proporción al número de jóvenes de 15 años que asisten al centro. Esto podría significar, por ejemplo, que los centros con el nivel superior de secundaria a los que sólo acudan alumnos con edad superior a la de PISA, 15, no deberían quedar incluidos en el marco muestral de centros. Del mismo modo, los centros con nivel inferior de secundaria sin alumnos de 15 años tampoco deberían incluirse en dicho marco muestral.

De este modo, ni la población de centros del nivel inferior de secundaria, ni la población del nivel superior representan a la población de alumnos de 15 años. Dicho de otro modo, la población objetivo de centros en PISA no coincidirá necesariamente con la red de centros dentro de un determinado país.

Esta ausencia de un perfecto ajuste entre la red de centros habitual del país y la población de centros de PISA influye en cómo deberían analizarse los datos de los centros. Para evitar sesgos en las estimaciones poblacionales, los datos de los centros deben importarse a los archivos de datos de alumnos y deben analizarse con el peso final de alumnos. Esto significa, por ejemplo, que no se estimará el porcentaje de centros públicos frente al de centros privados, sino que se estimará el porcentaje de alumnos de 15 años que asiste a centros públicos frente al de centros privados. Desde el punto de vista de la pedagogía y de la política educativa, lo que verdaderamente importa no es el porcentaje de centros que presente tales características, sino el porcentaje de alumnos que resulta influido por ellas, esto es, el porcentaje de alumnos que acude a un centro con tales características.

Fusión de los archivos de datos de centros y alumnos

El cuadro 9.1 proporciona la sintaxis de SPSS® para fusionar el archivo de datos de alumnos y el archivo de datos de centros. Ambos archivos deben ordenarse primero según las variables de identificación, esto es, CNT, SCHOOLID y STIDSTD en el archivo de datos de alumnos y CNT

y SCHOOLID en el archivo de datos de centros. Después, los dos archivos de datos ordenados pueden fusionarse según las variables de identificación comunes, CNT y SCHOOLID.

Cuadro 9.1. Sintaxis de SPSS® para fusionar el archivo de datos de alumnos y el de centros

```
get file 'c:\pisa\data2003\INT_schi_2003.sav'.
sort cases by subnatio schoolid.
save outfile='c:\pisa\data2003\INT_schi_2003.sav'.

get file='c:\pisa\data2003\INT_stui_2003.sav'.
sort cases by subnatio schoolid.
match files file=* /table='c:\pisa\data2003\INT_schi_2003.sav'
  /by subnatio schoolid.
Select if cnt='DEU'.
Save outfile='c:\pisa\data2003\DEU_sch.sav'.
```

Análisis de las variables de centros

Después de fusionar el archivo de datos de alumnos y el de centros, los datos de los centros pueden analizarse como cualquier variable propia de los alumnos, ya que las variables de los centros se consideran ahora atributos de los alumnos. Sin embargo, en este caso es aún más decisivo utilizar los pesos replicados para calcular los errores de muestreo. No hacerlo así produciría inferencias totalmente engañosas.

El resto de este capítulo explica los métodos para calcular los porcentajes de alumnos según la ubicación del centro y sus respectivos errores típicos, así como el rendimiento medio de los alumnos en la escala de matemáticas según la ubicación del centro.

El cuadro 9.2 presenta la pregunta que aparecía en el cuestionario del centro sobre el tipo de entorno urbano en el que se encuentra ubicado.

Cuadro 9.2. Pregunta a los centros sobre su ubicación en PISA 2003

P1 **¿Cuál de las siguientes opciones describe mejor la comunidad en la que se encuentra situado este centro?**

(Por favor, marque solo una casilla.)

Un pueblo, aldea o población rural (de menos de 3.000 personas) ..... ☐ 1

Un pueblo pequeño (de 3.000 a 15.000 personas) ..... ☐ 2

Una ciudad de tamaño medio (de 15.000 a 100.000 personas) ..... ☐ 3

Una ciudad grande (de 100.000 a 1.000.000 de personas) ..... ☐ 4

Una ciudad grande de más de 1.000.000 de personas ..... ☐ 5

El cuadro 9.3 proporciona la sintaxis de SPSS®. Como se indicó previamente, la macro de SPSS® podría consumir tiempo de ejecución, de modo que se recomienda mantener sólo las variables indispensables para los análisis.

Cuadro 9.3. Sintaxis de SPSS® para calcular el porcentaje de alumnos y el rendimiento medio en matemáticas según la ubicación del centro

```
get file "c:\pisa\data2003\DEU_sch.sav".
select if (not missing(sc01q01)).
Save outfile='c:\pisa\data2003\DEU_sch2.sav'.

* IMPORTAR LAS MACROS.
Include file 'c:\pisa\macros\mcr_SE_GrpPct.sps'.
Include file 'C:\PISA\macros\mcr_SE_pv.sps'.

* EJECUTAR LAS MACROS.
GRPPCT nrep = 80/
       within = cnt/
       grp = sc01q01/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\pisa\data2003\DEU_sch2.sav'/.
save outfile='c:\temp\schoolpct.sav'.

PV nrep = 80/
   stat = mean/
   dep = math/
   grp = cnt sc01q01/
   wgt = w_fstuwt/
   rwgt = w_fstr/
   cons = 0.05/
   infile = 'c:\pisa\data2003\DEU_sch2.sav'/.
save outfile='c:\temp\schoolmean.sav'.
```

Las tablas 9.2 y 9.3 presentan la estructura de los archivos de datos de salida.

Tabla 9.2. Estructura del archivo de resultados schoolpct.sav

CNT	SC01Q01	STAT	SE
DEU	1	5,04	1,37
DEU	2	24,61	2,70
DEU	3	38,76	3,75
DEU	4	19,53	2,77
DEU	5	12,06	1,98

Tabla 9.3. Estructura del archivo de resultados schoolmean.sav

CNT	SC01Q01	STAT	SE
DEU	1	489,65	14,81
DEU	2	507,46	6,19
DEU	3	496,74	8,92
DEU	4	510,24	13,19
DEU	5	507,07	14,13

Recordemos que los datos de los centros se analizaron como si fueran de los alumnos y se ponderaron con el peso final de estos. Por lo tanto, los resultados deberían interpretarse así: el 5,04% de los jóvenes de 15 años acude a un centro situado en un pueblo con menos de 3.000 personas; el 25% de los alumnos acuden a un centro situado en una pequeña población (de 3.000 a 15.000 personas) y así sucesivamente. Los alumnos que estudian en un centro situado en un pueblo pequeño tienen un rendimiento medio de 489,65 y así sucesivamente.

Puesto que los porcentajes para algunas categorías podrían ser pequeños, el error típico será grande para las estimaciones de las medias.

Todas las macros de SPSS® descritas en los capítulos anteriores pueden utilizarse sobre las variables de los centros, una vez que se hayan importado al archivo de datos de los alumnos.

Conclusiones

Por razones estadísticas y pedagógicas, los datos recopilados mediante el cuestionario del centro, así como las variables derivadas de este instrumento, deben analizarse a nivel de los alumnos.

Todas las macros de SPSS® pueden utilizarse sin modificaciones. La interpretación de los resultados debería establecer claramente el nivel del análisis; por ejemplo, el porcentaje de alumnos que acude a un centro situado en un pueblo pequeño y no el porcentaje de centros situado en un pueblo pequeño.

Capítulo 10

El error típico de una diferencia

Introducción

Supongamos que *X* representa la puntuación de los alumnos en una prueba de matemáticas e *Y*, la puntuación en una prueba de ciencias de la misma muestra de alumnos. Para resumir la distribución de las puntuaciones en ambas pruebas, se puede calcular:

- $\mu_{(X)}, \mu_{(Y)}$, que representan respectivamente la media de *X* y la media de *Y*,
- $\sigma^2_{(X)}, \sigma^2_{(Y)}$, que representan respectivamente la varianza de *X* y la varianza de *Y*.

Puede demostrarse que:

$$\mu_{(X+Y)} = \mu_{(X)} + \mu_{(Y)}, \text{ y}$$

$$\sigma^2_{(X+Y)} = \sigma^2_{(X)} + \sigma^2_{(Y)} + 2\text{cov}(X,Y).$$

Si se calcula una puntuación total simplemente sumando las puntuaciones de matemáticas y ciencias, entonces, según estas dos fórmulas, la media de la puntuación total será la suma de las dos medias iniciales, y la varianza de la puntuación total será igual a la suma de la varianza de las dos variables iniciales *X* e *Y*, más dos veces la covarianza entre *X* e *Y*. Esta covarianza representa la relación entre *X* e *Y*. Normalmente, los alumnos con alto rendimiento en matemáticas suelen tenerlo también en ciencias, por lo que en este ejemplo concreto deberíamos esperar una covarianza positiva y alta.

De modo parecido,

$$\mu_{(X-Y)} = \mu_{(X)} - \mu_{(Y)}, \text{ y}$$

$$\sigma^2_{(X-Y)} = \sigma^2_{(X)} + \sigma^2_{(Y)} - 2\text{cov}(X,Y).$$

Dicho de otro modo, la varianza de una diferencia es igual a la suma de las varianzas de las dos variables iniciales menos dos veces la covarianza entre esas dos variables iniciales.

Como se describió en el capítulo 3, una distribución muestral tiene las mismas características que cualquier distribución, excepto que las unidades consisten en estimaciones a partir de muestras y no en observaciones. Por lo tanto,

$$\sigma^2_{(\hat{\mu}_X - \hat{\mu}_Y)} = \sigma^2_{(\hat{\mu}_X)} + \sigma^2_{(\hat{\mu}_Y)} - 2\text{cov}(\hat{\mu}_X, \hat{\mu}_Y).$$

La varianza muestral de una diferencia es igual a la suma de las dos varianzas muestrales iniciales, menos dos veces la covarianza entre las dos distribuciones muestrales de las estimaciones.

Supongamos que queremos determinar si el rendimiento de las chicas es, en promedio, más alto que el de los chicos. Como para todos los análisis estadísticos, es preciso poner a prueba la hipótesis nula. En este ejemplo concreto, consistirá en calcular la diferencia entre la media del rendimiento de los chicos y la media del rendimiento de las chicas, o a la inversa. La hipótesis nula será:

$$H_0 : \hat{\mu}_{(chicos)} - \hat{\mu}_{(chicas)} = 0.$$

Para comprobar esta hipótesis nula, debe calcularse el error típico de esta diferencia y, después, compararlo con la diferencia observada. Los respectivos errores típicos de la estimación de la media para los chicos y para las chicas ($\sigma_{(\hat{\mu}_{chicos})}, \sigma_{(\hat{\mu}_{chicas})}$) son fáciles de calcular.

¿Qué nos dice la covarianza entre las dos variables, es decir, $\hat{\mu}_{(chicos)}, \hat{\mu}_{(chicas)}$? Una covarianza positiva significa que si $\hat{\mu}_{(chicos)}$ aumenta, también lo hará $\hat{\mu}_{(chicas)}$. Una covarianza igual o cercana a 0 significa que a medida que $\hat{\mu}_{(chicos)}$ aumenta o disminuye, $\hat{\mu}_{(chicas)}$ permanecerá invariable. Por último, una covarianza negativa significa que si $\hat{\mu}_{(chicos)}$ aumenta, $\hat{\mu}_{(chicas)}$ disminuirá, y a la inversa.

¿Cómo se correlacionan $\hat{\mu}_{(chicos)}$ y $\hat{\mu}_{(chicas)}$? Supongamos que, en la muestra de centros, se sustituye un centro mixto donde estudian alumnos de bajo rendimiento por un centro mixto donde estudian alumnos de alto rendimiento. La media del país aumentará ligeramente, así como las medias de los chicos y de las chicas. Si continúa el proceso de sustitución, $\hat{\mu}_{(chicos)}$ y $\hat{\mu}_{(chicas)}$ aumentarán probablemente siguiendo un patrón similar. De hecho, en un centro mixto donde acuden chicos con alto rendimiento también suelen estudiar chicas con alto rendimiento. Por lo tanto, la covarianza entre $\hat{\mu}_{(chicos)}$ y $\hat{\mu}_{(chicas)}$ será positiva.

Supongamos ahora que los centros no son mixtos. Un centro de chicos puede sustituir a uno de chicas en la muestra, por lo que $\hat{\mu}_{(chicos)}$ y $\hat{\mu}_{(chicas)}$ cambiarán. Si el género se usa como variable de estratificación, es decir, todos los centros de chicas se asignan a un estrato determinado y todos los centros de chicos se asignan a otro estrato determinado, entonces un centro de chicas sólo puede sustituirse por otro centro de chicas. En este caso, sólo cambiará $\hat{\mu}_{(chicas)}$ y, como el cambio no afectará a $\hat{\mu}_{(chicos)}$, el valor esperado de la covarianza entre $\hat{\mu}_{(chicos)}$ y $\hat{\mu}_{(chicas)}$ será 0.

Por último, una covarianza negativa significa que, si a un centro acuden chicos con alto rendimiento, en ese centro también estudian chicas con bajo rendimiento, o a la inversa. Esta situación no es muy verosímil.

En resumen, el valor esperado de la covarianza será igual a 0 si las dos submuestras son independientes. Si las dos submuestras no son independientes, el valor esperado de la covarianza podría diferir de 0.

En PISA, así como en los estudios de la IEA, las muestras de los países son independientes. Por lo tanto, para cualquier comparación entre dos países, el valor esperado de la covarianza será igual a 0, de modo que el error típico de la estimación es:

$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}}$, donde θ es cualquier estadístico.

Por ejemplo, en la escala de competencia matemática de PISA 2003, la media de Alemania es igual a 503, con un error típico de 3,3, y la media de Bélgica es igual a 529, con un error típico de 2,3. Por lo tanto, la diferencia entre Bélgica y Alemania será 529 − 503 = 26, y el error típico de esta diferencia es:

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}} = \sqrt{(3{,}3)^2 + (2{,}3)^2} = \sqrt{10{,}89 + 5{,}29} = \sqrt{16{,}18} = 4{,}02\,.$$

La diferencia dividida por su error típico, es decir, 26 / 4,02 = 6,46, es superior a 1,96, por lo que resulta significativa. Esto quiere decir que el rendimiento de Bélgica es superior al de Alemania.

De modo similar, el porcentaje de alumnos por debajo del nivel 1 es igual a 9,2 en Alemania (con un error típico de 0,8) y a 7,2 en Bélgica (con un error típico de 0,6). La diferencia es igual a 9,2 − 7,2 = 2,0, y su error típico es igual a:

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}} = \sqrt{(0{,}6)^2 + (0{,}8)^2} = \sqrt{0{,}36 + 0{,}64} = \sqrt{1} = 1\,.$$

La diferencia tipificada es igual a 2 (es decir, 2/1), por lo que también resulta significativa. Así, el porcentaje de alumnos por debajo del nivel 1 es mayor en Alemania que en Bélgica.

Dentro de un determinado país, cualquier submuestra se considerará independiente si la variable categórica empleada para definir las submuestras se utilizó como variable de estratificación explícita. Por ejemplo, puesto que Canadá utilizó las provincias como variable de estratificación explícita, estas submuestras son independientes, y cualquier comparación entre dos provincias no exige la estimación de la covarianza entre las distribuciones muestrales.

Como regla general, cualquier comparación entre países no exige la estimación de la covarianza, pero se recomienda decididamente estimar la covarianza entre las distribuciones muestrales para cualquier comparación dentro de un país.

Como ya se ha descrito en esta sección, la estimación de la covarianza entre, por ejemplo, $\hat{\mu}_{(chicos)}$ y $\hat{\mu}_{(chicas)}$ exigiría la selección de varias muestras y, después, el análisis de la variación de $\hat{\mu}_{(chicos)}$ en unión de $\hat{\mu}_{(chicas)}$. Por supuesto, tal procedimiento no es realista. Así pues, como para cualquier cálculo de un error típico en PISA, se utilizarán métodos de replicación que empleen los pesos replicados incluídos en la base de datos para estimar el error típico de una diferencia.

El error típico de una diferencia sin valores plausibles

Supongamos que un investigador desea comprobar si las chicas alemanas tienen expectativas profesionales más altas que los chicos.

Como ya se describió en el capítulo 6, la macro UNIVAR de SPSS® puede utilizarse para calcular las expectativas profesionales medias de chicos y de chicas, respectivamente.

El cuadro 10.1 presenta la sintaxis de SPSS® para calcular la media de las expectativas profesionales a los 30 años (BSMJ) según el género del alumnado. La tabla 10.1 presenta la estructura del archivo de salida, así como los resultados según el género.

Cuadro 10.1. Sintaxis de SPSS® para calcular la media de las expectativas profesionales según el género

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
Select if cnt='DEU'.
Select if (not missing(st03q01)).
Save outfile='c:\PISA\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
Include file 'C:\PISA\macros\mcr_SE_univ.sps'.

* EJECUTAR LA MACRO.
univar nrep = 80/
       stat = mean/
       dep = bsmj/
       grp = cnt st03q01/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\PISA\Data2003\DEU.sav'.
```

Tabla 10.1. Estructura del archivo de resultados a partir del cuadro 10.1

CNT	ST03Q01	STAT	SE
DEU	1	53,05	0,57
DEU	2	50,58	0,69

En promedio, la expectativa profesional es de 53,05 para las chicas y de 50,58 para los chicos. Como los centros alemanes suelen ser mixtos y el género no se utiliza como una variable de estratificación explícita, el valor esperado de la covarianza podría diferir de 0.

Para calcular el error típico según el género, es necesario calcular la estimación de la media para cada uno de los 80 pesos replicados. La tabla 10.2 presenta la estimación de la media por peso y por género.

Tabla 10.2. Estimaciones de la media para los pesos finales y los 80 replicados por género

Peso	Estimación de la media para las chicas	Estimación de la media para los chicos		Peso	Estimación de la media para las chicas	Estimación de la media para los chicos
Peso final	**53,05**	**50,58**				
Replicación 1	53,29	50,69		Replicación 41	52,69	50,55
Replicación 2	53,16	50,53		Replicación 42	53,28	51,23
Replicación 3	53,16	50,45		Replicación 43	53,07	50,39
Replicación 4	53,30	50,70		Replicación 44	52,95	49,72
Replicación 5	52,79	50,28		Replicación 45	53,31	51,04
Replicación 6	53,14	50,76		Replicación 46	53,72	50,80
Replicación 7	53,04	50,36		Replicación 47	52,91	51,03
Replicación 8	52,97	50,11		Replicación 48	53,10	50,53
Replicación 9	53,28	51,37		Replicación 49	53,05	50,81
Replicación 10	53,01	50,55		Replicación 50	53,79	50,90
Replicación 11	53,26	50,70		Replicación 51	52,65	50,15
Replicación 12	53,16	49,86		Replicación 52	53,30	50,45
Replicación 13	52,81	50,94		Replicación 53	52,68	50,12
Replicación 14	53,21	50,71		Replicación 54	52,74	50,01
Replicación 15	53,39	50,23		Replicación 55	53,50	50,11
Replicación 16	53,06	50,46		Replicación 56	52,54	50,58
Replicación 17	53,34	50,48		Replicación 57	53,31	51,03
Replicación 18	52,71	50,42		Replicación 58	53,13	50,34
Replicación 19	53,18	50,87		Replicación 59	52,72	50,37
Replicación 20	52,82	50,44		Replicación 60	53,49	51,43
Replicación 21	53,36	50,74		Replicación 61	53,13	50,71
Replicación 22	53,15	50,72		Replicación 62	53,61	51,27
Replicación 23	53,24	50,65		Replicación 63	52,74	50,15
Replicación 24	52,68	50,51		Replicación 64	53,19	50,25
Replicación 25	5276	50,44		Replicación 65	53,28	51,04
Replicación 26	52,79	50,43		Replicación 66	52,91	50,94
Replicación 27	53,01	50,58		Replicación 67	53,25	50,85
Replicación 28	53,24	50,12		Replicación 68	53,12	50,74
Replicación 29	52,86	50,68		Replicación 69	53,08	50,31
Replicación 30	52,85	50,02		Replicación 70	52,92	50,44
Replicación 31	52,90	50,85		Replicación 71	53,35	50,63
Replicación 32	53,25	50,60		Replicación 72	53,25	50,75
Replicación 33	53,32	50,54		Replicación 73	52,54	50,42
Replicación 34	52,42	50,55		Replicación 74	52,58	50,20
Replicación 35	52,91	50,72		Replicación 75	52,49	49,75
Replicación 36	53,06	50,36		Replicación 76	52,98	50,96
Replicación 37	52,67	50,73		Replicación 77	53,04	50,24
Replicación 38	53,36	50,16		Replicación 78	53,30	50,44
Replicación 39	52,57	50,36		Replicación 79	52,93	50,36
Replicación 40	53,07	50,58		Replicación 80	52,98	50,76

La estimación de la diferencia final será la diferencia entre las dos estimaciones finales, es decir: 53,05 − 50,58 = 2,47.

El procedimiento para estimar el error típico final es bastante directo. Es exactamente igual al procedimiento descrito en el capítulo 6, excepto que θ es ahora una diferencia, no una media ni un coeficiente de regresión. Los diferentes pasos son:

- la diferencia entre las medias de las chicas y los chicos se calculan por cada réplica;
- cada una de las 80 estimaciones de la diferencia se compara con la estimación de la diferencia final y después se eleva al cuadrado;
- se calcula la suma de los cuadrados y después se divide por 20 para obtener la varianza muestral de la diferencia;
- el error típico es la raíz cuadrada de la varianza muestral.

Estos pasos pueden resumirse así:

$\sigma_{(\hat{\theta})} = \sqrt{\frac{1}{20}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2}$, donde θ es una diferencia.

Concretamente:

- Para la primera replicación, la diferencia entre la estimación de la media de las chicas y la estimación de la media de los chicos es igual a (53,29 − 50,69) = 2,60. Para la segunda replicación, la estimación de la diferencia será igual a (53,16 − 50,53) = 2,63, y así sucesivamente para las 80 replicaciones. Todas estas estimaciones de la diferencia se presentan en la tabla 10.3.
- Cada una de las estimaciones de la diferencia replicadas se compara con la estimación de la diferencia final, y esta diferencia se eleva al cuadrado. Para la primera replicación, será $(2,60 - 2,47)^2 = 0,0164$. Para la segunda replicación, será $(2,63 - 2,47)^2 = 0,0258$. Estas diferencias al cuadrado también se presentan en la tabla 10.3.
- Se suman las diferencias al cuadrado. Esta suma es igual a:

  (0,0164 + 0,0258 + ⋯ + 0,0641) = 9,7360. La varianza muestral de la diferencia es, por tanto, igual a:

  $$\frac{9,7360}{20} = 0,4868 .$$

- El error típico es igual a la raíz cuadrada de 0,4868, es decir, 0,6977.

Puesto que $\frac{2,47}{0,6977}$ es mayor que 1,96, estadísticamente las expectativas profesionales de las chicas son superiores a las de los chicos en Alemania.

Si el investigador hubiera considerado independientes las dos submuestras alemanas, habría obtenido para el error típico de esta diferencia:

$$\sigma_{(\hat{\theta}_i - \hat{\theta}_j)} = \sqrt{\sigma^2_{(\hat{\theta}_i)} + \sigma^2_{(\hat{\theta}_j)}} = \sqrt{(0,57)^2 + (0,69)^2} = 0,895$$

En este caso concreto, la diferencia entre la estimación sin sesgo del error típico (es decir, 0,698) y la estimación sesgada del error típico (es decir, 0,895) es bastante pequeña. Como se mostrará más tarde en este capítulo, la diferencia entre las estimaciones con y sin sesgo del error típico pueden ser considerables.

Tabla 10.3. Estimaciones de la diferencia para los pesos finales y los 80 replicados

Peso	Diferencia entre chicos y chicas (chicas menos chicos)	Diferencia entre la replicación y las estimaciones finales al cuadrado		Peso	Diferencia entre chicos y chicas (chicas menos chicos)	Diferencia entre la replicación y las estimaciones finales al cuadrado
Peso final	**2,47**					
Replicación 1	2,60	0,0164		Replicación 41	2,14	0,1079
Replicación 2	2,63	0,0258		Replicación 42	2,05	0,1789
Replicación 3	2,72	0,0599		Replicación 43	2,68	0,0440
Replicación 4	2,61	0,0180		Replicación 44	3,23	0,5727
Replicación 5	2,51	0,0011		Replicación 45	2,28	0,0373
Replicación 6	2,39	0,0067		Replicación 46	2,92	0,2038
Replicación 7	2,68	0,0450		Replicación 47	1,88	0,3488
Replicación 8	2,86	0,1483		Replicación 48	2,56	0,0084
Replicación 9	1,92	0,3085		Replicación 49	2,23	0,0567
Replicación 10	2,46	0,0002		Replicación 50	2,89	0,1768
Replicación 11	2,57	0,0089		Replicación 51	2,49	0,0004
Replicación 12	3,30	0,6832		Replicación 52	2,85	0,1440
Replicación 13	1,87	0,3620		Replicación 53	2,56	0,0072
Replicación 14	2,50	0,0009		Replicación 54	2,73	0,0667
Replicación 15	3,16	0,4756		Replicación 55	3,39	0,8520
Replicación 16	2,60	0,0173		Replicación 56	1,96	0,2631
Replicación 17	2,87	0,1577		Replicación 57	2,28	0,0351
Replicación 18	2,29	0,0327		Replicación 58	2,79	0,1017
Replicación 19	2,31	0,0269		Replicación 59	2,35	0,0158
Replicación 20	2,38	0,0078		Replicación 60	2,05	0,1749
Replicación 21	2,62	0,0221		Replicación 61	2,42	0,0027
Replicación 22	2,43	0,0014		Replicación 62	2,34	0,0164
Replicación 23	2,59	0,0142		Replicación 63	2,59	0,0137
Replicación 24	2,17	0,0901		Replicación 64	2,94	0,2230
Replicación 25	2,32	0,0227		Replicación 65	2,24	0,0539
Replicación 26	2,36	0,0132		Replicación 66	1,97	0,2524
Replicación 27	2,43	0,0015		Replicación 67	2,40	0,0050
Replicación 28	3,12	0,4225		Replicación 68	2,38	0,0089
Replicación 29	2,18	0,0844		Replicación 69	2,76	0,0848
Replicación 30	2,84	0,1333		Replicación 70	2,48	0,0002
Replicación 31	2,06	0,1709		Replicación 71	2,72	0,0609
Replicación 32	2,65	0,0312		Replicación 72	2,50	0,0006
Replicación 33	2,78	0,0970		Replicación 73	2,12	0,1217
Replicación 34	1,87	0,3611		Replicación 74	2,39	0,0073
Replicación 35	2,19	0,0809		Replicación 75	2,73	0,0693
Replicación 36	2,69	0,0490		Replicación 76	2,02	0,2031
Replicación 37	1,94	0,2825		Replicación 77	2,80	0,1058
Replicación 38	3,20	0,5355		Replicación 78	2,86	0,1519
Replicación 39	2,21	0,0683		Replicación 79	2,57	0,0091
Replicación 40	2,48	0,0001		Replicación 80	2,22	0,0641
				Suma de las diferencias al cuadrado		9,7360

Se ha desarrollado una macro de SPSS® para el cálculo de errores típicos de las diferencias. El cuadro 10.2 presenta la sintaxis de SPSS® para ejecutar esta macro.

Cuadro 10.2. Sintaxis de SPSS® para calcular los errores típicos de las diferencias

```
* IMPORTAR LA MACRO.
include file 'C:\PISA\macros\mcr_SE_dif.sps'.

* EJECUTAR LA MACRO.
difNOpv nrep = 80/
        dep = bsmj/
        stat = mean/
        within = cnt/
        compare = st03q01/
        categ = 12/
        wgt = w_fstuwt/
        rwgt = w_fstr/
        cons = 0.05/
        infile = 'c:\PISA\Data2003\DEU.sav'/.
```

Además de los argumentos comunes a todas las macros de SPSS®, es preciso especificar otros cuatro argumentos:

- El argumento DEP informa a la macro sobre la variable numérica de la que se calculará una media o una desviación típica por cada valor de una variable categórica. En el ejemplo, DEP es igual a BSMJ.
- El argumento COMPARE especifica las variables categóricas en las que se basarán los contrastes.
- El argumento CATEG especifica los valores de las variables categóricas para las que se requieren contrastes. Como el sexo tiene sólo dos categorías, llamadas 1 y 2, la afirmación CATEG se establece como «12/». No debería haber espacios ni otros caracteres entre las categorías del argumento CATEG y la barra (/) debería seguir directamente a la última categoría. Si una variable categórica tiene cuatro categorías, y si estas cuatro categorías se especifican en la afirmación CATEGORY (como CATEG = 1234/), la macro calculará el error típico de la diferencia entre:
  - la categoría 1 y la categoría 2;
  - la categoría 1 y la categoría 3;
  - la categoría 1 y la categoría 4;
  - la categoría 2 y la categoría 3;
  - la categoría 2 y la categoría 4;
  - la categoría 3 y la categoría 4.
  - Esta macro tiene algunas limitaciones:
  - puede especificarse un máximo de 9 categorías de la variable «compare»;
  - la variable categórica «compare» debería definirse como numérica, por lo que las variables cadena deberían convertirse en numéricas;
  - los valores de las categorías deberían ser de un solo dígito. Los valores de dos dígitos deben recodificarse primero.
- El argumento STAT especifica el estadístico buscado. Véase el capítulo 6 [tabla 6.4] para los estadísticos disponibles.

Tabla 10.4. Estructura del archivo de resultados a partir del cuadro 10.2

CNT	STAT	SE
DEU	2,47	0,6977

Merece la pena hacer notar que, para las variables dicotómicas, el error típico de la diferencia también puede calcularse mediante un modelo de regresión.

Cuadro 10.3. Una sintaxis alternativa de SPSS® para calcular el error típico de una diferencia para una variable dicotómica

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if (cnt='DEU' & not missing(st03q01)).
compute gender=0.
if (st03q01=1) gender=1.
sort cases by cnt.
save outfile='c:\pisa\Data2003\DEU.sav'.

* IMPORTAR LA MACRO.
include file='C:\PISA\macros\mcr_SE_reg.sps'.

* EJECUTAR MACRO.
REGnoPV nrep = 80/
        ind = gender/
        dep = bsmj/
        grp = cnt/
        wgt = w_fstuwt/
        rwgt = w_fstr/
        cons = 0.05/
        infile = 'c:\pisa\Data2003\DEU.sav'/.
```

Tabla 10.5. Estructura del archivo de salida a partir del cuadro 10.3

CNT	CLASS	STAT	SE
DEU	b_0	50,58	0,686
DEU	gender	2,47	0,698

La estimación de la diferencia y su respectivo error típico son iguales a la estimación del coeficiente de regresión y su error típico. Para las variables categóricas politómicas, el uso de la macro de regresión exigiría la recodificación de las variables categóricas en $h - 1$ variables dicotómicas, donde h es igual al número de categorías. Además, la macro de regresión comparará cada categoría con la categoría de referencia (en la tabla de arriba, el grupo de referencia son los chicos), mientras que la macro DIFNOPV proporcionará todos los contrastes.

El error típico de una diferencia con valores plausibles

El procedimiento para calcular el error típico de una diferencia que implique a los valores plausibles consiste en:

- Utilizar cada valor plausible y, para el peso final y los 80 pesos replicados, el estadístico buscado, por ejemplo, una media, debe calcularse por cada valor de las variables categóri-

cas.

- Calcular por contraste, por valor plausible y por peso replicado la diferencia entre las dos categorías. Habrá 405 estimaciones de diferencias; la tabla 10.6 presenta su estructura.
- Calcular una estimación de diferencia final igual a la media de las cinco estimaciones de diferencias.
- Calcular, por cada valor plausible, la varianza muestral comparando la estimación de la diferencia final con las 80 estimaciones replicadas.
- Calcular una varianza muestral final igual a la media de las cinco varianzas muestrales.
- Calcular la varianza de imputación, también llamada *varianza del error de medida.*
- Combinar la varianza muestral y la varianza de imputación para obtener la varianza de error final.
- Calcular el error típico que será igual a la raíz cuadrada de la varianza de error.

Tabla 10.6. Estimaciones de diferencias según el género y sus respectivas varianzas muestrales en la escala de matemáticas

Peso	PV1	PV2	PV3	PV4	PV5
Final	−8,94	−9,40	−8,96	−7,46	−10,12
Replicación 1	−9,64	−10,05	−10,29	−8,74	−11,45
.....	.....	.....	.....	.....	.....
Replicación 80	−8,56	−8,52	−8,85	−7,70	−9,84
Varianza muestral	$(4,11)^2$	$(4,36)^2$	$(4,10)^2$	$(4,31)^2$	$(4,28)^2$

Se ha desarrollado una macro de SPSS® para calcular los errores típicos de las diferencias donde intervengan valores plausibles. El cuadro 10.4 proporciona la sintaxis de SPSS®. En este ejemplo, se calcula el error típico de la diferencia entre el rendimiento de los chicos y de las chicas en la escala combinada de competencia matemática.

Cuadro 10.4. Sintaxis de SPSS® para calcular errores típicos de las diferencias donde intervienen valores plausibles

```
* IMPORTAR LA MACRO.
include file 'C:\PISA\macros\mcr_SE_dif_PV.sps'.

* EJECUTAR LA MACRO.
dif_pv nrep = 80/
       dep = math/
       stat = mean/
       within = cnt/
       compare = st03q01/
       categ = 12/
       wgt = w_fstuwt/
       rwgt = w_fstr/
       cons = 0.05/
       infile = 'c:\pisa\Data2003\DEU.sav'/.
```

La tabla 10.7 presenta la estructura del archivo de salida.

Tabla 10.7. Estructura del archivo de resultados a partir del cuadro 10.4

CNT	STAT	SE
DEU	−8,98	4,37

Como la razón entre la estimación de la diferencia y su respectivo error típico es mayor que 1,96, se rechaza la hipótesis nula. Así pues, la media del rendimiento de las chicas es inferior a la del rendimiento de los chicos en Alemania. Asimismo, merece la pena destacar que estos resultados podrían obtenerse también mediante la macro de regresión para valores plausibles.

La tabla 10.8 proporciona las estimaciones de diferencias según el género para todos los países de PISA 2003, así como los errores típicos sin sesgo y los errores típicos con sesgo.

Tabla 10.8. Diferencias según el género en la escala de matemáticas, con errores típicos sin sesgo y con sesgo

País	Diferencia entre las medias	Error típico no sesgado	Error típico sesgado		País	Diferencia entre las medias	Error típico no sesgado	Error típico sesgado
AUS	−5,34	3,75	4,04		KOR	−23,41	6,77	6,90
AUT	−7,57	4,40	5,59		LIE	−28,84	10,92	9,58
BEL	−7,51	4,81	4,69		LUX	−17,17	2,81	2,40
BRA	−16,26	4,06	7,49		LVA	−2,81	3,97	5,97
CAN	−11,17	2,13	2,78		MAC	−21,26	5,83	5,83
CHE	−16,63	4,87	5,98		MEX	−10,90	3,94	5,91
CZE	−14,97	5,08	6,11		NLD	−5,12	4,29	5,36
DEU	−8,98	4,37	5,59		NOR	−6,22	3,21	4,04
DNK	−16,58	3,20	4,50		NZL	−14,48	3,90	4,23
ESP	−8,86	2,98	4,02		POL	−5,59	3,14	4,18
FIN	−7,41	2,67	3,24		PRT	−12,25	3,31	5,41
FRA	−8,51	4,15	4,60		RUS	−10,12	4,36	6,75
GBR	−6,66	4,90	4,84		SVK	−18,66	3,65	5,30
GRC	−19,40	3,63	6,11		SWE	−6,53	3,27	4,30
HKG	−4,06	6,64	7,96		THA	4,02	4,24	5,22
HUN	−7,79	3,54	4,69		TUN	−12,17	2,51	4,01
IDN	−3,34	3,39	6,02		TUR	−15,13	6,16	10,33
IRE	−14,81	4,19	4,54		URY	−12,09	4,15	5,51
ISL	15,41	3,46	3,15		USA	−6,25	2,89	4,65
ITA	−17,83	5,89	5,96		YUG	−1,21	4,36	6,14
JPN	−8,42	5,89	7,04					

En casi todos los países, el error típico sin sesgo es menor que el error típico con sesgo, lo que refleja una covarianza positiva entre las dos distribuciones muestrales. En algunos países, la diferencia entre los dos errores típicos es pequeña, pero es considerable en otros, como Brasil, Grecia, Indonesia y Turquía.

Comparaciones múltiples

En el capítulo 3, se advirtió que toda inferencia estadística se asocia con lo que suele llamarse un *error de tipo I*. Este error representa la probabilidad de rechazar equivocadamente una hipótesis nula que sea cierta.

Supongamos que en la población no hay diferencia en el rendimiento de chicos y chicas en cuanto a competencia lectora. Se toma una muestra y se calcula la diferencia según el género. Como esta diferencia se basa en una muestra, debe calcularse un error típico de la diferencia. Si la diferencia tipificada, esto es, la diferencia según el género dividida por su error típico, es menor que −1,96 o mayor que 1,96, esa diferencia se consideraría significativa. De hecho, existen 5 posibilidades entre 100 de observar una diferencia tipificada menor que −1,96 o mayor que 1,96 y que la hipótesis nula siga siendo cierta. Dicho de otro modo, hay 5 posibilidades entre 100 de rechazar la hipótesis nula, cuando no existe verdadera diferencia entre géneros en la población.

Si en la encuesta internacional participan 100 países y se calcula la diferencia entre géneros para cada uno de ellos, estadísticamente se espera que 5 de las 100 diferencias sean significativas, aun cuando no existen verdaderas diferencias a nivel de la población.

Para cada país, el error de tipo I se establece en 0,05. Para dos países, como los países son muestras independientes, la probabilidad de no cometer un error de tipo I, es decir, aceptar ambas hipótesis nulas, es ahora igual a 0,9025 (0,95 por 0,95). La tabla 10.9 presenta la tabulación cruzada de las distintas probabilidades.

Tabla 10.9. La tabulación cruzada de las distintas probabilidades

		País A	
		0,05	0,95
País B	0,05	0,0025	0,0475
	0,95	0,0475	0,9025

Esta cuestión estadística se amplifica aún más para tablas de comparaciones múltiples de rendimiento. Supongamos que necesitamos comparar las medias de tres países. Esto llevará tres pruebas: el país A frente al país B, el país A frente al país C y el país B frente al país C. Por tanto, la probabilidad de no cometer un error de tipo I es igual a:

$$(1-\alpha)\cdot(1-\alpha)\cdot(1-\alpha)=(1-\alpha)^3 .$$

En términos generales, si se ponen a prueba X comparaciones, la probabilidad de no cometer un error de tipo I es igual a:

$$(1-\alpha)^X .$$

Dunn (1961) desarrolló un procedimiento general que resulta apropiado para poner a prueba un conjunto de hipótesis *a priori*, controlando la probabilidad de cometer un error de tipo I. Consiste en ajustar el valor α, dividiéndolo por el número de comparaciones. El resultado se utiliza como valor crítico para cada comparación.

En el caso de tres comparaciones, el valor crítico para una $\alpha = 0{,}05$ será, por tanto, igual a 2,24 en vez de 1,96. De hecho,

$$\frac{0{,}05}{3} = 0{,}01666\,.$$

Puesto que la probabilidad se reparte entre las dos colas de la distribución muestral, se debe encontrar la puntuación z que corresponda a la proporción acumulada de 0,008333. Una consulta de la función acumulada de la distribución normal tipificada proporcionará el valor −2,24.

Sin embargo, el investigador aún tiene que decidir cuántas comparaciones están implicadas. En PISA, se decidió que no se aplicaría ninguna corrección del valor crítico, excepto en tablas de comparaciones múltiples. Es más, en muchos casos, los lectores están sobre todo interesados en averiguar si un valor dado en un país determinado es distinto de un segundo valor en el mismo país o en otro; por ejemplo, si las mujeres de cierto país tienen mejor rendimiento que los varones del mismo país. Por lo tanto, como solo se lleva a cabo una prueba cada vez, no es necesario ningún ajuste.

Por otra parte, con tablas de comparaciones múltiples, el lector está interesado en comparar el rendimiento de un país con todos los demás países. Por ejemplo, si se desea comparar el rendimiento del país 1 con los demás, tendremos las siguientes comparaciones: el país 1 con el país 2, el país 1 con el país 3 y el país 1 con el país L. Así pues, el ajuste se basará en $L - 1$ comparaciones.

En PISA 2003, los informes iniciales publicaron los resultados de 40 países y el valor crítico estuvo basado en 39 comparaciones y fue igual a 3,2272. Puesto que en PISA 2003 participaron más países, este valor crítico es ligeramente más alto que el valor crítico para PISA 2000[1].

Conclusiones

Este capítulo ha estado dedicado al cálculo de los errores típicos de las diferencias. Después de una descripción de las cuestiones estadísticas relacionadas con tales estimaciones, se han presentado los distintos pasos para calcular dichos errores típicos. También se han descrito las macros de SPSS® que facilitan tales cálculos.

Se ha establecido claramente que cualquier comparación entre países no exige la estimación de la covarianza, por lo que pueden usarse las macros de los capítulos anteriores y pueden combinarse los errores típicos. Sin embargo, se recomienda encarecidamente estimar la covarianza entre las distribuciones muestrales para cualquier comparación dentro de un país. Para estas estimaciones, se han introducido nuevas macros.

Por último, se ha comentado la corrección del valor crítico para comparaciones múltiples.

---

[1] El valor crítico en las comparaciones múltiples de PISA 2000 fue 3,144.

Capítulo 11

Media de la OCDE y total de la OCDE

Introducción

En todos los informes iniciales y temáticos de PISA, la OCDE proporciona los resultados de cada país, pero también dos estimaciones agregadas adicionales: la media de la OCDE y el total de la OCDE.

La media de la OCDE, a veces también llamada *media de los países*, es la media de los valores de todos los países de la OCDE para los que existen o pueden estimarse datos. La media de la OCDE puede utilizarse para ver cómo un país se compara en un indicador dado con un país representativo de la OCDE. La media de la OCDE no tiene en cuenta el tamaño absoluto de la población en cada país, es decir, cada país contribuye igualmente a la media. La contribución del país más pequeño de la OCDE, Luxemburgo, es equivalente a la de uno de los mayores, Estados Unidos.

El total de la OCDE considera a todos los países de la OCDE una sola entidad, a la que cada país contribuye en proporción al número de jóvenes de 15 años que estudian en sus centros. Ilustra la comparación de un país con el conjunto de la OCDE.

En PISA 2003, participaron todos los países de la OCDE, así como varios países asociados. Sin embargo, es posible que para un ciclo particular no haya datos disponibles para indicadores específicos en uno o varios países de la OCDE. Por lo tanto, los investigadores deberían tener en cuenta que los términos *media de la OCDE* y *total de la OCDE* se refieren a los países incluidos en las respectivas comparaciones para cada ciclo y para una comparación determinada.

Para estadísticos simples como una media o un porcentaje, los estadísticos de la media de la OCDE y el total de la OCDE y sus respectivos errores típicos pueden calcularse matemáticamente. Si participaron *C* países de la OCDE, la media promedio y su respectiva varianza muestral son iguales a:

$$\hat{\mu} = \frac{\sum_{i=1}^{C} \hat{\mu}_i}{C} \quad \text{y} \quad \sigma^2_{(\hat{\mu})} = \frac{\sum_{i=1}^{C} \sigma^2_{(\hat{\mu}_i)}}{C}.$$

La media total de la OCDE y su respectiva varianza muestral son iguales a:

$$\hat{\mu} = \frac{\sum_{i=1}^{C} w_i \hat{\mu}_i}{\sum_{i=1}^{C} w_i} \quad \text{y} \quad \sigma^2_{(\hat{\mu})} = \frac{\sum_{i=1}^{C} w_i^2 \sigma^2_{(\hat{\mu}_i)}}{\left[\sum_{i=1}^{C} w_i\right]^2},$$

donde w_i es la suma de los pesos finales de los alumnos para un país determinado.

Si bien estas fórmulas pueden usarse para calcular una media o un porcentaje, no pueden utilizarse para la mayoría del resto de estadísticos. Éstos sólo pueden obtenerse directamente a partir del conjunto de datos.

Recodificación de la base de datos para la estimación del total de la OCDE y de la media de la OCDE

Como ya se indicó en el capítulo 3, la suma de los pesos finales de los alumnos por cada país es una estimación de la población de alumnos de 15 años en dicho país. Por lo tanto, el estadístico del total de la OCDE puede obtenerse fácilmente borrando los datos de los países asociados. Luego, el estadístico se calcula usando la variable `oecd` como variable de agrupación (la variable de la `oecd` es una constante, ya que los países asociados se han borrado del archivo) en lugar del país (CNT). El error típico se obtiene como de costumbre, utilizando las 80 replicaciones. El cuadro 11.1 proporciona la sintaxis de SPSS® para el cálculo del rendimiento en matemáticas según el género para el total de la OCDE y la tabla 11.1 presenta la salida del procedimiento.

Cuadro 11.1. Sintaxis de SPSS® para calcular el total de la OCDE para el rendimiento en matemáticas según el género

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
compute oecd=0.
if
 (cnt='AUS'|cnt='AUT'|cnt='BEL'|cnt='CAN'|cnt='CZE'|cnt='DNK'|
  cnt='FIN'|cnt='FRA'|cnt='DEU'|cnt='GRC'|cnt='HUN'|cnt='ISL'|
  cnt='IRL'|cnt='ITA'|cnt='JPN'|cnt='KOR'|cnt='LUX'|cnt='MEX'|
  cnt='NZL'|cnt='NOR'|cnt='POL'|cnt='PRT'|cnt='ESP'|cnt='SWE'|
  cnt='CHE'|cnt='GBR'|cnt='USA'|cnt='NLD'|cnt='SVK'|cnt='TUR') oecd=1.
select if (oecd=1 & not missing(st03q01)).
save outfile='c:\pisa\Data2003\OECD.sav'.

* IMPORTAR LA MACRO.
include file 'c:\pisa\macros\mcr_SE_pv.sps'.

* EJECUTAR LA MACRO.
PV  nrep = 80/
    stat = mean/
    dep = math/
    grp = oecd st03q01/
    wgt = w_fstuwt/
    rwgt = w_fstr/
    cons = 0.05/
    infile = 'c:\pisa\Data2003\OECD.sav'/.
```

Tabla 11.1. Estructura del archivo de salida a partir del cuadro 11.1

ST03Q01	STAT	SE
1	483,93	1,25
2	494,04	1,32

La media de la OCDE exige un paso adicional. Es necesario recodificar los pesos finales de los alumnos, de modo que la suma de los pesos finales de los alumnos por cada país sea igual a una constante, por ejemplo 1000. Esto puede implementarse con facilidad mediante el procedimiento descrito en el cuadro 11.2.[1] La tabla 11.2 presenta la salida del procedimiento. Las variables de agrupación son `oecd` y `st03q01`, como en el ejemplo de arriba, para calcular el total de la OCDE y su error típico. Los dos argumentos de pesos son diferentes. El peso total de los alumnos (`w_fstuwt`) se sustituye por el peso de la variable `senate`. Los pesos replicados se han

transformado mediante la misma transformación lineal que el peso total de los alumnos y ahora se llaman s_fstr1 a s_fstr80.

Cuadro 11.2. Sintaxis de SPSS® para calcular la media de la OCDE para el rendimiento en matemáticas según el género

```
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
compute oecd=0.
If
 (cnt='AUS'|cnt='AUT'|cnt='BEL'|cnt='CAN'|cnt='CZE'|cnt='DNK'|
  cnt='FIN'|cnt='FRA'|cnt='DEU'|cnt='GRC'|cnt='HUN'|cnt='ISL'|
  cnt='IRL'|cnt='ITA'|cnt='JPN'|cnt='KOR'|cnt='LUX'|cnt='MEX'|
  cnt='NZL'|cnt='NOR'|cnt='POL'|cnt='PRT'|cnt='ESP'|cnt='SWE'|
  cnt='CHE'|cnt='GBR'|cnt='USA'|cnt='NLD'|cnt='SVK'|cnt='TUR') oecd=1.
select if oecd=1.
sort cases by cnt.
aggregate outfile='c:\temp\population.sav' /break=cnt /pop=sum(w_fstuwt).
match files file=* /table='c:\temp\population.sav'/by cnt.

compute senate=(w_fstuwt/pop)*1000.

define mcr ().
!do !r=1 !to 80.
compute !concat('s_fstr',!r)=(!concat('w_fstr',!r)/pop)*1000.
!doend.
!enddefine.
MCR.
save outfile='c:\pisa\Data2003\OECD.sav'.

weight by senate.
fre cnt.

*=================================================================.
* IMPORTAR LA MACRO.
include file 'c:\pisa\macros\mcr_SE_pv.sps'.

* EJECUTAR LA MACRO.
PV  nrep = 80/
    stat = mean/
    dep = math/
    grp = oecd st03q01/
    wgt = senate/
    rwgt = s_fstr/
    cons = 0.05/
    infile = 'c:\pisa\Data2003\OECD.sav'/.
```

Tabla 11.2. Estructura del archivo de salida a partir del cuadro 11.2

ST03Q01	STAT	SE
1	494,41	0,76
2	505,53	0,75

Merece la pena hacer notar que el error típico es más alto para el total de la OCDE que para la media de la OCDE. En el caso del total de la OCDE, el 40% de los datos provienen de sólo dos países (Estados Unidos y Japón), los cuales no tienen grandes tamaños de muestra en comparación con otros países de la OCDE.

Duplicación de los datos para evitar tres ejecuciones del procedimiento

Si un investigador está interesado en las estimaciones de los países, así como en el total de la OCDE y la media de la OCDE, se necesita ejecutar el procedimiento tres veces: una para las estimaciones por país, otra para la estimación del total de la OCDE y una tercera para la estimación de la media de la OCDE.

Con objeto de evitar tales repeticiones, es posible duplicar tres veces los datos de los países de la OCDE, de manera que el procedimiento proporcione directamente las estimaciones para cada país, así como las estimaciones del total de la OCDE y de la media de la OCDE.

El cuadro 11.3 presenta la sintaxis de SPSS® para generar tales conjuntos de datos. Consiste en los pasos siguientes:

- Se calcula una nueva variable categórica, llamada `oecd`, que separa los países de la OCDE de los países asociados. Un valor de 1 para esta variable designa a los países de la OCDE, mientras que un valor de 4 designa a los países asociados. Se crea una segunda variable alfanumérica, llamada `country`, y cuyo contenido es igual a `cnt`.
- Se seleccionan los casos correspondientes a los países de la OCDE y se guardan en el archivo TEMP2. La variable `oecd` se fija en 2 y la variable `country` se fija en TOT.
- En el archivo TEMP2, se calcula la suma de los pesos finales de los alumnos por cada país mediante el procedimiento descrito en el cuadro 11.2. Los pesos finales se transforman de tal modo que la suma por cada país sea igual a 1000. Se aplica la misma transformación lineal a las 80 replicaciones. La variable `cnt` se fija en AVG y la variable `oecd` se fija en 3. Las nuevas variables de peso sobreescriben a las antiguas. Estos datos nuevos se guardan en el archivo TEMP3.
- Después de ordenar TEMP2 y TEMP3, los tres archivos temporales se fusionan y se guardan en un archivo final de datos de SPSS®.

Las macros de SPSS® presentadas en los capítulos anteriores pueden aplicarse a este nuevo archivo de datos. Las variables de agrupación son ahora `oecd` y `country`, en lugar de `cnt`. El archivo de resultados consistirá en 43 filas. Las primeras 30 filas serán los resultados de los países de la OCDE. Las siguientes dos filas presentarán las estimaciones del total de la OCDE y la media de la OCDE. Finalmente, las últimas 11 filas presentarán las estimaciones de los países asociados.

Comparaciones entre las estimaciones de la media de la OCDE o del total de la OCDE y la estimación de un país

Sólo los países de la OCDE que participen plenamente contribuyen a la estimación de la media de la OCDE y del total de la OCDE y sus respectivos errores típicos. Por lo tanto, el valor esperado de la covarianza entre la varianza muestral de un país y la varianza muestral agregada de la OCDE diferirán de 0 si los valores del país se incluyen en los valores agregados de la OCDE, porque ninguna de las dos es independiente. De hecho, si la varianza muestral de un país aumenta, la varianza muestral agregada de la OCDE también aumentará.

Si un investigador quiere poner a prueba la hipótesis nula entre un país de la OCDE y la estimación agregada de la OCDE, debería estimarse la covarianza, como se explica en el capítulo 10. Puesto que se espera que la covarianza sea positiva, la estimación correcta del error típico debería ser menor que el error típico obtenido a partir de las fórmulas.

Cuadro 11.3. Sintaxis de SPSS® para la creación de un conjunto triplicado de datos que permita en una sola ejecución calcular el total de la OCDE y la media de la OCDE

```
* PAÍSES.
*------------------------.
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
compute OECD=4.
if (cnt='AUS'|cnt='AUT'|cnt='BEL'|cnt='CAN'|cnt='CZE'|cnt='DNK'|cnt='FIN'
|cnt='FRA'|cnt='DEU'|cnt='GRC'|cnt='HUN'|cnt='ISL'|cnt='IRL'|cnt='ITA'
|cnt='JPN'|cnt='KOR'|cnt='LUX'|cnt='MEX'|cnt='NZL'|cnt='NOR'|cnt='POL'
|cnt='PRT'|cnt='ESP'|cnt='SWE'|cnt='CHE'|cnt='GBR'|cnt='USA'|cnt='NLD'
|cnt='SVK'|cnt='TUR') OECD=1.
save outfile='C:\TEMP\TEMP1.SAV'.

* TOTAL DE LA OCDE.
*------------------------.
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if
(cnt='AUS'|cnt='AUT'|cnt='BEL'|cnt='CAN'|cnt='CZE'|cnt='DNK'|cnt='FIN'
|cnt='FRA'|cnt='DEU'|cnt='GRC'|cnt='HUN'|cnt='ISL'|cnt='IRL'|cnt='ITA'
|cnt='JPN'|cnt='KOR'|cnt='LUX'|cnt='MEX'|cnt='NZL'|cnt='NOR'|cnt='POL'
|cnt='PRT'|cnt='ESP'|cnt='SWE'|cnt='CHE'|cnt='GBR'|cnt='USA'|cnt='NLD'
|cnt='SVK'|cnt='TUR').
compute OECD=2.
COMPUTE CNT='TOT'.
save outfile='C:\TEMP\TEMP2.SAV'.

* MEDIA DE LA OCDE.
*------------------------.
get file 'C:\PISA\Data2003\INT_stui_2003.sav'.
select if
(cnt='AUS'|cnt='AUT'|cnt='BEL'|cnt='CAN'|cnt='CZE'|cnt='DNK'|cnt='FIN'
|cnt='FRA'|cnt='DEU'|cnt='GRC'|cnt='HUN'|cnt='ISL'|cnt='IRL'|cnt='ITA'
|cnt='JPN'|cnt='KOR'|cnt='LUX'|cnt='MEX'|cnt='NZL'|cnt='NOR'|cnt='POL'
|cnt='PRT'|cnt='ESP'|cnt='SWE'|cnt='CHE'|cnt='GBR'|cnt='USA'|cnt='NLD'
|cnt='SVK'|cnt='TUR').
compute OECD=3.
sort cases by cnt.
aggregate outfile='c:\temp\population.sav' /break=cnt /pop=sum(w_fstuwt).
match files file=* /table='c:\temp\population.sav'/by cnt.
compute w_fstuwt=(w_fstuwt/pop)*1000.
define mcr ().
!do !r=1 !to 80.
compute !concat('w_fstr',!r)=(!concat('w_fstr',!r)/pop)*1000.
!doend.
!enddefine.
mcr.
weight by w_fstuwt.
fre cnt.
COMPUTE CNT='AVG'.
save outfile='C:\TEMP\TEMP3.SAV'.
                                                                    ...
```

```
* CONCATENAR LOS ARCHIVOS TEMPORALES.
*-----------------------------------.
add files file='C:\TEMP\TEMP1.SAV' /file='C:\TEMP\TEMP2.SAV'
/file='C:\TEMP\TEMP3.SAV'.
sort cases by oecd cnt.
FORMATS oecd (f1.0).
save outfile='c:\pisa\data2003\ALL3.sav'.
cros cnt by oecd.

* IMPORTAR LA MACRO.
include file 'c:\pisa\macros\mcr_SE_pv.sps'.

* EJECUTAR LA MACRO.
PV  nrep = 80/
    stat = mean/
    dep = math/
    grp = oecd cnt/
    wgt = w_fstuwt/
    rwgt = w_fstr/
    cons = 0.05/
    infile = 'c:\pisa\data2003\ALL3.sav'/.
```

Puesto que los países asociados no contribuyen en absoluto a las estimaciones agregadas de la OCDE, no es necesario estimar la covarianza. El error típico de la diferencia puede obtenerse directamente a partir del error típico del país y el error típico agregado.

La tabla 11.3 presenta:

- el rendimiento medio de los países en matemáticas, así como la media de la OCDE y el total de la OCDE;
- el error típico de estas estimaciones;
- la diferencia entre el país y el total de la OCDE;
- el error típico de esta diferencia, mediante la fórmula proporcionada en el capítulo 10, es decir, sin una estimación de la covarianza;
- el error típico de esta diferencia, mediante la replicación, es decir, con una estimación de la covarianza;
- la diferencia entre el país y la media de la OCDE;
- el error típico de esta diferencia, mediante la replicación, es decir, con una estimación de la covarianza.

Las estimaciones correctas del error típico están en negrita. Las diferencias entre las estimaciones con y sin sesgo para los países de la OCDE no son muy grandes, excepto en el caso de Estados Unidos y Alemania para el total de la OCDE.

Tampoco las diferencias de los países asociados son muy grandes. Como la covarianza esperada para los países asociados es 0, ninguno de los dos errores típicos está sesgado por término medio. Sin embargo, se recomienda emplear el error típico directamente obtenido con la fórmula.

Tabla 11.3. Medias del rendimiento de los países en matemáticas y sus respectivos errores típicos, diferencia de los países con el total de la OCDE y con la media de la OCDE, y sus respectivos errores típicos sin y con estimación de la covarianza

	País		Total OCDE			Media OCDE		
	Media	ET	DIF	ET sin	ET con	DIF	ET sin	ET con
AUS	524,27	2,15	35,27	2,40	**2,12**	24,27	2,24	**2,03**
AUT	505,61	3,27	16,61	3,44	**3,49**	5,61	3,33	**3,27**
BEL	529,29	2,29	40,29	2,52	**2,42**	29,29	2,37	**2,23**
CAN	532,49	1,82	43,49	2,11	**2,08**	32,49	1,92	**1,96**
CHE	526,55	3,38	37,56	3,55	**3,48**	26,55	3,44	**3,38**
CZE	516,46	3,55	27,46	3,70	**3,90**	16,46	3,60	**3,51**
DEU	502,99	3,32	13,99	3,49	**3,42**	2,99	3,38	**3,30**
DNK	514,29	2,74	25,29	2,95	**2,99**	14,29	2,82	**2,67**
ESP	485,11	2,41	−3,89	2,64	**2,60**	−14,89	2,49	**2,47**
FIN	544,29	1,87	55,29	2,15	**2,07**	44,29	1,97	**1,91**
FRA	510,80	2,50	21,80	2,72	**2,45**	10,80	2,58	**2,46**
GBR	508,26	2,43	19,26	2,65	**2,41**	8,26	2,51	**2,39**
GRC	444,91	3,90	−44,09	4,05	**3,94**	−55,09	3,95	**3,81**
HUN	490,01	2,84	1,02	3,03	**3,20**	−9,99	2,91	**2,95**
IRL	502,84	2,45	13,84	2,67	**2,56**	2,84	2,53	**2,41**
ISL	515,11	1,42	26,11	1,78	**1,78**	15,11	1,56	**1,48**
ITA	465,66	3,08	−23,33	3,26	**3,11**	−34,34	3,14	**2,98**
JPN	534,14	4,02	45,14	4,16	**3,88**	34,14	4,06	**3,94**
KOR	542,23	3,24	53,23	3,41	**3,34**	42,23	3,30	**3,16**
LUX	493,21	0,97	4,21	1,45	**1,48**	−6,79	1,16	**1,20**
MEX	385,22	3,64	−103,78	3,80	**3,55**	−114,78	3,70	**3,64**
NLD	537,82	3,13	48,83	3,31	**3,19**	37,82	3,19	**3,10**
NOR	495,19	2,38	6,19	2,61	**2,69**	−4,81	2,46	**2,41**
NZL	523,49	2,26	34,49	2,50	**2,41**	23,49	2,34	**2,31**
POL	490,24	2,50	1,24	2,72	**2,82**	−9,76	2,58	**2,54**
PRT	466,02	3,40	−22,98	3,57	**3,30**	−33,98	3,46	**3,23**
SVK	498,18	3,35	9,19	3,51	**3,46**	−1,82	3,41	**3,31**
SWE	509,05	2,56	20,05	2,77	**2,48**	9,05	2,64	**2,40**
TUR	423,42	6,74	−65,58	6,82	**6,48**	−76,58	6,77	**6,46**
USA	482,88	2,95	−6,11	3,14	**2,38**	−17,12	3,02	**2,90**
TOT	489,00	1,07						
AVE	500,00	0,63						
BRA	356,02	4,83	−132,98	**4,95**	4,89	−143,98	**4,87**	4,77
HKG	550,38	4,54	61,39	**4,66**	4,80	50,38	**4,58**	4,68
IDN	360,16	3,91	−128,84	**4,05**	4,03	−139,84	**3,96**	3,88
LIE	535,80	4,12	46,80	**4,26**	4,16	35,80	**4,17**	4,13
LVA	483,37	3,69	−5,62	**3,84**	3,88	−16,62	**3,74**	3,76
MAC	527,27	2,89	38,27	**3,08**	3,13	27,27	**2,95**	2,85
RUS	468,41	4,20	−20,59	**4,33**	4,47	−31,59	**4,24**	4,33
THA	416,98	3,00	−72,02	**3,18**	3,38	−83,02	**3,06**	3,20
TUN	358,73	2,54	−130,26	**2,75**	2,55	−141,27	**2,61**	2,57
URY	422,20	3,29	−66,80	**3,46**	3,41	−77,80	**3,35**	3,30
YUG	436,87	3,75	−52,13	**3,90**	3,85	−63,13	**3,81**	3,78

Conclusiones

Este capítulo ha estado dedicado a los conceptos de *total de la OCDE* y *media de la OCDE*. Para estadísticos simples como una media o un porcentaje, estas estimaciones agregadas y sus res-

pectivos errores típicos pueden obtenerse directamente a partir de las estimaciones individuales de los países.

De todas formas, en la mayoría de los casos, la estimación y su error estándar sólo pueden calcularse a partir de la base de datos. Se ha proporcionado la sintaxis de SPSS®.

Con objeto de evitar tres ejecuciones para obtener estimaciones individuales de cada país, así como las estimaciones agregadas de la OCDE, también se ha proporcionado la sintaxis de SPSS® para crear un conjunto mayor de datos.

Por último, siguiendo las cuestiones planteadas en el capítulo anterior dedicado a las comparaciones, se trató acerca de cualquier comparación que implique a un país determinado y a una estimación agregada de la OCDE.

---

[1] Como alternativa, también puede emplearse un peso de país, CNTFAC1.

Capítulo 12

Las tendencias

Introducción

Los responsables de las políticas educativas y los investigadores necesitan información sobre cómo cambian los indicadores con el tiempo. Cuanto más largo sea el período temporal, más fiable será el indicador de tendencia. Un ejemplo sería un análisis del impacto de las reformas en el sistema educativo, donde los responsables políticos procurarían medir los cambios en un área determinada para evaluar hasta qué punto han sido eficaces sus medidas. Por ejemplo, a comienzos de la década de 1960, la mayoría de los países de la OCDE puso en marcha reformas educativas para facilitar el acceso a la educación terciaria, sobre todo mediante ayudas económicas. Un indicador del impacto de estas reformas sería calcular el porcentaje de la población con una calificación terciaria durante varios años, para mostrar cómo ha evolucionado. Calcular este indicador de tendencia es una operación estadística directa, puesto que la medida (es decir, si una persona ha terminado o no estudios terciarios) es bastante objetiva y la información se encuentra disponible, en la mayoría de los casos, al nivel de la población. Sin embargo, tales medidas pueden estar ligeramente sesgadas, a causa, por ejemplo, de distintos grados de inmigración durante el período considerado o de programas de intercambio de estudiantes, etcétera.

Por supuesto, las tendencias de un indicador particular a lo largo del tiempo necesitan una interpretación cuidadosa. Los responsables políticos deberían también tener en cuenta los cambios en el contexto económico del país, como las tasas crecientes de desempleo. Por ejemplo, un aumento en el porcentaje de personas que hayan terminado estudios terciarios no prueba necesariamente que la reforma del sistema educativo haya sido eficaz. Además, cuando se comparan los indicadores de tendencias entre varios países, es importante considerar hasta qué punto es comparable la definición del indicador de un país a otro; por ejemplo, *educación terciaria* podría significar algo distinto en cada país.

El proyecto PISA ofrece una oportunidad única de extender el cálculo de indicadores de tendencia en los resultados del sistema educativo, examinando el rendimiento de los alumnos en lectura, matemáticas y ciencias.

Para que las medidas de tendencias sean fiables, la comparabilidad de la población objetivo, los procedimientos de recogida de datos y el marco teórico deben ser constantes a lo largo del tiempo. Uno de los principales propósitos del proyecto es que puedan utilizarse los resultados de PISA como indicadores de tendencias.

PISA 2000 y PISA 2003 utilizaron los mismos marcos teóricos y los procedimientos de recogida de datos quedaron básicamente inalterados. En PISA 2000, la población objetivo se definió como los jóvenes de 15 años en el quinto curso o superior. En PISA 2003, se definió como los jóvenes de quince años en el séptimo curso o superior. En PISA 2000, sólo un porcentaje muy pequeño de alumnos de 15 años se encontraban en los cursos quinto o sexto (Austria = 0,03%, Canadá = 0,03%, República Checa = 0,06%, Alemania = 0,02%, Hungría = 0,59%, Letonia = 0,27%, Portugal = 1,25% y Rusia = 0,04%). Por lo tanto, excepto para Portugal, el cambio de la población de referencia no debería afectar significativamente a los indicadores de tendencias.

Otras cuestiones que deben incluirse en la interpretación de los indicadores de tendencias son las tasas de participación de alumnos y centros y los índices de cobertura. Una tasa de participación de centros más alta o más baja podría explicar parcialmente las diferencias observadas.

Más allá de estas precauciones preliminares, el cálculo de los indicadores de tendencias en PISA plantea dos cuestiones estadísticas:

- PISA recoge datos a partir de una muestra, por lo que cualquier estadístico debe asociarse con un error muestral. En la siguiente sección se expondrá cómo calcular dicho error muestral para un indicador de tendencia.
- En la evaluación de 2003, se incluyeron entre 20 y 30 ítems por área tomados de la evaluación de 2000, para garantizar una equiparación psicométrica. Estos ítems de anclaje se usaron para escalar las evaluaciones de PISA 2000 y PISA 2003 en una escala común. Como puede imaginarse fácilmente, si se hubieran seleccionado otros ítems de anclaje se habrían obtenido resultados ligeramente distintos para los indicadores de tendencias en el rendimiento. De ahí que cualquier comparación entre dos ciclos de PISA en cuanto al rendimiento de los alumnos exija la adición de otro componente de error: el error debido al muestreo de ítems.

Cálculo del error típico de los indicadores de tendencias en las variables que no son de rendimiento

Para cualquier país, las muestras de PISA 2000 y PISA 2003 son independientes. Por tanto, el error típico de cualquier indicador de tendencia que no esté relacionado con variables de rendimiento puede calcularse así:

$$\sigma_{(\hat{\theta}_{2003}-\hat{\theta}_{2000})} = \sqrt{\sigma^2_{(\hat{\theta}_{2003})} + \sigma^2_{(\hat{\theta}_{2000})}}$$, donde θ representa cualquier estadístico.

Sin embargo, el cálculo de una diferencia entre PISA 2000 y PISA 2003 y su error típico sólo son relevantes si las dos medidas son idénticas. Por ejemplo, en las bases de datos de PISA 2000 y PISA 2003, hay varios índices procedentes de los cuestionarios de alumnos que tienen exactamente los mismos nombres de variables (por ejemplo, HEDRES para recursos educativos familiares, BELONG para el sentimiento de pertenencia al centro por parte del alumno, etcétera). Las preguntas utilizadas para obtener estos índices no han cambiado, pero como el escalamiento se hizo de forma independiente en el 2000 y en el 2003, no hay garantía de que las medidas de 2000 y 2003 sean comparables. Además, estos índices se tipificaron a nivel de la OCDE para obtener una media de 0 y una desviación típica de 1. La tipificación del 2000 difiere de la del 2003. Por tanto, no se recomienda calcular indicadores de tendencias a partir de los índices de los cuestionarios.

En el caso de los índices de los cuestionarios de PISA, como las preguntas no se han modificado, los conceptos subyacentes son similares. Por lo tanto, los coeficientes de correlación entre estos índices y el rendimiento de los alumnos pueden compararse directamente. Ahora bien, como los parámetros de los ítems se calcularon en el 2003 sin ninguna relación con los datos de PISA 2000, las medidas de las escalas podrían ser ligeramente distintas; por ejemplo, un aumento absoluto en el sentimiento de pertenencia podría ser simplemente un resultado del escalamiento o de la tipificación, sin que haya habido ningún cambio en la actitud de los alumnos. Por las mismas razones, los coeficientes de regresión de los índices obtenidos a partir de los datos de los cuestionarios de los alumnos no pueden compararse entre 2000 y 2003.

La variable HISEI, (*Highest International Social and Economic Index*, índice internacional social y económico más alto [de entre ambos padres]) satisface todas las condiciones para el cálculo de los indicadores de tendencias. De hecho, las preguntas no se cambiaron y la transformación utilizada en las categorías de la ISCO[a] en el 2000 se utilizó en el 2003 sin ninguna modificación.

La tabla 12.1 presenta, por país, la estimación de la media de HISEI y de su error típico para PISA 2000 y PISA 2003, así como la diferencia entre las dos estimaciones, el error típico de esta diferencia y la diferencia tipificada, es decir, la diferencia dividida por su error típico.

Para Alemania (DEU), las medias de HISEI en el 2000 y en el 2003 son, respectivamente, iguales a 48,85 y 49,33. La diferencia entre estas dos recogidas de datos es, por tanto, igual a:

$$49{,}33 - 48{,}55 = 0{,}48.$$

Los errores típicos de estas estimaciones de las medias son iguales a 0,32 y 0,42. El error típico de la estimación de la diferencia es igual a:

$$\sigma_{(\hat{\theta}_{2003}-\hat{\theta}_{2000})} = \sqrt{\sigma^2_{(\hat{\theta}_{2003})} + \sigma^2_{(\hat{\theta}_{2000})}} = \sqrt{(0{,}32)^2 + (0{,}42)^2} = 0{,}53\,.$$

La diferencia tipificada, es decir, la estimación de la diferencia dividida por su error típico, es igual a:

$$\frac{0{,}48}{0{,}53} = 0{,}91\,.$$

Como la diferencia tipificada está incluida en el intervalo [−1,96; 1,96], la diferencia de la estimación de la media de HISEI entre el 2000 y el 2003 no es estadísticamente distinta de 0, con un error de tipo I de 0,05.

La tabla 12.1 muestra que la diferencia es estadísticamente distinta de 0 en 13 países: Austria, Bélgica, Brasil, la República Checa, Indonesia, Islandia, Corea, Liechtenstein, Luxemburgo, México, Tailandia, el Reino Unido y Estados Unidos.

Sería poco realista considerar estas diferencias solamente como un reflejo de los cambios sociales y económicos en estos 13 países. En un período de tres años, pueden producirse ciertos cambios, pero no explican por sí solos el tamaño de los aumentos o disminuciones observados.

[a] ISCO: International Standard Classification of Occupations («CIUO: Clasificación Internacional Uniforme de Ocupaciones»).

Tabla 12.1. Indicadores de tendencias entre PISA 2000 y PISA 2003 para HISEI por país

	PISA 2000		PISA 2003		Diferencia		
	Media	ET	Media	ET	Estimación	ET	Diferencia tipificada
AUS	52,25	(0,50)	52,59	(0,30)	0,34	(0,58)	0,59
AUT	49,72	(0,29)	47,06	(0,52)	−2,66	(0,59)	−4,49
BEL	48,95	(0,39)	50,59	(0,38)	1,65	(0,54)	3,05
BRA	43,93	(0,59)	40,12	(0,64)	−3,81	(0,87)	−4,39
CAN	52,83	(0,22)	52,58	(0,27)	−0,25	(0,35)	−0,73
CHE	49,21	(0,53)	49,30	(0,43)	0,09	(0,68)	0,13
CZE	48,31	(0,27)	50,05	(0,34)	1,74	(0,44)	3,98
DEU	48,85	(0,32)	49,33	(0,42)	0,48	(0,53)	0,91
DNK	49,73	(0,43)	49,26	(0,45)	−0,47	(0,63)	−0,75
ESP	44,99	(0,62)	44,29	(0,58)	−0,70	(0,85)	−0,83
FIN	50,00	(0,40)	50,23	(0,36)	0,23	(0,54)	0,42
FRA	48,27	(0,44)	48,66	(0,47)	0,39	(0,64)	0,61
GBR	51,26	(0,35)	49,65	(0,39)	−1,61	(0,52)	−3,07
GRC	47,76	(0,60)	46,94	(0,72)	−0,83	(0,93)	−0,88
HUN	49,53	(0,47)	48,58	(0,33)	−0,95	(0,57)	−1,65
IDN	36,38	(0,77)	33,65	(0,61)	−2,73	(0,98)	−2,77
IRL	48,43	(0,48)	48,34	(0,49)	−0,09	(0,69)	−0,13
ISL	52,73	(0,28)	53,72	(0,26)	0,99	(0,38)	2,62
ITA	47,08	(0,31)	46,83	(0,38)	−0,24	(0,49)	−0,50
JPN	50,54	(0,62)	49,98	(0,31)	−0,56	(0,69)	−0,80
KOR	42,80	(0,42)	46,32	(0,36)	3,52	(0,55)	6,36
LIE	47,46	(0,94)	50,73	(0,75)	3,27	(1,21)	2,71
LUX	44,79	(0,27)	48,17	(0,22)	3,38	(0,35)	9,76
LVA	50,15	(0,54)	50,28	(0,52)	0,13	(0,75)	0,18
MEX	42,48	(0,68)	40,12	(0,68)	−2,37	(0,96)	−2,46
NLD	50,85	(0,47)	51,26	(0,38)	0,42	(0,61)	0,68
NOR	53,91	(0,38)	54,63	(0,39)	0,72	(0,54)	1,33
NZL	52,20	(0,37)	51,46	(0,36)	−0,74	(0,51)	−1,45
POL	46,03	(0,47)	44,96	(0,34)	−1,07	(0,58)	−1,85
PRT	43,85	(0,60)	43,10	(0,54)	−0,75	(0,81)	−0,92
RUS	49,38	(0,45)	49,86	(0,38)	0,49	(0,59)	0,82
SWE	50,57	(0,39)	50,64	(0,38)	0,07	(0,55)	0,12
THA	33,02	(0,57)	36,01	(0,43)	2,99	(0,72)	4,18
USA	52,40	(0,79)	54,55	(0,37)	2,15	(0,87)	2,47

También es posible que la calidad de las muestras explique algunas de las diferencias. Puesto que la disposición de un alumno a participar positivamente se correlaciona con su currículo académico, y como por término medio los alumnos con menor rendimiento proceden de contextos sociales más bajos que los alumnos con mayor rendimiento, un aumento o una disminución en las tasas de participación de los alumnos podría afectar a la media de HISEI.

Un cambio en el porcentaje de datos perdidos para la variable HISEI sería otra explicación que puede verificarse fácilmente. En promedio, los alumnos que no aportan el dato de la profesión de sus padres tienen un rendimiento inferior. Por lo tanto, deberían esperarse características de entorno social bajo de tal modo que un aumento de datos perdidos podría asociarse con un aumento en la media de HISEI y viceversa.

La tabla 12.2 proporciona los porcentajes de datos perdidos para la variable HISEI en las bases de datos de PISA 2000 y PISA 2003. En realidad, estos resultados no confirman la hipótesis. Por ejemplo, en Estados Unidos los porcentajes de datos perdidos eran, respectivamente, de alrededor de un 14% en el 2000 y alrededor de un 6% en el 2003; las medias de HISEI eran, respectivamente, 52,40 y 54,55. En 9 de los 13 países donde las medias de HISEI difieren significativamente, o bien un aumento de la media de HISEI se asocia con una disminución del porcentaje de datos ausentes, o bien al revés. En los otros tres países —Bélgica, la República Checa y México—, la relación es coherente con la hipótesis.

Tabla 12.2. Porcentajes de datos perdidos en la variable HISEI

	PISA 2000		PISA 2003		Diferencia		
	%	ET	%	ET	Estimación	ET	Diferencia tipificada
AUS	4,15	(0,38)	7,91	(1,56)	3,76	(1,61)	2,33
AUT	2,06	(0,20)	3,62	(0,32)	1,56	(0,38)	4,13
BEL	5,02	(0,45)	6,11	(0,48)	1,09	(0,66)	1,66
BRA	7,90	(0,62)	8,75	(1,03)	0,86	(1,20)	0,71
CAN	3,00	(0,18)	12,34	(0,76)	9,34	(0,78)	11,93
CHE	3,36	(0,32)	3,06	(0,26)	−0,30	(0,41)	−0,72
CZE	1,90	(0,42)	5,65	(1,19)	3,75	(1,26)	2,97
DEU	3,05	(0,34)	9,92	(0,63)	6,87	(0,72)	9,55
DNK	7,12	(0,85)	2,73	(0,37)	−4,40	(0,92)	−4,76
ESP	4,48	(0,49)	3,70	(0,37)	−0,78	(0,62)	−1,27
FIN	1,96	(0,22)	1,44	(0,16)	−0,52	(0,27)	−1,92
FRA	6,23	(0,51)	4,61	(0,45)	−1,61	(0,68)	−2,37
GBR	5,15	(0,44)	7,23	(1,17)	2,07	(1,25)	1,66
GRC	4,04	(0,57)	5,81	(0,41)	1,78	(0,70)	2,53
HUN	3,02	(0,36)	5,39	(0,42)	2,37	(0,55)	4,31
IDN	6,99	(0,64)	8,67	(0,53)	1,67	(0,83)	2,03
IRL	3,23	(0,34)	4,32	(0,57)	1,09	(0,66)	1,65
ISL	2,19	(0,24)	2,30	(0,25)	0,11	(0,35)	0,31
ITA	2,73	(0,46)	2,47	(0,28)	−0,26	(0,54)	−0,48
JPN	62,52	(3,47)	11,25	(0,81)	−51,27	(3,56)	−14,41
KOR	7,34	(0,49)	2,36	(0,21)	−4,97	(0,54)	−9,29
LIE	5,49	(1,41)	3,02	(0,85)	−2,47	(1,64)	−1,50
LUX	9,55	(0,50)	3,62	(0,29)	−5,92	(0,58)	−10,27
LVA	5,02	(0,52)	3,34	(0,39)	−1,68	(0,66)	−2,56
MEX	8,51	(0,59)	5,07	(0,44)	−3,43	(0,74)	−4,65
NLD	3,07	(0,65)	7,64	(1,34)	4,57	(1,49)	3,07
NOR	2,44	(0,31)	3,18	(0,39)	0,74	(0,50)	1,49
NZL	3,92	(0,39)	14,13	(0,43)	10,22	(0,58)	17,60
POL	6,90	(0,79)	2,33	(0,30)	−4,57	(0,85)	−5,39
PRT	3,72	(0,42)	2,76	(0,28)	−0,96	(0,50)	−1,90
RUS	3,16	(0,33)	2,14	(0,30)	−1,02	(0,45)	−2,27
SWE	2,48	(0,30)	2,63	(0,31)	0,15	(0,43)	0,35
THA	10,95	(1,38)	5,85	(0,64)	−5,09	(1,52)	−3,35
USA	14,58	(1,95)	5,88	(0,38)	−8,70	(1,99)	−4,38

Este sencillo ejemplo muestra que la interpretación de los indicadores de tendencia es bastante complicada. La estructura social y económica de un país debería permanecer inalterada durante un período de tres años, de modo que no ocurran diferencias entre dos ciclos. Sin embargo,

como se ha mostrado, esta diferencia aparece como significativa en los 13 países.

A veces, estas diferencias significativas pueden explicarse por cambios en las tasas de participación de centros o de alumnos y en la distribución de los datos perdidos. Por lo tanto, se recomienda llevar a cabo alguna verificación antes de intentar interpretar las diferencias calculadas como un verdadero cambio en las características de la población.

Cálculo del error típico de los indicadores de tendencias en las variables del rendimiento

Anclaje de las escalas de rendimiento de PISA 2000 y PISA 2003

La base de datos de PISA 2000 contiene cinco valores plausibles para cada una de las siguientes áreas o subáreas:

- matemáticas;
- lectura;
  - recuperación de la información;
  - interpretación;
  - reflexión;
- ciencias.

La base de datos de PISA 2003 también contiene cinco valores plausibles para cada una de las siguientes áreas o subáreas:

- matemáticas,
  - espacio y forma,
  - cambio y relaciones,
  - incertidumbre,
  - cantidad,
- solución de problemas,
- lectura,
- ciencias.

Los procedimientos psicométricos utilizados para equiparar las escalas de rendimiento de PISA 2000 y PISA 2003 son distintos para las matemáticas que para la lectura y las ciencias.

La lectura fue el área principal en el 2000 y 28 de los 140 ítems desarrollados para la evaluación del 2000 se usaron también en la del 2003. Los datos del 2003, por tanto, se expresaron en la escala de lectura del 2000. Los datos de evaluación de ciencias de 2003 también se expresan en la escala de ciencias del 2000, ya que 25 de los 30 ítems desarrollados para la evaluación del 2000 se usaron en la del 2003.

A las matemáticas, como área principal, se les dedicó un gran trabajo de desarrollo en PISA 2003. Además, la evaluación de las matemáticas en el 2000 sólo cubrió dos de las cuatro áreas de contenido (espacio y forma y cambio y relaciones). 20 ítems de los 85 utilizados en la evaluación de 2003 provienen de la de 2000. Debido a esta ampliación de la evaluación, se creyó que era inapropiado expresar las puntuaciones de matemáticas de PISA 2003 en la escala de PISA 2000.

Sin embargo, para proporcionar a los países algunos indicadores de tendencias, las subescalas espacio y forma y cambio y relaciones de PISA 2000 se expresaron en las escalas de PISA 2003[1].

Los pasos para el anclaje de los datos de lectura y ciencias de PISA 2003 en las escalas de PISA 2000 son:

1. Calibración de los datos de lectura y ciencias del 2003 para obtener los parámetros de los ítems, es decir, la dificultad relativa de los ítems en la escala de Rasch.

2. Basándose en estos parámetros de los ítems, generación de valores plausibles para lectura y ciencias a partir de los datos de PISA 2003.

3. Basándose en los parámetros de los ítems del paso 1, pero sólo en los ítems de anclaje, generación de valores plausibles para lectura y ciencias a partir de los datos de PISA 2000. En este momento, hay dos conjuntos de valores plausibles para PISA 2000: el conjunto original de valores plausibles incluido en la base de datos de PISA 2000 y el conjunto de valores plausibles basados en los parámetros de los ítems de PISA 2003. Por desgracia, la media y la desviación típica del nuevo conjunto de valores plausibles serán ligeramente distintas de los valores plausibles originales de PISA 2000. Estas diferencias reflejan los cambios en la dificultad de los ítems de anclaje entre 2000 y 2003. Recordemos que la media y la desviación típica para la media de la OCDE se establecieron respectivamente en 500 y 100 en el 2000. Supongamos que el nuevo conjunto de valores plausibles proporciona una media de 505 y una desviación típica de 110. El nuevo conjunto de valores plausibles para los datos de PISA 2000 debe transformarse, de modo que su media y su desviación típica sean respectivamente iguales a 500 y 100.

4. Este paso consiste en calcular la transformación lineal que garantizará que la media y la desviación típica del nuevo conjunto de valores plausibles a partir de los datos de PISA 2000 tenga una media de 500 y una desviación típica de 100. Esta transformación lineal puede escribirse como:

$$PV_{cal\_2000} = \alpha + \beta \cdot PV_{cal\_2003}, \text{ con } \beta = \frac{\sigma_{cal\_2000}}{\sigma_{cal\_2003}} \text{ y } \alpha = \left(\mu_{cal\_2000} - \beta \cdot \left(\mu_{cal\_2003}\right)\right).$$

   En el ejemplo, $\beta = \frac{100}{110} = 0{,}909$ y $\alpha = \left(500 - \left(0{,}909 \cdot 505\right)\right) = 40{,}955$ [2].

5. Esta transformación lineal se aplica a los valores plausibles de PISA 2003. Aplicada a los valores plausibles de lectura o ciencias de PISA 2003, garantiza que el rendimiento de los alumnos en el 2003 sea comparable al del 2000.

Como ya se indicó más atrás, con otro conjunto de ítems de anclaje la transformación lineal habría sido distinta. Por consiguiente, existe una incertidumbre en la transformación debida al muestreo de los ítemes de anclaje, denominada *error de equiparación*.

Los pasos para anclar las dos subescalas de matemáticas de PISA 2000 en las subescalas de PISA 2003 son:

- Calibración de los datos de matemáticas del 2003 para obtener los correspondientes parámetros de los ítems de PISA 2003.

- Basándose en estos parámetros de los ítems, generación de los valores plausibles para PISA 2003.
- Basándose en los parámetros de los ítems del 2003, generación de valores plausibles para los datos de matemáticas de PISA 2000.

Del mismo modo, la estimación de la tendencia habría sido levemente distinta con otro conjunto de ítems de anclaje en lectura y ciencias. Por lo tanto, es importante integrar este componente de error en el error típico del indicador de tendencia.

Inclusión del error de equiparación en el cálculo del error típico

Para cada ítem de anclaje, tenemos dos estimaciones de parámetros del ítem que se encuentran ahora en la misma métrica: el parámetro del ítem del 2000 y el parámetro del ítem del 2003. Algunos de estos ítems de anclaje muestran un aumento de la dificultad relativa, otros una disminución, pero por término medio la diferencia es igual a 0. Esto significa que algunos ítems parecen más difíciles en el 2003 que en el 2000 o viceversa.

Ya que el subconjunto de ítems de anclaje puede considerarse una muestra aleatoria simple de una población infinita de ítems de anclaje, el error de equiparación puede calcularse así:

$$\sigma_{(error\_de\_equiparación)} = \sqrt{\frac{\sigma^2}{n}},$$

donde σ^2 representa la varianza de las diferencias de parámetros de los ítems y n, el número de ítems de anclaje utilizados.

Si los parámetros de los ítems procedentes de la calibración del 2003 casaran perfectamente con los parámetros de los ítems del 2000, la dificultad relativa de los ítems de anclaje no habría cambiado. Todas las diferencias entre la dificultad relativa en el 2000 y en el 2003 serían iguales a 0 y, por tanto, el error de equiparación sería igual a 0.

A medida que las diferencias de los parámetros de los ítems aumentan, la varianza de estas diferencias también lo hará, así como el error de equiparación. Parece lógico que la incertidumbre alrededor de la tendencia sea proporcional a los cambios en los parámetros de los ítems.

Así mismo, la incertidumbre alrededor de los indicadores de tendencia es inversamente proporcional al número de ítems de anclaje. Desde un punto de vista teórico, sólo hace falta un ítem para medir una tendencia, pero con sólo un ítem, la incertidumbre será muy grande. Si el número de ítems de anclaje aumenta, la incertidumbre disminuirá.

La tabla 12.3 proporciona los parámetros de los ítems centrados (es decir, las diferencias de dificultad entre los ítems) para los ítems de anclaje en lectura de PISA 2000 y de PISA 2003, así como las diferencias entre los dos conjuntos de estimaciones.

Tabla 12.3. Estimaciones de los parámetros de los ítems en el 2000 y el 2003 para los ítems de anclaje en lectura

Nombre del ítem	*Delta* centrado en 2003	*Delta* centrado en 2003	Diferencia
R055Q01	−1,28	−1,347	−0,072
R055Q02	0,63	0,526	−0,101
R055Q03	0,27	0,097	−0,175
R055Q05	−0,69	−0,847	−0,154
R067Q01	−2,08	−1,696	0,388
R067Q04	0,25	0,546	0,292
R067Q05	−0,18	0,212	0,394
R102Q04A	1,53	1,236	−0,290
R102Q05	0,87	0,935	0,067
R102Q07	−1,42	−1,536	−0,116
R104Q01	−1,47	−1,205	0,268
R104Q02	1,44	1,135	−0,306
R104Q05	2,17	1,905	−0,267
R111Q01	−0,19	−0,023	0,164
R111Q02B	1,54	1,395	−0,147
R111Q06B	0,89	0,838	−0,051
R219Q01T	−0,59	−0,520	0,069
R219Q01E	0,10	0,308	0,210
R219Q02	−1,13	−0,887	0,243
R220Q01	0,86	0,815	−0,041
R220Q02B	−0,14	−0,114	0,027
R220Q04	−0,10	0,193	0,297
R220Q05	−1,39	−1,569	−0,184
R220Q06	−0,34	−0,142	0,196
R227Q01	0,40	0,226	−0,170
R227Q02T	0,16	0,075	−0,086
R227Q03	0,46	0,325	−0,132
R227Q06	−0,56	−0,886	−0,327

La varianza de la diferencia es igual a 0,047486. El error de equiparación, por tanto, es igual a:

$$\sigma_{(error\_de\_equiparación)} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{0{,}047486}{28}} = 0{,}041\,.$$

En la escala de lectura de PISA, con una media de 500 y una desviación típica de 100, corresponde a 3,75.

Los errores de equiparación entre PISA 2000 y PISA 2003 son:

- lectura: 3,75;
- ciencias: 3,02;
- forma y espacio (matemáticas): 6,01;
- cambio y relaciones (matemáticas): 4,84.

A partir de los ítems de anclaje, se ha estimado una transformación común, que se aplica a todos los países participantes. Se deduce que cualquier incertidumbre que se introduzca a través de la equiparación es común a todos los alumnos y todos los países. Así, por ejemplo, supongamos que el error de equiparación desconocido entre PISA 2000 y PISA 2003 en lectura produ-

jo una sobreestimación de las puntuaciones de los alumnos de dos puntos en la escala de PISA 2000. Por tanto, todas las puntuaciones de los alumnos se sobreestimarían en dos puntos. Esta sobreestimación influiría en ciertos estadísticos, no en todos, calculados a partir de los datos de PISA 2003. Por ejemplo, consideremos:

- La media de cada país se sobreestimaría en una cantidad igual al error de equiparación; en nuestro ejemplo, esto son dos puntos.
- El rendimiento medio de cualquier subgrupo se sobreestimaría en una cantidad igual al error de equiparación; en nuestro ejemplo, esto son dos puntos.
- La desviación típica de las puntuaciones de los alumnos no se vería afectada, porque la sobreestimación de cada alumno por un error común no cambia la desviación típica.
- La diferencia entre las puntuaciones medias de dos países en PISA 2003 no resultaría influida, porque la sobreestimación de cada alumno por un error común habría distorsionado la media de cada país en la misma cantidad.
- La diferencia entre las puntuaciones medias de dos grupos (por ejemplo, varones y mujeres) en PISA 2003 no resultaría afectada, porque la sobreestimación de cada alumno por un error común habría distorsionado la media de cada grupo en la misma cantidad.
- La diferencia entre el rendimiento de un grupo de estudiantes (por ejemplo, un país) entre PISA 2000 y PISA 2003 cambiaría, porque el error influiría en la puntuación de cada alumno en PISA 2003.
- Un cambio en la diferencia entre dos grupos de PISA 2000 a PISA 2003 no resultaría afectado, porque ninguno de los componentes de esta comparación, que son diferencias de puntuación en el 2000 y el 2003 respectivamente, queda influido por un error común que se añade a todas las puntuaciones de los alumnos en PISA 2003.

En términos generales, el error de equiparación sólo tendría que considerarse cuando se establezcan comparaciones entre los resultados de PISA 2000 y PISA 2003, y entonces, sólo cuando se comparen las medias de grupos.

El ejemplo más obvio de una situación en la que es necesario usar el error de equiparación es en la comparación del rendimiento medio de un país entre PISA 2000 y PISA 2003.

En PISA 2000, la media en lectura para Alemania es igual a 483,99, con un error típico de 2,47. En PISA 2003, la media de Alemania es igual a 491,36, y el error típico es igual a 3,39. La diferencia entre el 2000 y el 2003 es, por tanto, igual a 491,36 − 483,99 = 7,37. El rendimiento promedio de los estudiantes alemanes, por tanto, ha aumentado en 7,37 puntos en la escala de lectura de PISA 2000.

El error típico de esta diferencia, como se ha mencionado más arriba, está influido por el error de equiparación. Por consiguiente, el error típico es igual a:

$$ET = \sqrt{\sigma^2_{(\hat{\mu}_{2000})} + \sigma^2_{(\hat{\mu}_{2000})} + \sigma^2_{(error\_de\_equiparación)}}$$

$$ET = \sqrt{(2{,}47)^2 + (3{,}39)^2 + (3{,}75)^2} = 5{,}63$$

Puesto que la diferencia tipificada entre PISA 2000 y PISA 2003 (7,37/5,63) está incluida en el intervalo [−1,96; 1,96], la hipótesis nula (es decir, no hay ninguna diferencia) no se rechaza. Dicho de otro modo, el rendimiento de Alemania en lectura no ha cambiado entre 2000 y 2003.

La tabla 12.4 proporciona las estimaciones de la competencia en lectura en Alemania según el género en 2000 y 2003, con sus respectivos errores típicos, así como las estimaciones de la diferencia y sus respectivos errores típicos.

Tabla 12.4. Rendimiento medio en la lectura según el género en Alemania

		Rendimiento en lectura	Error típico
2003	Chicas	512,93	3,91
	Chicos	470,80	4,23
	Diferencia	42,13	4,62
2000	Chicas	502,20	3,87
	Chicos	467,55	3,17
	Diferencia	34,65	5,21

Puesto que la comparación para un país determinado entre el 2000 y el 2003 resulta afectada por el error de equiparación, ocurre lo mismo en la comparación para un subgrupo determinado entre el 2000 y el 2003. Por lo tanto, el error típico debe incluir el error de equiparación.

Los indicadores de tendencia para los chicos alemanes y las chicas alemanas son, respectivamente, iguales a:

$$Tendencia_{chicas} = 512,93 - 502,20 = 10,73$$

$$ET_{chicas} = \sqrt{(3,91)^2 + (3,87)^2 + (3,75)^2} = 6,66$$

$$Tendencia_{(chicos)} = 470,80 - 467,55 = 3,25$$

$$ET_{(chicos)} = \sqrt{(4,23)^2 + (3,17)^2 + (3,75)^2} = 6,48$$

Ambas diferencias no son estadísticamente distintas de 0.

Por otra parte, la diferencia según el sexo en 2003 no se ve afectada por el error de equiparación. Es más, las estimaciones de ambos subgrupos serán sobreestimadas o infraestimadas en la misma cantidad, por lo que el cálculo de la diferencia neutralizará esta diferencia. Por consiguiente, el indicador de tendencia de la diferencia según el sexo y su error típico será igual a:

$$Tendencia_{(dif.\_género)} = 42,813 - 34,65 = 7,43$$

$$ET_{(dif.\_género)} = \sqrt{(4,62)^2 + (5,21)^2} = 6,96$$

Esto significa que el cambio de la diferencia según el sexo en Alemania, para la lectura, entre el 2000 y el 2003 no fue estadísticamente significativo, incluso aunque de la tabla 12.4 se deduce que aumentó considerablemente.

En los informes iniciales de PISA 2000 y PISA 2003, el rendimiento de los alumnos también se

describe mediante niveles de aptitud (véase el capítulo 8). Puesto que el error de equiparación afecta a la estimación de la media del país, los porcentajes de alumnos en cada nivel también resultarán afectados. Sin embargo, una sobreestimación o una infraestimación de los resultados de PISA 2003 de *X* puntos en la escala de PISA tendrá un impacto distinto en los porcentajes de alumnos en cada nivel de aptitud por cada país. Si el porcentaje es pequeño, el impacto será pequeño. Si el porcentaje es grande, el impacto será mayor. Habría sido demasiado complicado proporcionar por cada país y por cada nivel de aptitud un error de equiparación. Así pues, se decidió no tener en cuenta el error de equiparación para la comparación de porcentajes de alumnos en cada nivel de aptitud entre PISA 2000 y PISA 2003. Esto significa que los errores típicos de la diferencia entre el 2000 y el 2003 están infraestimados.

Conclusiones

Este capítulo ha estado dedicado al cálculo del error típico de los indicadores de tendencia. La comparación de cualquier variable distinta de las variables del rendimiento es bastante directa, ya que las muestras de PISA 2000 y PISA 2003 son independientes. Sin embargo, como se ha indicado con anterioridad, tales comparaciones sólo son relevantes si las medidas de 2000 y 2003 son comparables.

La comparación de las estimaciones de la media del rendimiento es más compleja, puesto que podría exigir la inclusión del error de equiparación en el error típico, según el estadístico de que se trate. Por ejemplo, la figura 2.6d en *Learning for Tomorrow's World – First Results from PISA 2003* (OCDE, 2004*a*) presenta las tendencias del rendimiento promedio en espacio y forma entre el 2000 y el 2003. El indicador de tendencia ha integrado el error de equiparación en su error típico. En el mismo informe, la figura 2.6c presenta las tendencias entre 2000 y 2003 en los percentiles 5.º, 10.º, 25.º, 75.º, 90.º y 95.º, y el error de equiparación no estaba integrado en el error típico de las tendencias. En términos generales, el informe inicial de PISA 2003 ha integrado el error de equiparación sólo en las tablas donde se compara el rendimiento medio de un país entre el 2000 y el 2003.

Debido al creciente interés acerca de los indicadores de tendencia y su impacto político, es fundamental interpretar con precaución los cambios significativos. Éstos podrían deberse simplemente a una diferencia en la tasa de participación de los centros o los alumnos o en el patrón de los datos perdidos.

---

[1] La base de datos de PISA 2000 se ha actualizado para integrar este nuevo conjunto de valores plausibles.

[2] En realidad, la transformación lineal se aplicó a los valores plausibles antes de su transformación en la escala de PISA, con una media de 500 y una desviación típica de 100. Además, se aplicaron distintas transformaciones según el género (por ejemplo, chicas, chicos y género ausente). Las transformaciones lineales según el género son:

(1) chicas: 2000_PV = 0,0970 + (0,8739 ·2003_PV);

(2) chicos: 2000_PV = 0,0204 + (0,8823 ·2003_PV);

(3) género ausente: 2000_PV = 0,0552 + (0,8830 ·2003_PV).

En ciencias, la transformación lineal es: 2000_PV = –0,01552 + (1,0063 ·2003_PV).

Capítulo 13

El análisis multinivel

Introducción

Durante las últimas dos décadas, los datos de las encuestas sobre educación se han analizado cada vez más con modelos multinivel. De hecho, puesto que los modelos de regresión lineal no permiten tener en cuenta los efectos que puede provocar el modo en que los alumnos se asignan a centros o a clases dentro de los centros, pueden proporcionar una representación incompleta o engañosa de la eficiencia de los sistemas educativos. Por ejemplo, en algunos países, el contexto socioeconómico de un alumno puede determinar en parte el tipo de centro al que asiste y, por tanto, habrá poca variación en el contexto socioeconómico de los alumnos dentro de cada centro. En otros países o sistemas, los centros pueden seleccionar alumnos procedentes de un amplio espectro de contextos socioeconómicos, pero dentro del centro el contexto socioeconómico del alumno influye en el tipo de clase al que se lo asigna y, como resultado, en la varianza dentro del centro. Un modelo de regresión lineal que no tenga en cuenta la estructura jerárquica de los datos no diferenciará entre estos dos sistemas.

El uso de modelos multinivel (Goldstein, 1995), también llamados *modelos lineales jerárquicos* (Bryk y Raudenbush, 1992), reconoce el hecho de que los estudiantes están anidados dentro de clases y centros. La variación relativa de la medida del resultado —entre alumnos, dentro del mismo centro y entre centros— puede por tanto evaluarse.

La regresión lineal simple

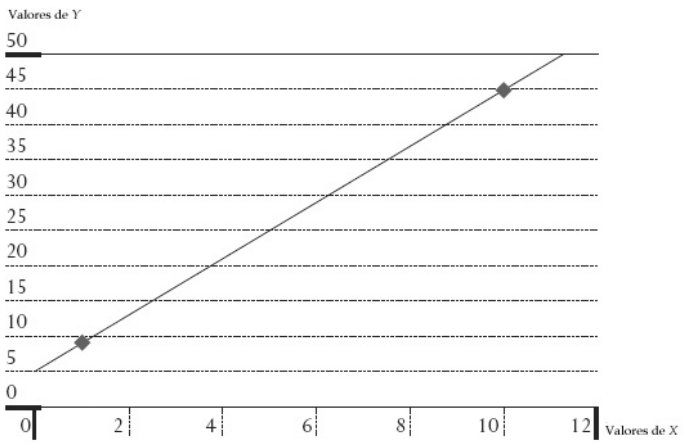
Una ecuación lineal puede representarse siempre mediante una línea recta. Una ecuación con dos variables se representará en un espacio bidimensional; una ecuación con tres variables, en un espacio tridimensional, y así sucesivamente.

La siguiente ecuación se representa gráficamente en la figura 13.1.

$$Y = 5 + 4X$$

Puesto que todas las ecuaciones lineales se representan mediante una línea recta, sólo es necesario identificar dos puntos que pertenezcan a la línea para poder dibujarla. Si X es igual a 1, entonces Y será igual a 9. Si X es igual a 10, entonces Y será igual a 45. La línea recta con los puntos (1, 9) y (10, 45) corresponde a la ecuación.

Figura 13.1. Representación gráfica de una ecuación lineal



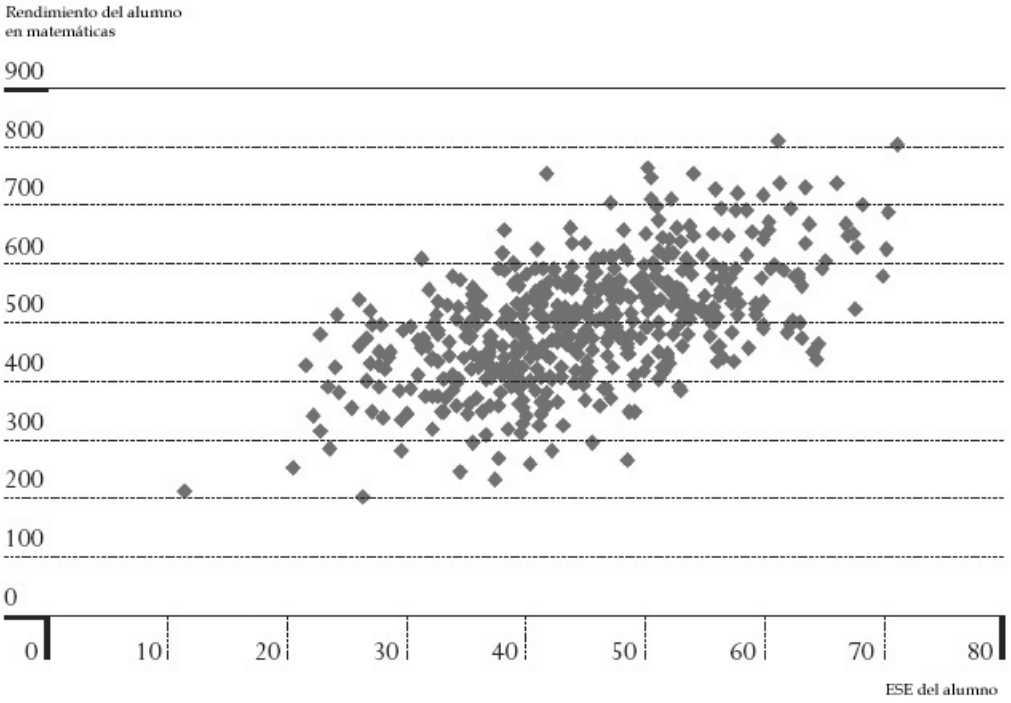
La figura 13.1 muestra la representación gráfica de la ecuación $Y = 5 + 4X$. Como muestra la figura, la línea cruza el eje Y en 5. El punto (0, 5) se llama *intercepto* y da el valor de Y cuando X es igual a 0. El factor de X o, en términos estadísticos, el *coeficiente de regresión*, da la pendiente de la línea recta. Nos informa sobre el aumento de Y para una unidad adicional en el eje X. En el ejemplo considerado, si X aumenta en una unidad, Y lo hará en cuatro unidades.

La expresión general de una ecuación lineal con dos variables es:

$Y = a + bX$, donde a es el intercepto y b, el coeficiente de regresión.

Aunque los procesos humanos pueden describirse también con un enfoque similar, son menos deterministas. Representaremos gráficamente la relación que podría existir entre el entorno socioeconómico de los alumnos (ESE) y su rendimiento académico.

Figura 13.2. Relación entre el entorno socioeconómico de los alumnos y su rendimiento académico

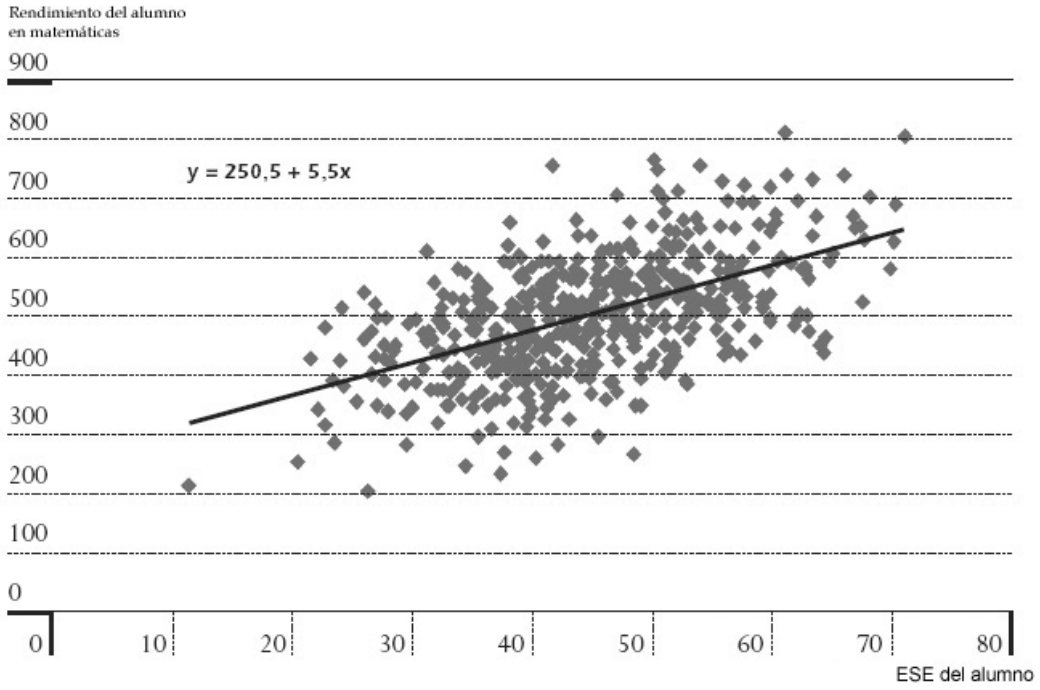


Como muestra la figura 13.2, existe una relación positiva entre el entorno socioeconómico del alumno y el rendimiento académico. El rendimiento de los alumnos de entornos socioeconómicos más altos tiende a ser mayor. Sin embargo, a diferencia de una ecuación lineal, no todos los puntos están en una línea recta, lo que significa que los alumnos de un nivel socioeconómico bajo podrían tener un buen rendimiento académico, y viceversa.

Los estadísticos usan un análisis de regresión lineal para cuantificar tales relaciones. El proceso en este ejemplo concreto es similar a una ecuación lineal con dos variables. Consiste en calcular una ecuación $Y_i = \alpha + \beta X_i$, donde Y_i es el rendimiento académico del alumno i, y X_i es el entorno socioeconómico familiar. Esta ecuación también puede representarse mediante una línea, recta en este caso, llamada *línea de regresión.*

La línea de regresión de la figura 13.3 corresponde a la ecuación de regresión $Y_i = 250{,}5 + 5{,}5X_i$. Una medida del estatus socioeconómico utilizada en PISA 2000 y PISA 2003 (Ganzeboom y otros, 1992) es el índice HISEI o estatus profesional más alto de ambos padres. Este índice va desde 16 a 90, con una media de aproximadamente 50 y una desviación típica de alrededor de 15. El rendimiento en matemáticas tiene una media internacional de 500 y una desviación típica de 100. Esta ecuación muestra que el aumento de una unidad en la escala de HISEI se asocia, por término medio, con un aumento de 5,5 puntos en la escala de matemáticas de PISA.

Figura 13.3. Línea de regresión del entorno socioeconómico en el rendimiento de los alumnos en matemáticas



Esta ecuación de regresión también puede utilizarse para predecir el rendimiento en matemáticas de un alumno si se conoce su entorno socioeconómico. Por ejemplo, esta ecuación de regresión predecirá para cada alumno con un valor de HISEI de 60 una puntuación de 250,5 + (5,5 · 60) = 580,5. Dicho de otro modo, cualquier alumno con un HISEI de 60 tendrá una puntuación predicha de 580,5. Sin embargo, como muestra la figura 13.3, algunos de estos alumnos tienen un rendimiento muy cercano a esta puntuación predicha, normalmente llamada $\hat{Y}_i$, pero los demás tienen un rendimiento mejor o peor.

Antes de calcular la ecuación de regresión, cada alumno de la muestra podía caracterizarse mediante su HISEI *(X_i)* y su rendimiento en matemáticas *(Y_i)*. Ahora, también puede caracterizarse a los alumnos mediante su puntuación predicha, $\hat{Y}_i$, y mediante la diferencia entre la puntuación observada y la puntuación predicha, $(Y_i - \hat{Y}_i)$, normalmente llamada *residuo* (o ε_i).

Tabla 13.1. HISEI, rendimiento en matemáticas, puntuación predicha y residuo

Alumnos	HISEI	Puntuación observada	Puntuación predicha	Residuo
1	49	463	520	−57
2	53	384	542	−158
3	51	579	531	+48
4	42	404	481,5	−77,5
5	42	282	481,5	−199,5

El primer alumno tiene un valor HISEI de 49 y un rendimiento en matemáticas de 463. Basándose en su entorno socioeconómico, habríamos predicho una puntuación de 520. Este alumno tiene, por tanto, un rendimiento inferior al que se esperaba. El residuo es igual a −57. Por otra parte, el tercer alumno tiene un rendimiento de 579 y una puntuación esperada de 531. Este alumno tiene un rendimiento mejor de lo esperado.

La tabla 13.1 muestra que las puntuaciones observadas, las puntuaciones predichas y las puntuaciones residuales presentan cierta variabilidad, a partir de la que pueden calcularse coeficientes de varianza. La ecuación de regresión y la línea de regresión están elaboradas de tal modo que minimizan la varianza del residuo, llamada *varianza residual*. Esto significa que:

- la ecuación de regresión debe incluir el punto (μ_x, μ_y);
- la media de la puntuación predicha es igual a la media de la puntuación observada ($\mu_y = \mu_{\hat{y}}$);
- la media del residuo debe ser igual a 0.

Por último, un análisis de regresión puede extenderse a diversas variables explicativas. Si se incorporan *k* variables predictoras en la regresión, la ecuación se escribirá así:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Por ejemplo, el rendimiento de matemáticas en la prueba de PISA puede explicarse según el entorno familiar del estudiante, el género, el tiempo que estudie en casa durante la semana, el interés por las matemáticas y así sucesivamente.

Cuadro 13.1. Interpretación de un coeficiente de regresión y de un intercepto

Un coeficiente de regresión refleja el cambio de las unidades en el eje *Y* (la variable dependiente; en este caso particular, el aumento en la escala de matemáticas) por cada cambio de una unidad del eje *X*. La interpretación de un coeficiente de regresión depende de la unidad de medida de la variable independiente. Por tanto, el efecto estadístico de distintas variables independientes no puede compararse, a menos que estas variables independientes estén expresadas en las mismas unidades de medida.

Para conseguir esto, las variables independientes pueden tipificarse, de modo que las unidades de medida se conviertan en la desviación típica. Si todas las variables tienen una desviación típica de 1, los coeficientes de regresión de las distintas variables pueden compararse directamente. Los coeficientes de regresión reflejarán el aumento en la escala matemática de las variables independientes por cada desviación típica.

Supongamos que dos variables independientes, llamadas X_1 y X_2, se utilizan para explicar el rendimiento en matemáticas de los alumnos en dos países. Las tablas siguientes proporcionan los coeficientes de regresión y la desviación típica de X_1 y X_2 antes y después de tipificar las variables independientes.

Antes de la tipificación				
	X_1		X_2	
	β_1	$\sigma_{(x_1)}$	β_1	$\sigma_{(x_2)}$
País A	10	2	15	3
País B	5	1	7,5	1,5

Después de la tipificación				
	X_1		X_2	
	β_1	$\sigma_{(x_1)}$	β_1	$\sigma_{(x_2)}$
País A	5	1	5	1
País B	5	1	5	1

Los resultados son bastante distintos. Basándose en los coeficientes de regresión después de la tipificación, parece que las dos variables independientes tienen el mismo efecto estadístico sobre el rendimiento matemático en ambos países. Supongamos que X_1 representa el tiempo dedicado al estudio en casa. En el país A, el aumento de una hora dedicada al estudio se asocia con un aumento de 10 puntos en la escala de matemáticas, mientras que en el país B, una hora adicional se asocia con un aumento de 5 puntos. Si bien la tipificación de las variables permite realizar comparaciones, la interpretación de determinado un coeficiente de regresión se hace más compleja, puesto que ya no se refiere a la escala de medida original.

Por tanto, no hay un único algoritmo para solucionar este problema. Depende de la naturaleza de la variable independiente y del propósito de los análisis.

La interpretación del intercepto es incluso más compleja, ya que depende de la desviación típica y de la media de las variables independientes. Supongamos que HISEI se tipifica a una media de 0 y una desviación típica de 1. El coeficiente de regresión reflejaría el aumento en matemáticas por desviación típica en la escala de estatus socioeconómico. Por tanto, el intercepto representaría el rendimiento de un alumno con una puntuación HISEI transformada de 0. En un modelo con sólo variables tipificadas, reflejaría el rendimiento de un alumno hipotético que obtiene puntuaciones iguales a la media en todas las variables independientes.

El análisis de regresión lineal simple frente al de regresión multinivel

La regresión lineal simple precedente ha demostrado la relación entre el entorno socioeconómico y el rendimiento en matemáticas al nivel de la población, es decir, de los alumnos de 15 años escolarizados en una institución educativa.

Una relación entre el contexto socioeconómico del alumno y el rendimiento en matemáticas no implica necesariamente que los países más ricos tengan una media de rendimiento superior a la de los países en desarrollo. Además, la relación observada a nivel de los alumnos en los centros no implica necesariamente que se produzca el mismo fenómeno dentro de cada centro.

Los análisis de regresión multinivel tienen en cuenta que las unidades muestrales están anidadas dentro de unidades más amplias. En lugar de calcular una ecuación de regresión sobre el conjunto de datos entero, el análisis de regresión multinivel calculará una ecuación de regresión por cada unidad más amplia. Por tanto, un análisis de regresión multinivel calculará una ecuación de regresión por cada centro.

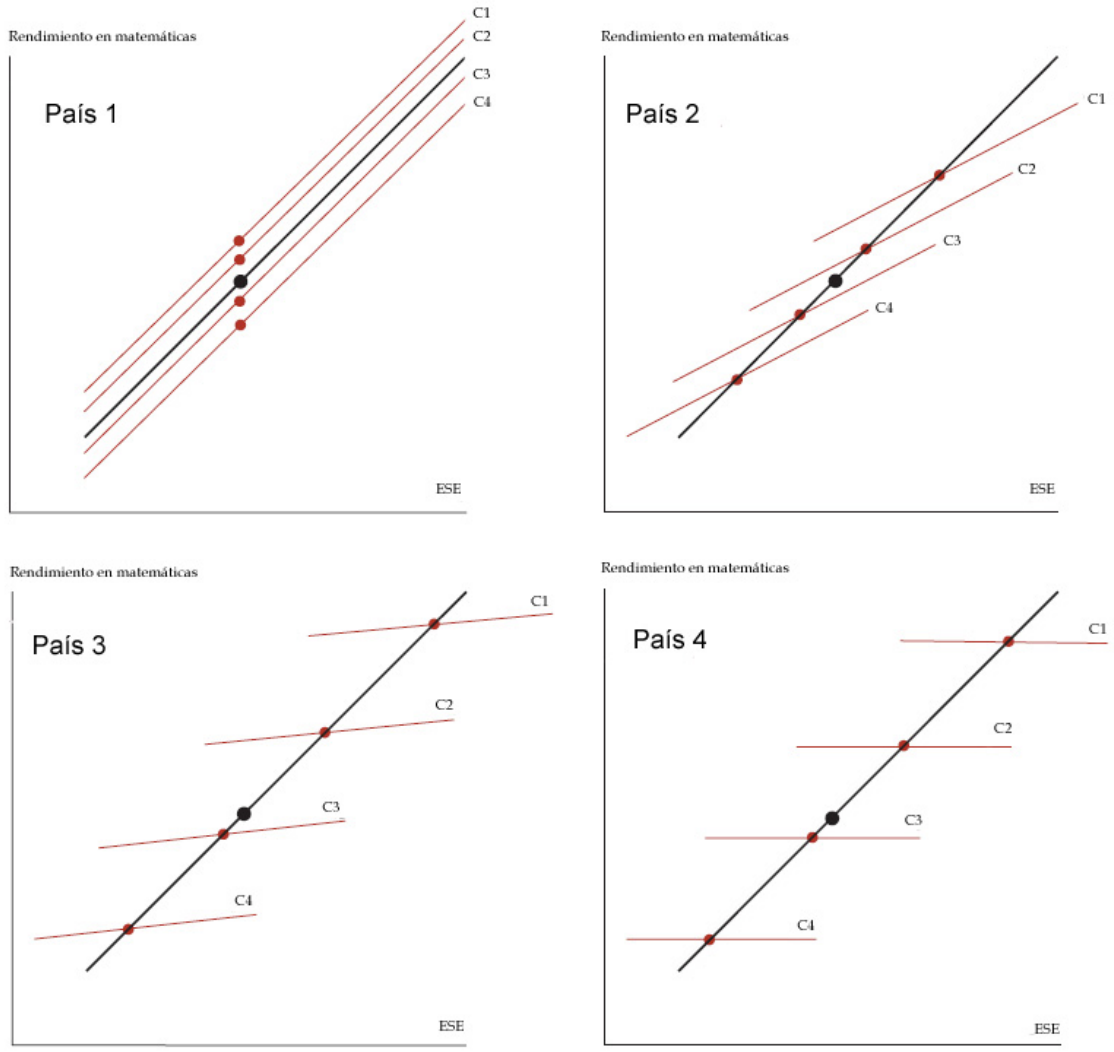
La figura 13.4 muestra cuatro gráficos que destacan la distinción entre una regresión lineal y un modelo de regresión lineal multinivel. Estos cuatro gráficos representan la relación entre los entornos socioeconómicos de los alumnos (ESE) y sus estimaciones de rendimiento en matemáticas, en distintos países.

La línea gruesa representa la línea de regresión, cuando no se tiene en cuenta la estructura jerárquica de los datos. Las líneas finas representan la relación entre estas dos varibles dentro de determinados centros. Para cada centro, existe una línea de regresión (una línea fina en este ejemplo concreto). El punto más grande en las líneas de regresión gruesas representa el punto con la media de *X* e *Y* como coordenadas, (μ_x, μ_y), y el punto más pequeño en las líneas de regresión multinivel representa el punto con la media de los centros de *X* e *Y* como coordenadas, (μ_{xi}, μ_{yi}).

El análisis de regresión lineal simple, gráficamente representado por las líneas gruesas, demuestra que la puntuación esperada de un alumno procedente de un entorno socioeconómico más alto es considerablemente más elevada que la puntuación esperada de un alumno procedente de un entorno socioeconómico más bajo. La comparación entre los cuatro gráficos muestra la similitud de la relación entre el entorno socioeconómico y el rendimiento del alumno en cada país. Basándonos en análisis de regresión lineal simples, concluiríamos que la relación entre el entorno socioeconómico y el rendimiento del alumno es idéntica en los distintos países.

Sin embargo, los análisis de regresión multinivel distinguen claramente la relación entre las dos variables en los cuatro países.

Figura 13.4. Análisis de regresión lineal frente a análisis de regresión multinivel



En el país 1, las líneas de regresión multinivel son similares y cercanas a la línea de regresión lineal simple. Esto significa que:

- En cuanto al entorno socioeconómico del alumno (eje *X*):
  - A los diferentes centros asisten alumnos procedentes de una amplia variedad de entornos socioeconómicos. Todas las líneas de regresión intra-centro cubren la gama completa de valores en el eje *X*.
  - En los centros se matriculan alumnos del mismo entorno socioeconómico. De hecho, las proyecciones de los puntos pequeños en el eje *X* están muy cercanas entre sí.
- En cuanto al rendimiento del alumno en matemáticas (eje *Y*):
  - En cada centro, existen alumnos con bajo, medio y alto rendimiento. Todas las líneas de regresión intra-centro cubren el eje Y.

- Por término medio, los centros tienen un nivel similar de rendimiento. De hecho, las proyecciones de los puntos pequeños en el eje Y están muy cercanas entre sí. También significa que la varianza inter-centro es bastante pequeña.

- En cuanto a la relación entre el contexto socioeconómico y el rendimiento en matemáticas:
  - En cada centro existe una fuerte relación entre el entorno socioeconómico y el rendimiento. Dentro de todos los centros, los alumnos con un nivel socioeconómico bajo tienen un rendimiento bastante inferior al de los alumnos con mayor nivel socioeconómico. La pendiente de la línea de la regresión intra-centro indica la intensidad de la relación.

Cada centro del país 1 puede considerarse, por tanto, una muestra aleatoria simple de la población, y cada centro refleja las relaciones que existen a nivel poblacional.

La situación opuesta al país 1 se representa gráficamente en el país 4. Las líneas de regresión multinivel difieren considerablemente de la línea de regresión lineal simple. En este caso particular, significa que:

- En cuanto al entorno socioeconómico del alumno (eje *X*):
  - Los centros no cubren la variedad de entornos socioeconómicos que existen a nivel de la población. Al centro 1 acuden principalmente alumnos de alto nivel socioeconómico, mientras que al centro 4 acuden alumnos de bajo nivel socieconómico.
  - Por lo tanto, los centros reciben alumnos de distinto entorno socioeconómico, como mostrarían las proyecciones de los puntos pequeños en el eje *X*. Dicho de otro modo, existe una segregación significativa al nivel de los centros.

- En cuanto al rendimiento del alumno en matemáticas (eje *Y*):
  - Los centros no cubren los distintos niveles de rendimiento de los alumnos que existen al nivel de la población. Al centro 1 asisten sobre todo alumnos de alto rendimiento y al centro 4, de bajo rendimiento.
  - La diferencia entre el nivel medio de los centros es considerable, como muestran las proyecciones de los puntos pequeños en el eje *Y*. En el país 4, la varianza del rendimiento de los centros es, por tanto, muy importante.

- En cuanto a la relación entre el entorno socioeconómico y el rendimiento en matemáticas:
  - En ningún centro existe relación entre el entorno socioeconómico y el rendimiento.
  - Dentro de un centro determinado, el entorno socioeconómico del alumno no importa. Lo que sí importa es el centro al que asiste. Sin embargo, el entorno socioeconómico del alumno decidirá a qué centro acudirá.

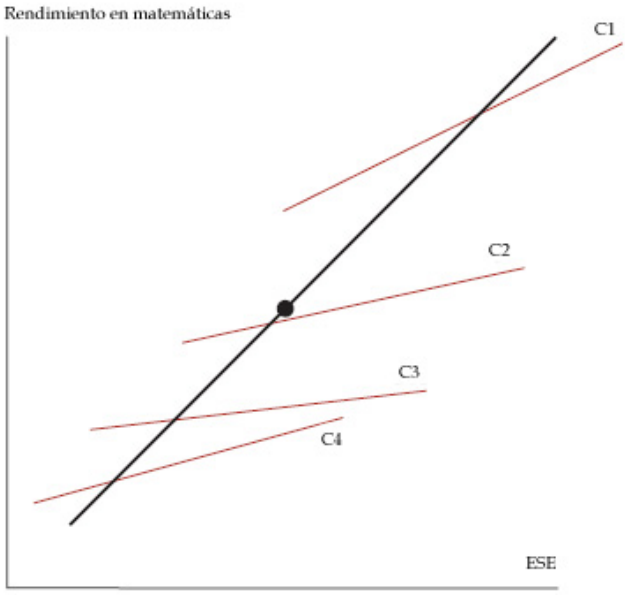
Los países 2 y 3 presentan situaciones intermedias entre estos ejemplos extremos.

El efecto fijo frente al efecto aleatorio

En los casos examinados hasta el momento, las líneas de regresión dentro del centro eran todas paralelas, pero los análisis de regresión multinivel también permiten que la pendiente de regre-

sión sea distinta. En el primer caso, el efecto de *X* se considerará fijo, mientras que en el último caso, el efecto se considerará aleatorio. La figura 13.5 ilustra un caso con efecto aleatorio.

Figura 13.5. Un modelo aleatorio multinivel



Matemáticamente, en el caso de una sola variable predictora, los dos modelos pueden diferenciarse de la forma siguiente:

- para un efecto fijo:

$$Y_{ij} = \alpha_j + \beta X_{ij} + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + U_{0j}$$

- para un efecto aleatorio:

$$Y_{ij} = \alpha_j + \beta_j X_{ij} + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + U_{0j},$$

$$\beta_j = \gamma_{10} + U_{1j}$$

El subíndice *i* de las ecuaciones se refiere al alumno[1] (también llamado nivel 1 en la literatura sobre los modelos multinivel) y el subíndice *j* se refiere al centro (o nivel 2). En una ecuación, la presencia del subíndice *j* para un coeficiente de regresión significa que puede variar de un centro a otro.

El término ε_{ij} designa el residuo de la ecuación, es decir, la diferencia entre la puntuación observada Y_{ij} y la puntuación predicha $\hat{Y}_{ij}$. Este residuo se distribuye normalmente con una media de 0 y una varianza constante en el nivel 1 (es decir, el nivel del alumno), normalmente ex-

presada como σ^2.

Como muestran estas dos ecuaciones, el intercepto α_j se considera siempre como un efecto aleatorio. Considerar el intercepto como un parámetro fijo reduciría el modelo multinivel a un análisis de regresión lineal. El intercepto α_j puede dividirse en una parte fija (es decir, γ_{00} expresa el intercepto general y es igual a la media de los interceptos de centros α_j) y, en segundo lugar, en una parte aleatoria (es decir, U_{0j} expresa la distancia al centro desde el intercepto general). Se asume que esta distancia al centro U_{0j} tiene una media de 0 y una varianza de τ_0^2.

El coeficiente β de la primera ecuación no tiene subíndice *j*, lo que significa que el efecto de *X* no puede variar de un centro a otro. Por lo tanto, las líneas de regresión son paralelas y el efecto de *X* se considera fijo. En cambio, el coeficiente β de la segunda ecuación tiene un subíndice *j*, lo que significa que puede variar de un centro a otro. Las líneas de regresión ya no son paralelas y el efecto de *X* se considera ahora de tipo aleatorio. Igual que antes, este coeficiente de regresión β_j puede dividirse en una parte fija y otra parte aleatoria. La parte fija γ_{10} se llama *coeficiente de regresión global* y corresponde a la media de los coeficientes de regresión β_j. La parte aleatoria U_{0j} es la distancia al centro desde el coeficiente de regresión global. Tiene una media de 0 y una varianza llamada τ_1^2.

Los efectos aleatorios y los efectos fijos pueden combinarse en un único análisis de regresión multinivel. Por ejemplo, en la siguiente ecuación, se introducen en el modelo dos variables predictoras del resultado de los alumnos: una, X_1, considerada fija, y otra, X_2, considerada aleatoria:

$$Y_{ij} = \alpha_j + \beta_1 X_{1ij} + \beta_{2j} X_{2ij}$$

Algunos ejemplos con SPSS®

Normalmente hay dos tipos de índices relevantes en los análisis multinivel: los coeficientes de regresión y la descomposición de la varianza entre los distintos niveles, es decir, el nivel de alumnos (o nivel 1) y el nivel de centros (o nivel 2).

Los análisis de regresión multinivel siempre informan de la varianza del residuo en los distintos niveles —la varianza entre centros y la varianza dentro del centro— que no se explican por medio de las variables predictoras incluidas en el modelo.

Sin embargo, los informes científicos suelen mostrar la varianza explicada. La conversión de la varianza del residuo en un porcentaje de varianza explicada sólo exige la comparación de los coeficientes de varianza de los centros y de los alumnos con sus respectivos coeficientes de varianza residual.

Ejemplo 1

La descomposición de la varianza total puede obtenerse fácilmente con un modelo de regresión multinivel. El siguiente modelo [designado habitualmente como modelo vacío o *nulo* – N. del T.]:

$$Y_{ij} = \alpha_j + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + U_{0j}$$

proporcionará estimaciones sin sesgo de la varianza entre centros y dentro del centro. Como el modelo de regresión no tiene variables predictoras, los interceptos de los centros, es decir, α_j, serán por tanto iguales a las medias de los centros o cercanos a ellas. La varianza de U_{0j} será igual a la varianza entre centros. Puesto que a cada alumno se le asignará la media de su centro como puntuación predicha, la varianza de ε_{ij} será igual a la varianza dentro del centro.

SPSS® ofrece dos procedimientos para los análisis multinivel: VARCOMP permite a los investigadores realizar una descomposición multinivel de la varianza; MIXED es un procedimiento que puede emplearse para trabajar con modelos multinivel. Los procedimientos MIXED y VARCOMP de SPSS® permiten calcular modelos de regresión multinivel. Sin embargo, ambos procedimientos exigen la normalización de los pesos, es decir, que la suma de los pesos sea igual al número de alumnos en el conjunto de datos. Si se usa la claúsula BY, la normalización se realizará según las categorías de la variable de agrupación.

El cuadro 13.2 presenta la sintaxis de SPSS® para esta normalización, así como un breve procedimiento de comprobación.[2] Adviértase que para guardar las estimaciones de la varianza en un archivo de salida, es necesario utilizar la sub-instrucción `outfile` de VARCOMP (véase el cuadro 13.3).

Cuadro 13.2. Normalización de los pesos finales de PISA 2003

```
GET FILE 'c:\PISA\Data2003\INT_stui_2003.sav'.
SORT CASES BY cnt schoolid stidstd.

*** CALCULAR LOS PESOS NORMALIZADOS ***.
WEIGHT OFF.
AGGREGATE OUTFILE= 'c:\temp\temp1.sav'
   /BREAK = cnt
   /popwgt=SUM(w_fstuwt)
   /smpsize=NU.
EXEC.

MATCH FILE FILE=*
    /TABLE = "c:\temp\temp1.sav"
    /BY cnt.
EXEC.

COMPUTE std_wgt=(w_fstuwt/popwgt)*smpsize.
EXEC.

*** VERIFICACIÓN ***.
WEIGHT OFF.
FREQ cnt.
WEIGHT BY std_wgt.
FREQ cnt.

WEIGHT OFF.
SAVE OUTFILE = 'c:\temp\INT_stui_2003.sav'.
```

El cuadro 13.3 presenta la sintaxis de SPSS® para un modelo de regresión multinivel, así como la sintaxis de SPSS® para el cálculo de la correlación intraclase.

Cuadro 13.3. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 1

```
*** VARCOMP ***.

WEIGHT OFF.
SPLIT FILE BY CNT.
VARCOMP
  pv1math BY schoolid
  /RANDOM = schoolid
  /METHOD = ML
  /OUTFILE = VAREST("c:\temp\decompvar.sav")
  /REGWGT = std_wgt
  /INTERCEPT = INCLUDE .
SPLIT FILE OFF.

GET FILE "c:\temp\decompvar.sav".
COMPUTE rho = vc1/(vc1+vc2).
FREQ rho.
RENAME VARS (vc1 vc2 = intcept residual).
SAVE OUTFILE = "c:\temp\rho.sav"/ KEEP cnt intcept residual rho.
```

La sub-instrucción `RANDOM` define el segundo nivel de los análisis. La primera cláusula, a continuación del nombre del procedimiento (o bien VARCOMP, o bien MIXED) especifica el modelo,

con la variable dependiente y después las variables predictoras a continuación de la palabra clave WITH. El procedimiento VARCOMP también exige definir el segundo nivel en la cláusula del modelo utilizando la palabra clave BY (por ejemplo: pv1math BY schoolid WITH HISEI).

En el ejemplo del cuadro 13.3, no hay variable predictora incluida en la cláusula que especifica el modelo. Por lo tanto, las varianzas residuales intra-centros e inter-centros serán iguales a las estimaciones de las varianzas intra-centros e inter-centros. La sub-instrucción RANDOM distingue entre variables predictoras fijas y aleatorias, como se ha explicado en la sección anterior. Debería advertirse que, cuando se utilice el procedimiento MIXED, siempre es necesario mencionar intercept. A continuación de la sub-instrucción REGWGT debería ir el peso normalizado a nivel del alumno. Con objeto de obtener resultados según el país, la instrucción debería ir precedida de SPLIT FILE BY (*variable de agrupación*) y seguida de SPLIT FILE OFF.

El procedimiento VARCOMP permite incluir una sub-instrucción OUTFILE que a su vez permite grabar las estimaciones de la varianza en otro archivo externo. Las estimaciones de la varianza se guardarán en el archivo *decompvar.sav*. Otras opciones para la sub-instrucción OUTFILE en VARCOMP son COVB (matriz de covarianza de las estimaciones de varianza) y CORB (matriz de correlación de las estimaciones de varianza).

La tabla 13.2 proporciona las estimaciones de la varianza entre centros y dentro de centro y la correlación intraclase. Estas estimaciones de la varianza se guardaron en el archivo *decompvar.sav*. Como se mostró en el cuadro 13.3, la correlación intraclase[3] es igual a:

$$\rho = \frac{\sigma^2_{entre\_centros}}{\sigma^2_{entre\_centros} + \sigma^2_{dentro\_del\_centro}} = \frac{\tau^2_0}{\tau^2_0 + \sigma^2},$$

donde $\sigma^2_{entre\_centros}$ o τ^2_0 es la varianza entre centros y $\sigma^2_{dentro\_del\_centro}$ o σ^2, la varianza dentro de los centros. En Australia, la varianza entre centros es igual a 1919,11[4] y la varianza dentro de los centros es igual a 7169,09. La correlación intraclase es, por tanto, el porcentaje de la varianza total del que es responsable el centro. Refleja cómo los centros difieren en el rendimiento promedio de los alumnos. En Australia, la correlación intraclase resulta ser igual a 1919,11/(1919,11 + 7169,09) = 0,21. La estimación de la correlación intraclase varía de 0,04 en Islandia a 0,63 en los Países Bajos.

Ejemplo 2

Los siguientes ejemplos se basan en los datos de Luxemburgo. El tamaño de la muestra de centros en Luxemburgo, 29, permitirá la presentación de las estimaciones de los parámetros de centro. En el ejemplo 2, el entorno socioeconómico del alumno, llamado HISEI, se introduce como factor fijo.

PREPARACIÓN DEL ARCHIVO DE DATOS

En las bases de datos de PISA no hay datos perdidos para el peso final ni para la estimación del rendimiento del alumno. Sin embargo, sí hay valores perdidos para variables que podrían usar-

se como predictoras en un modelo de regresión multinivel. Estos datos perdidos plantean dos cuestiones principales:

- La suma de los pesos difiere ligeramente del número de casos que usarán los modelos de regresión. Adviértase que los casos con valores perdidos se retiran automáticamente[5] de cualquier modelo de regresión.
- Las varianzas de centros y de alumnos a partir de modelos diferentes no pueden compararse, ya que los valores perdidos no siempre son aleatorios. Por ejemplo, es menos probable que los alumnos de un contexto socioeconómico bajo respondan a las preguntas acerca de las profesiones de ambos padres.

TABLA 13.2. Estimaciones de la varianza entre centros y dentro de los centros y de la correlación intraclase

País	Varianza entre centros	Varianza dentro de los centros	rho
AUS	1919,11	7169,09	0,21
AUT	5296,65	4299,71	0,55
BEL	7328,47	5738,33	0,56
BRA	4128,49	5173,60	0,44
CAN	1261,58	6250,12	0,17
CHE	3092,60	6198,65	0,33
CZE	4972,45	4557,50	0,52
DEU	6206,92	4498,70	0,58
DNK	1109,45	7357,14	0,13
ESP	1476,85	6081,74	0,20
FIN	336,24	6664,98	0,05
FRA	3822,62	4536,22	0,46
GBR	1881,09	6338,25	0,23
GRC	3387,52	5991,75	0,36
HKG	4675,30	5298,26	0,47
HUN	5688,56	4034,66	0,59
IDN	2769,48	3343,87	0,45
IRL	1246,70	6110,71	0,17
ISL	337,56	7849,99	0,04
ITA	4922,84	4426,67	0,53
JPN	5387,17	4668,82	0,54
KOR	3531,75	5011,56	0,41
LIE	3385,41	5154,08	0,40
LUX	2596,36	5806,97	0,31
LVA	1750,22	6156,52	0,22
MAC	1416,99	6449,96	0,18
MEX	2476,01	3916,46	0,39
NLD	5528,99	3326,09	0,62
NOR	599,49	7986,58	0,07
NZL	1740,61	7969,97	0,18
POL	1033,90	7151,46	0,13
PRT	2647,70	5151,93	0,34
RUS	2656,62	6021,44	0,31
SVK	3734,56	4873,69	0,43
SWE	986,03	8199,46	0,11
THA	2609,38	4387,08	0,37
TUN	2821,00	3825,36	0,42
TUR	6188,40	4891,13	0,56
URY	4457,08	5858,42	0,43

USA	2395,38	6731,45	0,26
YUG	2646,00	4661,59	0,36

Para evitar estos dos problemas, se recomienda borrar cualquier caso con valores perdidos para las diferentes variables predictoras que se usarán en los modelos de regresión antes de la normalización de los pesos. Puesto que los siguientes ejemplos de modelos de regresión multinivel utilizarán dos variables de los alumnos (HISEI para el contexto socioeconómico del alumno y ST03Q01 para su género) y dos variables de los centros (el porcentaje de chicas en el centro, PCGIRLS, y el tipo de centros, SCHLTYPE), los casos con valores perdidos en al menos una de estas cuatro variables deberán ser borrarados antes de normalizar los pesos.

El cuadro 13.4 presenta la sintaxis de SPSS®. Consiste en:

- fusionar el archivo de datos de alumnos y el archivo de datos de centros con las variables de interés;
- borrar los casos en que falte al menos un dato para la variable predictora;
- normalizar el peso.

Antes de borrar los casos con valores perdidos, hay 3923 registros en la base de datos de Luxemburgo. Después de borrar, quedan 3782. Se han borrado aproximadamente un 3,5% de los datos. Si se borran demasiados casos, por ejemplo, más de un 10%, o bien deberían retirarse del análisis las variables con demasiados datos perdidos, o bien deberían usarse métodos de imputación.

EJECUTAR DE NUEVO EL MODELO MULTINIVEL NULO

Después de borrar los casos con valores perdidos con la sintaxis del cuadro 13.4, se ejecuta el modelo multinivel nulo, es decir, un modelo de regresión multinivel sin ninguna variable predictora del cuadro 13.3, para obtener las estimaciones de las varianzas entre centros y dentro de los centros. Las estimaciones de las varianzas inter-centros e intra-centros, guardadas en el archivo *decompvar.sav*, son ahora respectivamente iguales a 2563,30 y 5734,35, en lugar de 2596,36 y 5806,97.

Cuadro 13.4. Sintaxis de SPSS® para normalizar los pesos finales de PISA 2003 con el borrado de casos con valores perdidos en Luxemburgo

```
GET FILE "c:\PISA\Data2003\INT_stui_2003.sav".
SELECT IF (cnt = 'LUX').
EXEC.

SORT CASES BY cnt schoolid stidstd.
SAVE OUTFILE = "c:\temp\LUXstui2003.sav"
    /KEEP cnt schoolid stidstd w_fstuwt pv1math hisei st03q01.

GET FILE "c:\PISA\Data2003\INT_schi_2003.sav".
SELECT IF (cnt = 'LUX').
EXEC.

SORT CASES BY cnt schoolid.
SAVE OUTFILE = "c:\temp\LUXsch2003.sav"
    /KEEP cnt schoolid schltype pcgirls.

MATCH FILE FILE = "c:\temp\LUXstui2003.sav"/
    /TABLE = "c:\temp\LUXsch2003.sav"
    /BY CNT SCHOOLID.
EXEC.

COUNT nbmiss = hisei st03q01 schltype (missing).
SELECT IF (nbmiss = 0).
EXEC.

*** CALCULAR LOS PESOS NORMALIZADOS ***.

WEIGHT OFF.
AGGREGATE OUTFILE= 'c:\temp\temp1.sav'
    /BREAK = cnt
    /popwgt = SUM(w_fstuwt)
    /smpsize = NU.
EXEC.

MATCH FILES FILE=*
    /TABLE = "c:\temp\temp1.sav"
    /BY cnt.
EXEC.

COMPUTE std_wgt = (w_fstuwt/popwgt)*smpsize.
EXEC.

*** VERIFICACIÓN ***.

WEIGHT OFF.
FREQ cnt.
WEIGHT BY std_wgt.
FREQ cnt.
WEIGHT OFF.

SAVE OUTFILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math
            /FIXED = intercept
            /PRINT = G SOLUTION
            /METHOD = ML
            /RANDOM = intercept | SUBJECT(schoolid)
            /REGWGT = std_wgt
            /SAVE = FIXPRED PRED.
```

Calcular el modelo nulo o «vacío» con MIXED sólo tendrá un parámetro fijo γ_{00}, que es 492,36 para los datos de Luxemburgo.

Por desgracia, SPSS® no produce archivos de salida con los parámetros aleatorios en las unidades de segundo nivel. Con un modelo nulo, estos parámetros aleatorios sólo incluirían la distancia residual del centro U_{0j}. La tabla 13.3 es un listado de un archivo de parámetros aleatorios que se calculó utilizando el programa SAS®. Contiene:

- las variables de agrupación usadas en el modelo, es decir, CNT;
- el efecto, es decir, el intercepto o, como se mostrará más tarde, la variable predictora aleatoria, la estimación;
- la variable de nivel superior, es decir, SCHOOLID;
- la estimación;
- el error típico de la estimación;
- el número de grados de libertad (el número de alumnos menos el número de centros);
- el estadístico *t;*
- la probabilidad de que las estimaciones difieran de 0.

Por ejemplo, la distancia residual del centro 00001 del intercepto global 492,36 es sólo 0,71. Esta distancia no difiere de 0, como demuestra el estadístico *t* y su valor de probabilidad asociado. Dicho de otro modo, el intercepto del centro 00001 no es significativamente distinto del intercepto global. Por otra parte, el intercepto del centro 00002 es significativamente más alto que el intercepto global.

FACTOR DE ENCOGIMIENTO (SHRINKAGE)

En el caso de un modelo nulo, podríamos considerar que la suma del intercepto global γ_{00} y una determinada distancia residual de centro U_{0j} deberían ser exactamente iguales a la media de rendimiento del centro.

Los modelos multinivel encogen las distancias residuales de los centros. Para ilustrar este proceso de encogimiento, supongamos que tenemos un sistema educativo con 100 centros. Supongamos que las medias de rendimiento de los centros son idénticas. Dicho de otro modo, la varianza entre centros es igual a 0. Si 20 alumnos se someten a una prueba dentro de cada centro, se espera que las estimaciones de la media del centro difieran ligeramente de las medias de los centros.

Tabla 13.3. Listado del archivo de parámetros aleatorios calculado con SAS®

CNT	Efecto	SCHOOLID	Estimación	Error típico de predicción	Grados de libertad	Valor de *t*	Probab.
LUX	Intercepto	00001	0,71	13,00	3753	0,05	0,96
LUX	Intercepto	00002	66,39	11,63	3753	5,71	0,00
LUX	Intercepto	00003	−23,71	11,03	3753	−2,15	0,03
LUX	Intercepto	00004	−44,68	12,18	3753	−3,67	0,00
LUX	Intercepto	00005	−8,56	10,68	3753	−0,80	0,42
LUX	Intercepto	00006	61,90	11,34	3753	5,46	0,00
LUX	Intercepto	00007	−68,69	12,39	3753	−5,54	0,00
LUX	Intercepto	00008	61,14	11,62	3753	5,26	0,00
LUX	Intercepto	00009	81,64	11,10	3753	7,36	0,00
LUX	Intercepto	00010	−62,00	11,37	3753	−5,45	0,00
LUX	Intercepto	00011	33,19	25,14	3753	1,32	0,19
LUX	Intercepto	00012	−11,35	12,54	3753	−0,91	0,37
LUX	Intercepto	00013	15,56	10,47	3753	1,49	0,14
LUX	Intercepto	00014	8,01	11,25	3753	0,71	0,48
LUX	Intercepto	00015	37,55	12,36	3753	3,04	0,00
LUX	Intercepto	00016	−46,59	10,95	3753	−4,26	0,00
LUX	Intercepto	00017	−33,61	10,98	3753	−3,06	0,00
LUX	Intercepto	00018	−76,02	12,54	3753	−6,06	0,00
LUX	Intercepto	00019	−70,43	12,96	3753	−5,43	0,00
LUX	Intercepto	00020	57,54	11,17	3753	5,15	0,00
LUX	Intercepto	00021	8,04	11,01	3753	0,73	0,47
LUX	Intercepto	00022	−0,67	25,14	3753	−0,03	0,98
LUX	Intercepto	00023	84,27	10,90	3753	7,73	0,00
LUX	Intercepto	00024	29,88	11,12	3753	2,69	0,01
LUX	Intercepto	00025	63,74	11,69	3753	5,45	0,00
LUX	Intercepto	00026	−33,65	11,15	3753	−3,02	0,00
LUX	Intercepto	00027	−8,29	11,53	3753	−0,72	0,47
LUX	Intercepto	00028	−36,89	13,84	3753	−2,66	0,01
LUX	Intercepto	00029	−84,43	10,96	3753	−7,71	0,00

De hecho, dentro de cada centro, pueden seleccionarse para la muestra sobre todo alumnos de alto o bajo rendimiento, de modo que la media del centro se sobreestime o se subestime, respectivamente. Conforme aumenta el número de alumnos seleccionados para la muestra dentro de los centros, es más probable que decrezca la diferencia entre la media del centro y su estimación. Por lo tanto, el factor de encogimiento es inversamente proporcional al número de alumnos seleccionados dentro de los centros.

El factor de encogimiento[6] es igual a:

$$\frac{n_j \sigma^2_{entre\_centros}}{n_j \sigma^2_{entre\_centros} + \sigma^2_{dentro\_del\_centro}},$$

donde n_j es el número de alumnos en el centro j seleccionados para la muestra (Goldstein, 1997).

La tabla 13.4 presenta, para cada centro, el rendimiento medio de matemáticas, el número de alumnos utilizados en el modelo de regresión multinivel, la distancia residual del intercepto global calculada mediante el modelo de regresión multinivel nulo, como se presentó en la tabla 13.1, y la suma del intercepto global γ_{00} y la distancia del centro U_{0j}. El cuadro 13.5 muestra cómo calcular la tabla 13.4 mediante SPSS® (los resultados, en el archivo de salida *schoolmeans.sav*).

Cuadro 13.5. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 2 (1)

```
GET FILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math
          /FIXED = intercept
          /PRINT = G SOLUTION
          /METHOD = ML
          /RANDOM = intercept | SUBJECT(schoolid)
          /REGWGT = std_wgt
          /SAVE = FIXPRED PRED.
compute schresid = PRED_1 - FXPRED_1.
exec.

weight by std_wgt.
aggregate outfile = "c:\temp\schoolmeans.sav"
  /break = schoolid
  /schmn dep fxpred pred = mean (pv1math, schresid, FXPRED_1, PRED_1)
  /nstud = nu.
exec.
```

Tabla 13.4. Rendimiento de los centros en matemáticas, número de alumnos por centro y media corregida

Centro	Media del centro	Número de alumnos	Distancia U_{0j}	$\gamma_{00} + U_{0j}$
00001	493,1	67	0,7	493,1
00002	560,0	120	66,4	558,8
00003	468,3	179	−23,7	468,6
00004	446,6	94	−44,7	447,7
00005	483,7	233	−8,6	483,8
00006	555,2	146	61,9	554,3
00007	421,8	83	−68,7	423,7
00008	554,6	116	61,1	553,5
00009	575,1	167	81,6	574,0
00010	429,4	131	−62,0	430,4
00011	535,2	8	33,2	525,6
00012	480,7	78	−11,3	481,0
00013	508,0	289	15,6	507,9
00014	500,5	150	8,0	500,4
00015	530,9	87	37,6	529,9
00016	445,2	184	−46,6	445,8
00017	458,3	183	−33,6	458,8
00018	414,2	73	−76,0	416,3
00019	419,6	66	−70,4	421,9
00020	550,7	162	57,5	549,9
00021	500,5	174	8,0	500,4
00022	491,5	8	−0,7	491,7
00023	577,6	185	84,3	576,6
00024	522,7	169	29,9	522,2
00025	557,3	117	63,7	556,1
00026	458,2	151	−33,7	458,7
00027	483,9	126	−8,3	484,1
00028	453,9	53	−36,9	455,5
00029	406,9	183	−84,4	407,9

Como se ha mostrado, la diferencia entre la media del centro y la suma $\gamma_{00} + U_{0j}$ es:

- directamente proporcional a la distancia residual del centro; es decir, el factor de encogimiento afecta principalmente a los centros con alto o bajo rendimiento;
- inversamente proporcional al número de alumnos observados en el centro.

INTRODUCCIÓN DE HISEI COMO EFECTO FIJO

Con la introducción de la variable del alumno HISEI como efecto fijo, la ecuación puede escribirse así:

$$Y_{ij} = \alpha_j + \beta_1(HISEI)_{ij} + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + U_{0j}$$

La sintaxis de SPSS® para este modelo se presenta en el cuadro 13.6 y algunos fragmentos de la salida de SPSS®, en el cuadro 13.7.

Cuadro 13.6. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 2 (2)

```
GET FILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math WITH hisei
            /FIXED = intercept hisei
            /PRINT = G SOLUTION
            /METHOD = ML
            /RANDOM = intercept | SUBJECT(schoolid)
            /REGWGT = std_wgt.
```

Cuadro 13.7. Salida de SPSS® (ejemplo 2)

Estimaciones de parámetros de covarianza[a,b]

Parámetro	Estimación	Error típico
Residuo	5551,5060563	128,1500411
Varianza del intercepto [unidad de análisis = SCHOOLID]	1950,3945680	530,9974935

Estimaciones de efectos fijos[a,b]

Parámetro	Estimación	Error típico	Grados de libertad	*t*	Sig.	Intervalo de confianza del 95%	
						Límite inferior	Límite superior
Intercepto	446,7649734	9,2614577	43,341	48,239	,000	428,0917086	465,4382383
HISEI	,9479007	,0823676	3780,841	11,508	,000	,7864114	1,1093899

[a] Variable dependiente: primer valor plausible en matemáticas.
[b] Residuo ponderado mediante std_wgt.

Sólo se ha introducido un cambio en comparación con la sintaxis presentada en el cuadro 13.5. El nombre HISEI se ha añadido a la cláusula que especifica el modelo.

El intercepto global γ_{00} es ahora igual a 446,76 y el coeficiente de regresión dentro de centro β_1 es igual a 0,9479. Esto significa que, dentro de un centro, un aumento de 1 unidad en la escala HISEI se asociará con un aumento de 0,9479 en la escala de matemáticas. En comparación, el coeficiente de regresión lineal de HISEI en el rendimiento en matemáticas es igual a 2,05. La relación entre el contexto socioeconómico y el rendimiento del alumno en el sistema educativo de Luxemburgo parece ser similar a la de los ejemplos hipotéticos para el país 2 o el país 3 de la figura 13.4.

Las estimaciones de la varianza residual entre centros y dentro del centro, llamadas respectivamente τ_0^2 y σ^2, son iguales a 1949,09 y 5551,53.

El porcentaje de varianza explicado por la variable HISEI puede calcularse así:

$$1 - \frac{1949,09}{2563,07} = 0,24$$ en el nivel de los centros y

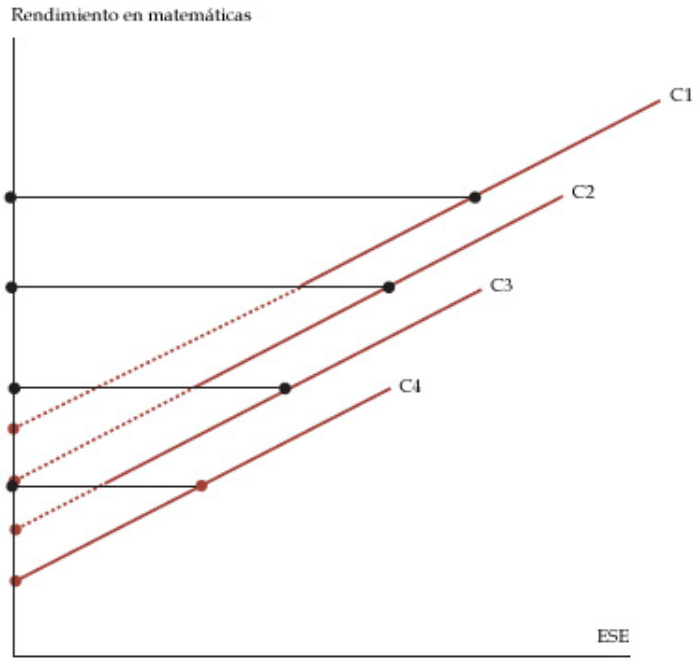
$$1 - \frac{5551,53}{5734,39} = 0,03$$ en el nivel de los alumnos.

¿Cómo puede una variable del alumno explicar alrededor de un 24% de la varianza entre centros y sólo el 3% de la varianza dentro de los centros? Esto refleja principalmente la segregación según el entorno socioeconómico en el centro. Algunos de los centros de Luxemburgo reciben sobre todo alumnos de entornos socioeconómicos altos, mientras que a otros centros acuden sobre todo alumnos de entornos socioeconómicos bajos.

La figura 13.6 proporciona una explicación gráfica de este fenómeno. La varianza entre centros, en cualquier caso, puede representarse gráficamente mediante la variabilidad de los interceptos de los centros en el eje *Y*.

Adviértase que la varianza entre centros puede obtenerse mediante un modelo de regresión multinivel nulo. En ese caso concreto, el intercepto está cercano a la proyección ortogonal del promedio de rendimiento de centros sobre el eje *Y*, como muestran las líneas negras horizontales continuas de la figura 13.6. Como se explicó en la sección anterior, la diferencia entre la media de centros y el intercepto proviene de la aplicación del factor de encogimiento.

Figura 13.6. Representación gráfica de la reducción de la varianza entre centros



La varianza residual entre centros puede obtenerse mediante la prolongación de la línea de regresión en el eje *Y*, como muestran las líneas de puntos de la figura 13.6. Como puede verse, la amplitud de los interceptos provenientes de las líneas horizontales continuas es superior a la de los interceptos provenientes de las líneas inclinadas de puntos.

En términos generales, una variable del alumnos influirá sobre la varianza entre centros si:

- los centros difieren en la media y en el rango de los alumnos en cuanto a esa variable (véanse los países 2, 3 y 4 de la figura 13.4);
- el coeficiente de regresión intra centros de esa variable es distinto de 0. El país 4 en la figura 13.4 ilustra un caso donde usar la variable HISEI al nivel de los alumnos en el modelo no reducirá la varianza entre centros. Por otra parte, la introducción del estatus socioeconómico del centro, es decir, la media en HISEI del centro, influirá considerablemente en la varianza entre centros en el caso del país 4.

Ejemplo 3

El ejemplo 3 es similar al ejemplo 2, excepto en que HISEI se considera ahora un efecto aleatorio. La sintaxis de SPSS® se presenta en el cuadro 13.8. La ecuación puede ahora escribirse así:

$$Y_{ij} = \alpha_j + \beta_{1j}(HISEI) + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j}$$

Cuadro 13.8. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 3

```
GET FILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math WITH hisei
            /FIXED = intercept hisei
            /PRINT = G SOLUTION
            /METHOD = ML
            /RANDOM = intercept hisei | SUBJECT(schoolid)
            /REGWGT = std_wgt.
```

La variable HISEI se ha añadido a la sub-instrucción `RANDOM`.

El archivo de parámetros fijos contiene el intercepto global γ_{00} y el coeficiente de regresión global de HISEI γ_{10}. Del mismo modo que los interceptos de centros, que están divididos en dos partes: un intercepto global y una distancia residual al centro, el coeficiente de regresión dentro de los centros está dividido en dos partes: un coeficiente de regresión global (la parte fija, llamada γ_{10}) y una distancia residual desde el coeficiente de regresión de los centros (la parte aleatoria, llamada U_{1j}).

El intercepto global y el coeficiente de regresión se presentan en la tabla 13.5. El intercepto global es igual a 449,59 y el coeficiente de regresión global de HISEI es igual a 0,89. Como muestran el estadístico *t* y su probabilidad asociada, ambos parámetros son significativamente distintos de 0.

Tabla 13.5. Resultados de los parámetros fijos en la salida del procedimiento

CNT	Efecto	Estimación	Error típico de predicción	Valor de *t*	Prob.
LUX	Intercepto	449,59	9,69	46,39	0,00
LUX	HISEI	0,89	0,11	8,17	0,00

El archivo de parámetros aleatorios lista las distancias residuales del centro:

- U_{0j} desde el intercepto γ_{00}, es decir, 449,59;
- U_{1j} desde el coeficiente de regresión γ_{10}, es decir, 0,89.

Puesto que HISEI se considera ahora un efecto aleatorio, no tiene sentido interpretar la distancia residual del centro desde el intercepto global. La tabla 13.6 presenta la distancia residual del centro desde el coeficiente de regresión HISEI global para los primeros 13 centros.[7]

Tabla 13.6. Resultados de los parámetros aleatorios calculados con SAS®

CNT	Efecto	Centro	Estimación	Error típico de predicción	Grados de libertad	Valor de *t*	Prob.
LUX	HISEI	00001	0,22	0,31	3724	0,71	0,48
LUX	HISEI	00002	0,04	0,26	3724	0,15	0,88
LUX	HISEI	00003	0,29	0,26	3724	1,13	0,26
LUX	HISEI	00004	−0,51	0,29	3724	−1,75	0,08
LUX	HISEI	00005	−0,08	0,25	3724	−0,31	0,76
LUX	HISEI	00006	0,07	0,28	3724	0,26	0,79
LUX	HISEI	00007	−0,04	0,29	3724	−0,13	0,90
LUX	HISEI	00008	−0,13	0,27	3724	−0,49	0,62

LUX	HISEI	00009	−0,29	0,25	3724	−1,19	0,23
LUX	HISEI	00010	−0,17	0,26	3724	−0,65	0,52
LUX	HISEI	00011	0,07	0,34	3724	0,19	0,85
LUX	HISEI	00012	−0,04	0,28	3724	−0,14	0,89
LUX	HISEI	00013	0,82	0,22	3724	3,66	0,00

El coeficiente de regresión HISEI para el centro 00001 es igual a 0,89 + 0,22 = 1,11, pero no puede considerarse significativamente distinto del intercepto global. Entre los 13 centros presentados en la tabla 13.6, sólo el centro 00013 presenta un coeficiente de regresión significativamente distinto del coeficiente global, como muestra la probabilidad del estadísticos *t*. El coeficiente de regresión HISEI es igual a 0,89 + 0,82 = 1,71 y, como muestran el estadístico *t* y su probabilidad, este coeficiente de regresión intra-centro es significativamente distinto del coeficiente de regresión global.

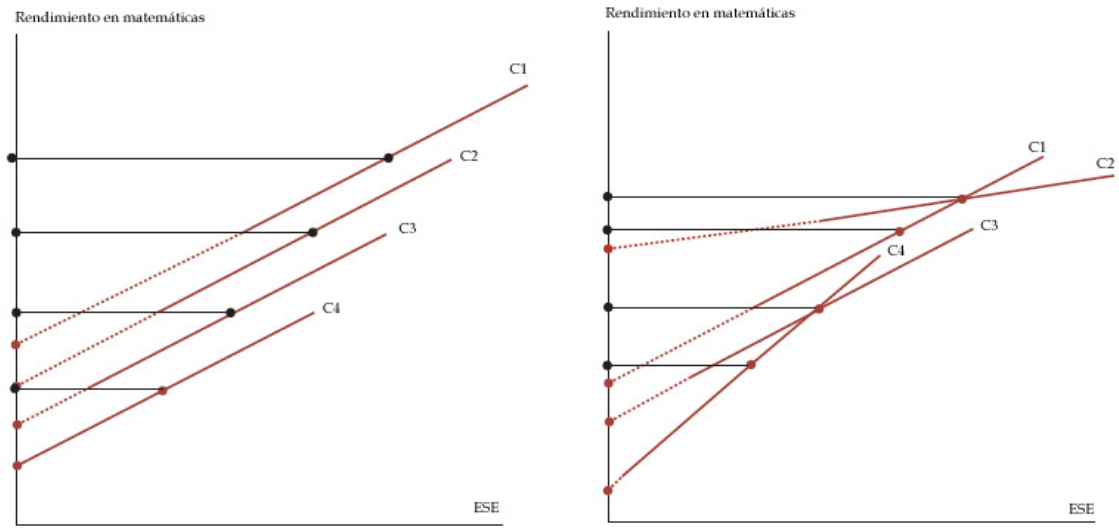
SPSS® proporciona ahora tres estimaciones de varianzas:

- la varianza residual inter-centros τ_0^2, es decir, 2147,64;
- la varianza residual intra-centro σ^2, es decir, 5509,34;
- la varianza de los coeficientes de regresión de HISEI τ_1^2, es decir, 0,1275. Esta también es la variabilidad de la distancia residual desde el coeficiente de regresión.

En comparación con el ejemplo 2, la varianza residual entre centros ha aumentado ligeramente y la varianza residual dentro de los centros ha decrecido ligeramente. La reducción de la varianza dentro de los centros no es sorprendente, ya que el efecto aleatorio sólo puede llevar a un mejor ajuste a los datos.

La figura 13.7 ayuda a comprender el aumento de la varianza residual entre centros. El coeficiente de regresión para el centro 00001 (C1) tiene una pendiente algo menos pronunciada, de modo que la prolongación de la línea de regresión caerá más alta que antes sobre el eje *Y*. Además, el coeficiente de regresión tiene una pendiente algo más pronunciada en el centro 00004 (C4), de modo que la extensión de la línea de regresión caerá algo más baja en el eje *Y*.

FIGURA 13.7. Reducción de la varianza residual entre centros para un modelo fijo y para uno aleatorio



Ejemplo 4

En el ejemplo 4, el género del alumno, llamado ST03Q01 en la base de datos de PISA, se añade al modelo anterior como factor fijo. La ecuación puede escribirse así:

$$Y_{ij} = \alpha_j + \beta_{1j}(HISEI)_{ij} + \beta_2(ST03Q01)_{ij}$$

$$\alpha_j = \gamma_{00} + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j}$$

El cuadro 13.9 presenta la sintaxis de SPSS®.

Cuadro 13.9. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 4

```
GET FILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math WITH hisei st03q01
            /FIXED = intercept hisei st03q01
            /PRINT = G SOLUTION
            /METHOD = ML
            /RANDOM = intercept hisei | SUBJECT(schoolid)
            /REGWGT = std_wgt.
```

Los parámetros fijos son, respectivamente, iguales a 419,68 para el intercepto global, 0,86 para el coeficiente de regresión de HISEI global y 20,7927 para el coeficiente de género global.

La varianza residual entre centros τ_0^2 es igual a 2167,41 y la varianza residual dentro de centro σ^2 es igual a 5415,34. Por último, la varianza del coeficiente de regresión de HISEI de centros

τ_1^2 es igual a 0,1313.

Este modelo explica $1-\frac{2167,41}{2563,07}=15,3\,\%$ de la varianza entre centros y $1-\frac{5415,34}{5734,39}=5,6\,\%$ de la varianza dentro del centro.

El coeficiente de regresión para el género de 20,8 refleja la diferencia según el género esperada dentro de cualquier centro, después de eliminar el efecto de HISEI.

Cuadro 13.10. Interpretación del coeficiente de regresión dentro del centro

La diferencia esperada según el género dentro del centro puede diferir en gran medida de la diferencia global según el género, sobre todo en un sistema con muchos itinerarios tempranos. Parece que las chicas tienden más a elegir un itinerario académico, mientras que es más probable que los chicos escojan un itinerario de formación profesional. El coeficiente de regresión lineal para el género en el rendimiento de los alumnos no tiene en cuenta esta escolarización diferenciada. Si los distintos itinerarios son impartidos por distintos centros, como en Alemania, por ejemplo, un modelo de regresión multinivel tendrá en cuenta esta escolarización diferenciada, de modo que el coeficiente de regresión multinivel de género será considerablemente distinto del coeficiente de regresión lineal. La tabla siguiente proporciona los coeficientes de regresión lineal y multinivel para el género a partir de los datos de PISA 2003 en Alemania.

A nivel de la población, los chicos superan a las chicas en 8,9 puntos en matemáticas, mientras que las chicas superan a los chicos en 42,1 puntos en lectura. Sin embargo, dentro de un determinado centro, las diferencias esperadas en matemáticas y lectura son, respectivamente, iguales a 30,7 y −19,3.

Diferencias según el sexo en Alemania

	Matemáticas	Lectura
Coeficiente de regresión lineal simple	8,9	−42,1
Coeficiente de regresión multinivel	30,7	−19,3

El género puede también considerarse un factor aleatorio. En ese caso, la ecuación puede escribirse así:

$$Y_{ij}=\alpha_j+\beta_{1j}\left(HISEI\right)_{ij}+\beta_{2j}\left(ST03Q01\right)_{ij}$$

$$\alpha_j=\gamma_{00}+U_{0j}$$

$$\beta_{1j}=\gamma_{10}+U_{1j}$$

$$\beta_{2j}=\gamma_{20}+U_{2j}$$

El cuadro 13.11 presenta la estimación de la varianza de los parámetros aleatorios, así como las estimaciones de los coeficientes de regresión de las partes fijas del modelo.

Cuadro 13.11. Salida de SPSS® (ejemplo 4)

Estimaciones de efectos fijos[a,b]

Parámetro	Estimación	Error típico	Grados de libertad	*t*	Significación
Intercepto	419,3613485	10,0172881	43,829	41,864	,000
HISEI	,8606827	,1097952	38,918	7,839	,000
ST03Q01	21,0238222	3,1530028	31,424	6,668	,000

Estimaciones de parámetros de covarianza[a,b]

Parámetro		Estimación	Error típico
Residuo		5400,8730832	125,3934221
Intercepto [unidad de análisis = SCHOOLID]	Varianza	1904,7788816	613,6594064
HISEI [unidad de análisis = SCHOOLID]	Varianza	,1348145	,0714020
ST03Q01 [unidad de análisis = SCHOOLID]	Varianza	70,5719428	56,7028751

[a] Variable dependiente: primer valor plausible en matemáticas.
[b] Residuo ponderado mediante std_wgt.

Como se muestra en el cuadro 13.11, la variabilidad de U_{2j}, es decir, la distancia residual del centro desde el coeficiente de regresión del géneros es bastante grande. Esto indica que las diferencias según el género varían de un centro a otro.

Ejemplo 5

La última ecuación del ejemplo 4 fue $Y_{ij} = \alpha_j + \beta_{1j}(HISEI)_{ij} + \beta_{2j}(ST03Q01)_{ij}$. Esta ecuación refleja el modelo de la variabilidad de rendimiento del alumno dentro de los centros al introducir variables predictoras del nivel de los alumnos. Sin embargo, debido al efecto de la segregación, estas variables predictoras del nivel de los alumnos pueden explicar parte de la varianza entre centros.

También es posible introducir una variable predictora del nivel de los centros. Supongamos que estamos interesados en el efecto del tipo de centro sobre el rendimiento medio del centro. La ecuación puede escribirse así:

$$Y_{ij} = \alpha_j + \beta_{1j}(HISEI)_{ij} + \beta_{2j}(ST03Q01)_{ij} + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + \gamma_{01}(SCHLTYPE)_j + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + U_{2j}$$

Dicho de otro modo, puesto que la variable 'tipo de centro' es idéntica para todos los alumnos dentro de un centro determinado, esta variable sólo influirá sobre los interceptos de centro.

Dados el entorno socioeconómico y la composición según el género de los centros, ¿explica el tipo de centro por qué algunos centros tienen mejor rendimiento del esperado y por qué en otros centros ocurre a la inversa?

La sintaxis de SPSS® se presenta en el cuadro 13.12.

Cuadro 13.12. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 5 (1)

```
GET FILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math WITH hisei st03q01 schltype
            /FIXED = intercept hisei st03q01 schltype
            /PRINT = G SOLUTION
            /METHOD = ML
            /RANDOM = intercept hisei | SUBJECT(schoolid)
            /REGWGT = std_wgt.
```

La tabla 13.7 presenta los resultados de los parámetros fijos.

Tabla 13.7. Parámetros fijos (ejemplo 5)

CNT	Efecto	Estimación	Error típico de predicción	Grados de libertad	Valor de *t*	Prob.
LUX	Intercepto	320,47	66,69	27	4,81	0,00
LUX	HISEI	0,86	0,11	28	7,84	0,00
LUX	ST03Q01	20,69	2,59	3723	7,98	0,00
LUX	SCHLTYPE	35,14	23,36	3723	1,50	0,13

Como muestra la tabla 13.7, la variable 'tipo de centro' no es significativa. Dicho de otro modo, no puede establecerse que los centros privados difieran de los centros públicos una vez que se elimina el efecto de las variables del entorno socioeconómico y del género de los alumnos.

Ejemplo 6

Por último, el modelo puede extenderse para comprender por qué varían los coeficientes de regresión de centro HISEI y ST03Q01. Las dos hipótesis que se pondrán a prueba son:

- los coeficientes de regresión de HISEI difieren entre los centros públicos y privados;
- los coeficientes de regresión de ST03Q01 están relacionados con el porcentaje de chicos y chicas del centro. La ecuación puede escribirse así:

$$Y_{ij} = \alpha_j + \beta_{1j}\left(HISEI\right)_{ij} + \beta_{2j}\left(ST03Q01\right)_{ij} + \varepsilon_{ij}$$

$$\alpha_j = \gamma_{00} + \gamma_{01}\left(SCHLTYPE\right)_j + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\left(SCHLTYPE\right)_j + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}\left(PCGIRLS\right)_j + U_{2j}$$

El cuadro 13.13 presenta la sintaxis de SPSS® para ejecutar este modelo. Comprobar si los coeficientes de regresión de HISEI difieren según el tipo de centro resulta parecido a comprobar la interacción entre el tipo de centro y los coeficientes de regresión de HISEI. Por lo tanto, en SPSS®, el término `hisei*schltype` debe añadirse en la sub-instrucción FIXED, así como `st03q01*pcgirls`. Adviértase que la especificación del modelo que sigue al nombre del procedimiento lista `schltype` y `pcgirls` sin los términos de interacción.

Cuadro 13.13. Sintaxis de SPSS® para un modelo de regresión multinivel: ejemplo 5 (2)

```
GET FILE = "c:\temp\LUX2003.sav".
weight off.
MIXED pv1math WITH hisei st03q01 schltype pcgirls
   /FIXED = intercept hisei st03q01 schltype hisei*schltype st03q01*pcgirls
   /PRINT = G SOLUTION
   /METHOD = ML
   /RANDOM = intercept hisei st03q01| SUBJECT(schoolid)
   /REGWGT = std_wgt.
```

Cuadro 13.14. Resultados de SPSS® (ejemplo 5)

Estimaciones de efectos fijos[a,b]

Parámetro	Estimación	Error típico	Grados de libertad	*t*	Significación
Intercepto	291,4022663	71,2240565	39,885	4,091	,000
HISEI	1,8607710	,9019687	51,684	2,063	,044
ST03Q01	19,8852339	11,7315332	50,628	1,695	,096
SCHLTYPE	45,1271348	24,8610903	39,203	1,815	,077
HISEI*SCHLTYPE	−,3504893	,3137546	50,175	−1,117	,269
ST03Q01*PCGIRLS	2,6190012	24,8681643	49,929	,105	,917

Estimaciones de parámetros de covarianza[a,b]

Parámetro		Estimación	Error típico
Residuo		5397,4444256	125,3180631
Intercepto [unidad de análisis = SCHOOLID]	Varianza	1757,3290219	578,2565538
HISEI [unidad de análisis = SCHOOLID]	Varianza	,1426724	,0735291
ST03Q01 [unidad de análisis = SCHOOLID]	Varianza	71,0821703	56,3930823

[a] Variable dependiente: primer valor plausible en matemáticas.
[b] Residuo ponderado mediante std_wgt.

La tabla 13.8 presenta los parámetros fijos de la ecuación. Como muestran los resultados de SPSS® en el cuadro 13.14, el coeficiente de regresión de HISEI aleatorio no se asocia significativamente con el tipo de centro.

Tabla 13.8. Estimaciones de los parámetros fijos (ejemplo 6)

Efecto	Estimación del coeficiente	Coeficiente
Intercepto	291,40	γ_{00}
HISEI	1,86	γ_{10}
ST03Q01	19,88	γ_{20}
SCHLTYPE	45,12	γ_{01}
HISEI*SCHLTYPE	−0,35	γ_{11}
ST03Q01*PCGIRLS	2,62	γ_{21}

Tabla 13.9. Estimaciones de la varianza de los parámetros aleatorios (ejemplo 6)

Efecto	Estimación de la varianza	Coeficiente
Intercepto	1757,37	U_{0j}
HISEI	0,1427	U_{1j}
ST03Q01	71,0154	U_{2j}
Residuo	5397,46	ε_{ij}

Como muestra la probabilidad obtenida, es preciso aceptar las dos hipótesis nulas; es decir, el tipo de centro no está asociado con las pendientes de HISEI y la diferencia según el género dentro de centro no está asociada con el porcentaje de chicas en el centro.

Limitaciones del modelo multinivel en el contexto de PISA

Esta sección pretende alertar a los posibles usuarios de los datos de PISA acerca de las limitaciones o los peligros de tales modelos en el contexto de PISA.

Dichos modelos están diseñados para descomponer la varianza de alumnos en:

- varianza entre centros;
- varianza dentro de los centros;
- varianza dentro de las clases.

Puesto que PISA selecciona por cada centro participante una muestra aleatoria de una población de la misma edad de entre todos los cursos y clases, permite la descomposición de la varianza en sólo dos niveles: una varianza entre centros y una varianza dentro de los centros. Además, se espera que la varianza global sea mayor con una muestra de edad que con una muestra de curso, a menos que la población de la misma edad estudie en un único curso, como ocurre en Islandia o Japón.

Para permitir comparaciones internacionales significativas, estos tipos de indicadores exigen una definición común para lo que se debe considerar un centro y una clase. Si bien no hay graves problemas para definir qué es un alumno, sí hay diferencias importantes de un país a otro en cuanto a qué es un centro y qué es una clase.

Las encuestas internacionales sobre educación están sobre todo interesadas en la muestra de alumnos, por lo que podríamos considerar la muestra de centros como un paso necesario para seleccionar una muestra eficiente de alumnos que minimice el coste de las pruebas. En este contexto, la definición de qué es un centro o qué es una clase no plantea mayores problemas. Sin

embargo, la cada vez mayor importancia y popularidad de los análisis multinivel exigen dedicar más atención a las definiciones.

PISA 2000 y PISA 2003 no dan una definición detallada de lo que se debe considerar un centro. Durante los procedimientos de muestreo se dió prioridad a desarrollar una lista de unidades que garantizara una cobertura completa de la población escolarizada de 15 años y que, además, aportara tasas de respuesta aceptables. Una vez que se seleccionaba un centro, también tenía que ser práctico seleccionar alrededor de 35 alumnos dentro de ese centro para someterlos a evaluación. De este modo, se construyó el marco muestral de centros teniendo en cuenta cuestiones de cobertura de alumnos e implementación práctica de PISA, más que consideraciones analíticas. Por tanto, en las bases de datos de PISA, es posible que lo que se considera como centro represente diferentes instituciones educativas que quizá no sean comparables sin tener en cuenta alguna restricción. Por ejemplo, en algunos países, los centros se definen como unidades administrativas que pueden consistir en diversos edificios, no necesariamente cercanos entre sí. Otros países utilizaron el edificio como unidad muestral de centros y, por último, algunos países definieron un centro como una rama o itinerario dentro de un edificio determinado. Es probable que cuanto mayores sean estas agrupaciones tanto menores serán las diferencias entre ellas y tanto mayores las diferencias dentro de ellas. En este contexto, esperaríamos observar altas correlaciones intraclase en estos países y un coeficiente de regresión dentro del centro no significativo para el entorno socioeconómico del alumno (Kirsch y otros, 2002).

Además de este problema de una definición internacional de lo que debe considerarse un centro, los usuarios de los datos deben ser conscientes de las siguientes cuestiones:

- La elección de una definición de centro en un país determinado puede estar dictada por la disponibilidad de los datos. De hecho, los centros nacionales de investigación deben incluir en el marco muestral de centros una medida del tamaño de la población de alumnos de 15 años (véase el capítulo 2). Esta información quizá esté disponible en el ámbito de la unidad administrativa, pero no en el ámbito de cada edificio. En los estados federales que cuentan con varios sistemas educativos, los datos disponibles podrían diferir entre un sistema y otro, de modo que el concepto de centro podría variar incluso dentro del mismo país.
- Por razones prácticas u operativas, el concepto de centro podría variar entre dos recogidas de datos de PISA. Por ejemplo, algunos países utilizaron las unidades administrativas en el marco muestral de centros de PISA 2000 y las unidades de edificios en el marco muestral de centros de PISA 2003. Tales cambios se implementaron para aumentar la tasa de participación de centros. Estos cambios conceptuales influirán en los resultados de cualquier descomposición de varianzas y también podrían afectar a los resultados de los modelos multinivel. Cambiar de una definición administrativa a una definición de edificios aumentará la correlación intraclase y debería disminuir la pendiente del coeficiente de regresión dentro de los centros. Si ocurren tales cambios en un país, se recomienda vivamente no calcular ninguna tendencia a partir de descomposiciones de las varianzas o regresiones multinivel.

Como demuestra este ejemplo, los análisis multinivel y los análisis de descomposición de varianzas deben interpretarse a la luz de:

- la estructura de los sistemas educativos;

- la definición de *centro* utilizada en el marco muestral de centros.

Dentro de las limitaciones comentadas en esta sección, los análisis de regresión multinivel son sin duda adecuados y apropiados para describir cómo se asignan los alumnos a los centros y cuáles son los criterios principales para tal asignación. Sin embargo, 10, incluso 20, variables de alumnos y centros nunca podrán captar la complejidad de un sistema educativo. Además, PISA está midiendo un proceso acumulativo de aproximadamente diez años de escolarización. Lo que hacemos hoy, sin duda, no puede explicar lo que somos hoy. Por consiguiente, las prácticas pedagógicas y el entorno escolar en que los jóvenes de 15 años aprenden actualmente es incapaz de explicar por completo el rendimiento actual de estos escolares. En este contexto, las recomendaciones sobre medidas políticas deberían realizarse e interpretarse con precaución.

Conclusiones

Este capítulo describe, en primer lugar, el concepto de análisis multinivel y cómo realizar tales modelos con SPSS®. Empieza con el modelo más sencillo, llamado modelo nulo, y después aumenta progresivamente la complejidad añadiendo variables. Por último, en el contexto de PISA, se comentan cuestiones metodológicas importantes que limitan la comparabilidad internacional de los resultados.

---

[1] Para mantener la consistencia con la literatura sobre regresión multinivel, los subíndices *i* y *j* se han invertido en relación con el uso que se hace de ellos en el capítulo 2.

[2] Multiplicar el peso total de alumnos W_FSTUWT por la variable CNTFAC2 produce los mismos pesos (COMPUTE std_wgt = w_fstuwt * cntfac2) que la sintaxis del cuadro 13.2. Sin embargo, los pesos normalizados resultantes sólo deberían usarse para modelos multinivel basados en variables sin ningún valor perdido. Al calcular modelos multinivel incluyendo variables con valores perdidos, un peso normalizado distinto debería calcularse después de borrar todos los casos con valores perdidos.

[3] Véase también la tabla 4.4 del capítulo 4.

[4] Adviértase que los resultados obtenidos a partir de las sintaxis de SPSS® en este capítulo pueden diferir a veces de los ejemplos del texto, que se calcularon utilizando el programa SAS®. Estas desviaciones menores se deben a ciertas diferencias en los algoritmos utilizados por ambos paquetes de *software* estadístico.

[5] Sin embargo, puede usarse como archivo de datos para un análisis de regresión lineal una matriz de correlación calculada con la opción de exclusión de casos según pareja.

[6] Este factor de encogimiento debe asociarse a la Media Cuadrática Esperada de los centros en un modelo ANOVA. De hecho, $E\left(MS_{centro}\right) = n_j \sigma^2_{entre\_centros} + \sigma^2_{dentro\_del\_centro}$.

[7] Adviértase que este archivo de salida no está disponible con SPSS®; los resultados se presentan aquí para servir de ilustración y se calcularon con SAS®.

Capítulo 14

Otras cuestiones estadísticas

Introducción

Los informes iniciales de PISA 2000 y PISA 2003 incluyeron descripciones de la relación entre índices de cuestionario y rendimiento de los alumnos, mediante la división de los índices de cuestionario en cuartiles y, después, presentando el rendimiento medio por cuartil. Los informes de PISA también incluían los conceptos estadísticos de riesgo relativo y de riesgo atribuible. Este capítulo se dedica a estas dos cuestiones.

Los análisis en cuartiles

Como ya se ha mencionado en el capítulo 4, los índices derivados de los datos de cuestionario se generaron mediante un modelo de Rasch y las estimaciones de los alumnos se presentaron mediante *estimaciones de verosimilitud ponderadas* (WLE). Ya se ha mencionado que una WLE es una variable discontinua.

La tabla 14.1 presenta la distribución del índice de cuestionario «interés y gusto por las matemáticas» a partir del conjunto de datos de PISA 2003 en Alemania. Esta tabla muestra claramente el carácter discontinuo de la variable.

Para dividir un índice de cuestionario en cuartiles, deben calcularse los percentiles 25º, 50º y 75º. Estos percentiles son, respectivamente, −0,6369, 0,029 y 0,973 para el índice «interés y gusto por las matemáticas» en Alemania.

Existen dos procedimientos de recodificación posibles: menor frente a igual o mayor, o menor o igual frente a mayor.

La sintaxis de SPSS® se presenta en el cuadro 14.1.

Cuadro 14.1. Dos procedimientos para recodificar en cuartiles mediante sintaxis de SPSS®

```
get file 'C:\PISA\Data2003\int_stui_2003.sav'.
Select if cnt = 'DEU'.
Weight by w_fstuwt.
FREQ VARIABLES = intmat /PERCENTILES= 25 50 75.

if (intmat <  -0.6369) q1=1.
if (intmat >= -0.6369 and intmat < 0.029) q1=2.
if (intmat >=  0.029 and intmat <  0.973) q1=3.
if (intmat >=  0.973) q1=4.

if (intmat <= -0.6369) q2=1.
if (intmat >  -0.6369 and intmat <= 0.029) q2=2.
if (intmat >   0.029 and intmat <=  0.973) q2=3.
if (intmat >   0.973) q2=4.
freq q1 q2.
```

Según el procedimiento adoptado, los porcentajes de alumnos en los cuartiles inferior, segundo tercero y superior son, respectivamente, iguales a 24,88, 21,39, 27,80 y 25,93 o 34,53, 21,60, 25,33 y 18,54.

Ninguno de estos dos procedimientos genera cuartiles que incluyan exactamente al 25% de los

alumnos. Puesto que los porcentajes de alumnos en cada cuartil pueden variar según los países, no pueden llevarse a cabo comparaciones internacionales.

Tabla 14.1. Distribución del índice de cuestionario «interés y gusto por las matemáticas» en Alemania

WLE	Porcentaje	Porcentaje acumulado		WLE	Porcentaje	Porcentaje acumulado
−1,783	10,20	10,20		0,477	0,10	64,30
−1,733	0,02	10,23		0,643	0,10	64,40
−1,700	0,02	10,25		0,643	9,53	73,93
−1,469	0,02	10,27		0,869	0,03	73,96
−1,258	7,53	17,80		0,912	0,04	74,00
−1,179	0,02	17,82		0,925	0,05	74,05
−1,147	0,02	17,85		0,940	0,02	74,07
−1,077	0,03	17,88		0,973	7,39	81,46
−0,971	0,08	17,95		1,044	0,03	81,49
−0,929	6,77	24,73		1,146	0,03	81,52
−0,739	0,15	24,88		1,299	5,27	86,79
−0,637	9,66	34,53		1,338	0,02	86,81
−0,619	0,13	34,66		1,346	0,04	86,85
−0,370	0,02	34,68		1,464	0,02	86,87
−0,335	0,07	34,74		1,568	0,04	86,91
−0,319	11,37	46,11		1,587	4,58	91,49
−0,250	0,01	46,13		1,702	0,01	91,51
−0,160	0,10	46,22		1,761	0,02	91,53
−0,045	0,05	46,27		1,792	0,04	91,57
0,029	9,86	56,13		1,817	0,05	91,62
0,057	0,04	56,17		1,827	0,03	91,64
0,068	0,08	56,25		1,891	4,72	96,37
0,132	0,07	56,32		2,091	0,04	96,41
0,229	0,06	56,39		2,119	0,02	96,43
0,300	0,02	56,41		2,161	0,07	96,50
0,345	7,75	64,15		2,335	0,04	96,54
0,448	0,02	64,17		2,373	3,46	100,00
0,462	0,02	64,20				

Por lo tanto, es necesario distribuir a los alumnos con un WLE igual a uno de los tres percentiles entre los dos cuartiles adyacentes. Por ejemplo, 7,39% de los alumnos obtienen una puntuación igual al percentil 75. Como el 74,07% de los alumnos obtienen una puntuación inferior, es necesario seleccionar una muestra del 0,93% de los alumnos con una puntuación igual al percentil 75 y asignarlos al tercer cuartil. El 6,46% restante se asignará al cuarto cuartil.

Este proceso aleatorio de submuestreo se puede llevar a cabo añadiendo una pequeña cantidad aleatoria al índice. Ese ruido aleatorio generará más categorías y, por tanto, los tres nuevos percentiles podrán dividir la variable del índice en cuartiles que contengan exactamente un 25% de los alumnos. El cuadro 14.2 presenta la sintaxis de SPSS® para la adición de una cantidad aleatoria, así como el cálculo de los percentiles y la recodificación en cuartiles.

Cuadro 14.2. Sintaxis de SPSS® para la recodificación en cuartiles de los índices

```
get file 'C:\PISA\Data2003\int_stui_2003.sav'.
Weight by w_fstuwt.
Select if cnt = 'DEU'.
Set seed = 1.
compute newindex = intmat + rv.normal(0,.01).

FREQ VARIABLES = newindex
  /FORMAT=NOTABLE
  /PERCENTILES = 25 50 75.

if (newindex <  -0.655) quart = 1.
if (newindex >= -0.655 and newindex < 0.0263) quart = 2.
if (newindex >=  0.0263 and newindex < 0.9622) quart = 3.
if (newindex >=  0.9622) quart = 4.

freq quart.
```

Los resultados del procedimiento FREQUENCIES mostrará que cada cuartil contiene un 25% de los casos.

Esta asignación aleatoria de algunas partes de la población a uno de los cuatro cuartiles añade un componente de error al error típico. De hecho, en nuestro ejemplo, la composición del 0,93% de los alumnos asignados al tercer cuartil y la composición del 6,46% restante asignado al cuarto cuartil podría diferir entre dos ejecuciones del procedimiento (a no ser que la semilla se establezca en un entero determinado, como en el cuadro 14.2).

Para tener en cuenta este nuevo componente de error, puede implementarse el enfoque estadístico adoptado para los análisis de valores plausibles. Por lo tanto, consistirá en:

- calcular para cada alumno un conjunto de cinco cuartiles plausibles;
- por cada cuartil plausible, calcular el estadístico buscado y su respectiva varianza muestral usando el peso final y los 80 replicados;
- promediar las cinco estimaciones y sus respectivas varianzas muestrales;
- calcular la varianza de imputación;
- combinar la varianza muestral y la varianza de imputación para obtener la varianza de error final.

Si la variable dependiente es un conjunto de valores plausibles, se empleará el procedimiento descrito en el capítulo 6, excepto que cada valor plausible se analizará con un cuartil plausible distinto. El cuadro 14.3 presenta la sintaxis de SPSS® para el cálculo del rendimiento promedio en matemáticas por cuartil de cualquier índice derivado de un cuestionario.

Cuadro 14.3. Sintaxis de SPSS® para el cálculo del rendimiento en matemáticas por cuartil de un índice derivado de los cuestionarios

```
get file 'C:\PISA\Data2003\int_stui_2003.sav'.
Select if cnt = 'DEU'.
Save outfile = 'c:\pisa\data2003\DEU.sav'.

* IMPORTAR LA MACRO.
Include file 'C:\PISA\macros\mcr_SE_PV_WLEQRT.sps'.

* EJECUTAR LA MACRO.
PVWLEQRT       nrep = 80/
               stat = mean/
               pv = math/
               wle = intmat/
               grp = cnt/
               wgt = w_fstuwt/
               rwgt = w_fstr/
               cons = 0.05/
               infile = 'c:\pisa\data2003\DEU.sav'/.

* COMPROBAR LOS CUARTILES.
get file = 'C:\temp\quarters.sav'.
weight by w_fstuwt.
FREQUENCIES VARIABLES = intmat1 intmat2 intmat3 intmat4 intmat5
   /FORMAT = NOTABLE
   /PERCENTILES = 25 50 75
   /ORDER = ANALYSIS.
MEANS intmat1 by quart1 /CELLS MIN MAX COUNT NPCT.
MEANS intmat2 by quart2 /CELLS MIN MAX COUNT NPCT.
MEANS intmat3 by quart3 /CELLS MIN MAX COUNT NPCT.
MEANS intmat4 by quart4 /CELLS MIN MAX COUNT NPCT.
MEANS intmat5 by quart5 /CELLS MIN MAX COUNT NPCT.
weight off.
```

Los diferentes pasos de este procedimiento son:

1. A partir del índice inicial del cuestionario, se crean cinco variables nuevas añadiendo un número aleatorio.
2. Para cada nueva variable, se calculan los percentiles 25º, 50º y 75º y después se exportan a un archivo de datos temporal.
3. Las cinco nuevas variables se comparan con sus respectivos percentiles y las asignaciones a los cuartiles se guardan en cinco variables categóricas (estas 10 variables nuevas se guardan en el archivo de datos temporal *C:\temp\quarters.sav*).
4. El estadístico se calcula para cada valor plausible del rendimiento en matemáticas mediante una de las cinco variables categóricas nuevas.
5. Se calculan la estimación final y el error típico final.

Los resultados de estos pasos se presentan en la tabla 14.2.

Tabla 14.2. Puntuaciones medias y errores típicos en la escala de matemáticas para cada cuartil del índice de «interés y gusto por las matemáticas»

CNT	cuarto (INTMAT)	estadístico (MATH)	SE
DEU	1	493	4,90
DEU	2	511	4,01
DEU	3	520	4,67
DEU	4	524	4,69

Los conceptos de riesgo relativo y de riesgo atribuible

Riesgo relativo

La noción de riesgo relativo es una medida de asociación entre un factor antecedente y un factor resultado (Cornfield, 1951). El *riesgo relativo* es simplemente la razón de dos riesgos, es decir, el riesgo de observar el resultado cuando el antecedente está presente, y el riesgo de observar el resultado cuando el antecedente no está presente. La tabla 14.3 presenta la notación que se utilizará.

Tabla 14.3. Representaciones utilizadas en una tabla de dos variables

		Medida del resultado		
		Sí	No	Total
Medida del antecedente	Sí	p_{11}	p_{12}	$p_{1.}$
	No	p_{21}	p_{22}	$p_{2.}$
	Total	$p_{.1}$	$p_{.2}$	$p_{..}$

$p_{..}$ es igual a $\frac{n..}{n..}$, donde $n_{..}$ es el número total de alumnos y $p_{..}$ es, por tanto, igual a 1; $p_{i.}$ y $p_{.j}$ representan, respectivamente, las probabilidades marginales para cada fila y cada columna. Las probabilidades marginales son iguales a las frecuencias marginales divididas por el número total de alumnos. Por último, los valores p_{ij} representan las probabilidades para cada celda y son iguales al número de observaciones en una celda particular dividido por el número total de observaciones.

En este documento, las convenciones para la tabla de dos variables serán las siguientes:

- Las filas representan el factor antecedente con:
  - la primera fila para «tener el antecedente»;
  - la segunda fila para «no tener el antecedente».
- Las columnas representan el resultado con:
  - la primera columna para «tener el resultado»;
  - la segunda columna para «no tener el resultado».

En estas condiciones, el riesgo relativo es igual a:

$$RR = \frac{(p_{11} / p_{1.})}{(p_{21} / p_{2.})}.$$

Supongamos que un psicólogo desea analizar el riesgo de que un alumno repita curso si los padres se han divorciado recientemente. El psicólogo selecciona una muestra aleatoria simple de alumnos del 10º curso. En este ejemplo concreto, la variable del resultado está presente si el chico está repitiendo el 10º curso y el factor antecedente se considera presente si los padres del alumno se divorciaron en los dos últimos años. Los resultados obtenidos se presentan en las tablas 14.4 y 14.5.

Tabla 14.4. Distribución de cien alumnos según estado civil de los padres y repetición de curso

	Repetición de curso	No repetición de curso	Total
Padres divorciados	10	10	20
Padres no divorciados	5	75	80
Total	15	85	100

Tabla 14.5. Probabilidades según estado civil de los padres y repetición de curso

	Repetición de curso	No repetición de curso	Total
Padres divorciados	0,10	0,10	0,20
Padres no divorciados	0,05	0,75	0,80
Total	0,15	0,85	1,00

El riesgo relativo, por tanto, es igual a:

$$RR = \frac{(p_{11} / p_{1.})}{(p_{21} / p_{2.})} = \frac{(0{,}10 / 0{,}20)}{(0{,}05 / 0{,}80)} = \frac{0{,}5}{0{,}0625} = 8$$

Esto significa que la probabilidad de repetir el 10º curso es ocho veces mayor si los padres se han divorciado recientemente que si no es así.

Riesgo atribuible

El riesgo atribuible es igual a:

$$RA = \frac{(p_{11}p_{22}) - (p_{12}p_{21})}{(p_{.1}p_{2.})}.$$

En el ejemplo anterior, el riesgo atribuible es igual a:

$$RA = \frac{(p_{11}p_{22}) - (p_{12}p_{21})}{(p_{.1}p_{2.})} = \frac{(0{,}10 \cdot 0{,}75) - (0{,}10 \cdot 0{,}05)}{(0{,}15 \cdot 0{,}80)} = 0{,}583.$$

El riesgo atribuible se interpreta como se indica a continuación. Si el factor de riesgo pudiera eliminarse, la tasa de ocurrencia de la variable resultado en la población se reduciría por este coeficiente. Con la siguiente versión de la fórmula, el significado del riesgo atribuible, es decir, una reducción del resultado si desaparece el factor de riesgo, es más obvio.

$$RA = \frac{(p_{.1}) - (p_{21} / p_{2.})}{(p_{.1})}$$

La expresión $p_{.1}$ representa la proporción de alumnos en la muestra entera que presentan el resultado. La expresión *(p21/p2.)* representa la proporción de alumnos que no corren riesgo (no tienen el antecedente), pero de todos modos padecen el resultado. La diferencia entre estas dos proporciones aporta la reducción absoluta si el riesgo (o antecedente) se eliminase. Dividir esta diferencia por la primera expresión transforma esta reducción absoluta en una reducción relativa o una reducción expresada como una proporción.

Estas dos fórmulas proporcionan el mismo coeficiente:

$$RA = \frac{(p_{.1}) - (p_{21} / p_{2.})}{(p_{.1})} = \frac{(0{,}15) - (0{,}05 / 0{,}80)}{0{,}15} = 0{,}583$$

Para expresar este coeficiente como un porcentaje, el coeficiente necesita ser multiplicado por 100.

Inestabilidad del riesgo relativo y del riesgo atribuible

El riesgo relativo y el riesgo atribuible se inventaron para variables dicotómicas. Pero cada vez con más frecuencia, estos dos coeficientes se extienden y utilizan con variables continuas. Para aplicar los coeficientes a variables continuas, es necesario establecer un punto de corte para cada variable y dicotomizando así las variables continuas.

Es importante reconocer que, cuando se aplican a variables dicotomizadas, los valores calculados de riesgo relativo y de riesgo atribuible dependerán del valor del punto de corte escogido.

Para demostrar la influencia del punto de corte sobre los riesgos relativo y atribuible, se generaron dos variables aleatorias con una correlación de 0,30. Estas dos variables se transformaron a continuación en variables dicotómicas, usando como puntos de corte los percentiles 10º, 15º, 20º , 25º y 30º. La tabla 14.6 presenta los riesgos relativo y atribuible para estos puntos de corte.

Tabla 14.6. Riesgo relativo y riesgo atribuible para distintos puntos de corte

Percentil	Riesgo relativo	Riesgo atribuible
10	2,64	0,13
15	2,32	0,16
20	1,90	0,15
25	1,73	0,15
30	1,63	0,15

La tabla 14.6 muestra que los coeficientes de riesgo relativo y, en menor grado, de riesgo atribuible dependen del establecimiento de los puntos de corte y, por tanto, es preciso interpretar su valor a la luz de esta observación.

Una comparación semejantre de los riesgos relativo y atribuible se calculó con los datos de PISA, para identificar los cambios según la situación de los puntos de corte. El factor antecedente era el nivel de estudios de la madre y la variable de resultado era la puntuación del alumno en lectura. Una puntuación baja en lectura («tener el resultado») se definió sucesivamente dentro de los países como tener puntuaciones inferiores a los percentiles 10º, 15º, 20º, 25º, 30º y 35º.

Los riesgos relativos para estos distintos puntos de corte son (para los países de la OCDE) respectivamente iguales a 2,20, 1,92, 1,75, 1,62, 1,53 y 1,46. Los riesgos atribuibles son respectivamente iguales a 0,25, 0,21, 0,19, 0,17, 0,15 y 0,14.

Sin embargo, las correlaciones entre los distintos riesgos relativos y atribuibles son bastante altas, como se muestra en la tabla 14.7.

Tabla 14.7. Correlación entre riesgos relativos y riesgos atribuibles en el percentil 10º con los percentiles 15º, 20º, 25º, 30º y 35º

	RR	RA
P15	0,96	0,98
P20	0,93	0,97
P25	0,92	0,96
P30	0,90	0,94
P35	0,87	0,92

En PISA, se decidió utilizar el percentil 25º como punto de corte para variables continuas al calcular los riesgos relativos y atribuibles.

Cálculo del riesgo relativo y del riesgo atribuible

Según las variables implicadas en el cálculo del riesgo relativo y el riesgo atribuible, el procedimiento podría variar. De hecho, estos dos conceptos estadísticos necesitan como datos dos variables dicotómicas, como el género (ST03Q01).

Sin embargo, la mayoría de las variables en las bases de datos de PISA no son dicotómicas, sino categóricas o continuas.

La recodificación de una variable categórica en dicotómica no plantea problemas especiales. Desde un punto de vista teórico, es preciso decidir el propósito de la comparación y, como resultado, la recodificación deseada. Por ejemplo, en PISA 2000, los niveles de estudios de los padres se presentan mediante la clasificación CINE (Clasificación internacional de niveles educativos; en inglés ISCED, *International Standard Classification of Education*) (OCDE, 1999*b*). Si el contraste se establece sobre la distinción entre estudios terciarios frente a estudios no terciarios, entonces la variable categórica puede recodificarse en una variable dicotómica. Los alumnos cuyos padres no tengan un título de estudios terciarios se considerarán como de riesgo.

Las variables numéricas también deben recodificarse en variables dicotómicas. Como ya se dijo anteriormente, la OCDE ha decidido dividir las variables numéricas tomando como punto de corte el percentil 25º.

En las bases de datos de PISA 2000 y PISA 2003, todas las variables numéricas, excepto las escalas de rendimiento, son variables discontinuas. Para garantizar que el percentil 25º divida las variables en dos categorías que incluyan, respectivamente, el 25% y el 75%, es necesario añadir una cantidad aleatoria a la variable inicial, como se describió en la sección dedicada a los análisis en cuartiles. Se calculan cinco estimaciones de riesgo relativo o cinco de riesgo atribuible y después se combinan.

Por último, si hay implicados valores plausibles, como medidas de resultados, también se calcularán cinco estimaciones y se combinarán después. Sin embargo, no es necesario añadir una cantidad aleatoria a la variable inicial, ya que constituyen una variable continua.

Conclusiones

Este capítulo se ha dedicado a algunas cuestiones estadísticas relacionadas con la forma en que la OCDE presentó los resultados de PISA 2000 y PISA 2003 en los informes iniciales, sobre todo los índices de cuestionario en cuartiles y los riesgos relativos y atribuibles.

Capítulo 15

Las macros de SPSS®

Introducción

Este capítulo presenta la sintaxis SPSS® de las macros utilizadas en los capítulos anteriores. Estas macros también están disponibles en un archivo comprimido que se puede descargar del mismo sitio *web* donde se descarga este manual.

Se describen doce macros, sintetizadas en la tabla 15.1. Los nombres de archivo están en negrita y los nombres de las macros y sus argumentos en un tipo de letra monoespaciado. Todas las macros tienen cinco argumentos comunes:

- `nrep =`
- `wgt =`
- `rwgt =`
- `cons =`
- `infile =`

Los restantes argumentos son propios de cada macro en particular. Estos argumentos propios fueron ampliamente explicados en los capítulos precedentes.

Estructura de las macros

Todas las macros tienen la misma estructura.

- El primer paso consiste en:
  - Leer los datos contenidos en el archivo especificado en INFILE y descartar las variables que no son necesarias para el análisis.
- El segundo paso constituye la parte iterativa de la macro:
  - El procedimiento SPSS® que calcula el estimador se ejecuta 81 o 405 veces,
  - En cada ejecución, los resultados se graban en un archivo temporal. Los resultados parciales del replicado y de los valores plausibles pueden ser también combinados en un archivo.
- El paso final se dedica al cálculo del estadístico final y de su error típico, es decir:
  - Se calculan las diferencias al cuadrado entre el estimador final y los 80 replicados,
  - Se calcula la suma de las diferencias al cuadrado y se divide por 20,
  - Se calculan los estimadores finales, los estimadores de varianza muestral y, en el caso de los valores plausibles, la variaza de imputación o de medida.

La sintaxis SPSS® se presenta a continuación.

Tabla 15.1. Síntesis de las doce macros de SPSS®

Estadístico deseado	Sin valores plausibles	Con valores plausibles

Mean, sd, sum, pgt, plt, pin, pout, fgt, flt, fin, fout	**mcr_SE_univ.sps** (cap. 6)
	`univar nrep = /`
	`stat = /`
	`dep = /`
	`grp = /`
	`wgt = /`
	`rwgt = /`
	`cons = /`
	`infile = ''/.`
		mcr_SE_pv.sps (cap. 7)
		`PV nrep = /`
		`stat = /`
		`dep = /`
		`grp = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`

Porcentaje	**mcr_SE_GrpPct.sps** (cap. 6)
	`GRPPCT nrep = /`
	`within = /`
	`grp = /`
	`wgt = /`
	`rwgt = /`
	`cons = /`
	`infile = ''/.`
		mcr_SE_PctLev.sps (cap. 8)
		`PCTLEV nrep = /`
		`within = /`
		`grp = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`

Coeficientes de regresión	**mcr_SE_reg.sps** (cap. 6)
	`REGnoPV nrep = /`
	`ind = /`
	`dep = /`
	`grp = /`
	`wgt = /`
	`rwgt = /`
	`cons = /`
	`infile = ''/.`
		mcr_SE_reg_PV.sps (cap. 7)
		`REG_PV nrep = /`
		`ind = /`
		`dep = /`
		`grp = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`

Coeficientes de correlación	**mcr_SE_cor.sps** (cap. 6)
	`CORnoPV nrep = /`
	`var1 = /`
	`var2 = /`
	`grp = /`
	`wgt = /`
	`rwgt = /`
	`cons = /`
	`infile = ''/.`
		mcr_SE_cor_1PV.sps (cap. 7)
		`COR_1PV nrep = /`
		`nopv = /`
		`pv = /`
		`grp = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`
		
		mcr_SE_cor_2PV.sps (cap. 7)
		`COR_2PV nrep = /`
		`pv1 = /`
		`pv2 = /`
		`grp = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`

Diferencias en:
mean, sd, sum, pgt, plt, pin, pout, fgt, flt, fin, fout
	mcr_SE_dif.sps (cap. 10)
	`difNOpv nrep = /`
	`dep = /`
	`stat = /`
	`within = /`
	`compare = /`
	`categ = /`
	`wgt = /`
	`rwgt = /`
	`cons = /`
	`infile = ''/.`
		mcr_SE_dif_PV.sps (cap. 10)
		`dif_pv nrep = /`
		`dep = /`
		`stat = /`
		`within = /`
		`compare = /`
		`categ = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`

Mean, sd, sum, pgt, plt, pin, pout, fgt, flt, fin, fout, en valores plausibles
 	
índices de cuartiles de WLE
		mcr_SE_PV_WLEQRT.sps (cap. 14)
		`PVWLEQRT nrep = /`
		`stat = /`
		`pv = /`
		`wle = /`
		`grp = /`
		`wgt = /`
		`rwgt = /`
		`cons = /`
		`infile = ''/.`

Cuadro 15.1. Sintaxis de la macro mcr_SE_univ.sps

```
define univar (nrep   = !charend('/')/
               stat   = !charend('/')/
               dep    = !charend('/')/
               grp    = !charend('/')/
               wgt    = !charend('/')/
               rwgt   = !charend('/')/
               cons   = !charend('/')/
               infile = !charend('/')).

get file !infile /keep !grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!dep).

*** COMPUTE ESTIMATE ***.

weight by !wgt.
aggregate outfile = !quote(!concat('c:\temp\',!stat,'_all.sav'))
   /break=!grp /stat=!stat(!dep).

* REPLICATES.

!do !i= 1 !to !nrep.
weight by !concat(!rwgt,!i).
aggregate outfile = !quote(!concat('C:\temp\',!stat,'_',!dep,!i,'.sav'))
   /break=!grp /statr=!stat(!dep).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!stat,'_',!dep,'1.sav')).

!Do !e = 2 !to !nrep.
add files file=* /file=!quote(!concat('C:\temp\',!stat,'_',!dep,!e,'.sav')).
!Doend.

sort cases by !grp.
match files file=* /table=!quote(!concat('c:\temp\',!stat,'_all.sav')) /by
!grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

compute var=(statr-stat)**2.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/ break=!grp/ stat= mean(stat)/ var=sum(var).

compute var=!cons*var.

*** COMPUTE STANDARD ERROR ***.

compute se=sqrt(var).
exec.

formats stat (f8.3)/ se (f8.3).
list cases/var= !grp stat se.

!enddefine.
```

Cuadro 15.2. Sintaxis de la macro mcr_SE_pv.sps

```
define PV  (nrep = !charend('/')/
            stat = !charend('/')/
            dep = !charend('/') /
            grp = !charend('/') /
            wgt = !charend('/') /
            rwgt = !charend('/') /
            cons = !charend('/')/
            infile = !charend('/')).

get file !infile /keep !grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!concat(pv,1,!dep)
  !concat(pv,2,!dep) !concat(pv,3,!dep) !concat(pv,4,!dep) !concat(pv,5,!dep)
.

*** COMPUTE STATISTIC ***.

weight by !wgt.
erase file='c:\temp\all.sav'.
aggregate outfile = 'c:\temp\all.sav' /break=!grp /
    stat1 stat2 stat3 stat4 stat5=!stat(!concat(pv1,!dep) !concat(pv2,!dep)
!concat(pv3,!dep) !concat(pv4,!dep) !concat(pv5,!dep)).

* REPLICATES.

!do !i= 1 !to !nrep.
weight by !concat(!rwgt,!i).
erase file=!quote(!concat('c:\temp\',!dep,!i,'.sav')).
aggregate outfile = !quote(!concat('c:\temp\',!dep,!i,'.sav'))/break=!grp /
    statr1 statr2 statr3 statr4 statr5=!stat(!concat(pv1,!dep) !con-
cat(pv2,!dep) !concat(pv3,!dep) !concat(pv4,!dep) !concat(pv5,!dep)).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('c:\temp\',!dep,'1','.sav')).
cache.

!Do !e = 2 !to !nrep.
add files/file=*/file=!quote(!concat('c:\temp\',!dep,!e,'.sav')).
!Doend.

sort cases by !grp.

match files file=*/table= 'c:\temp\all.sav'/by !grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
          b=stat1 to stat5/
          c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
         break=!grp/
          stat1 to stat5= mean(stat1 to stat5)/
          var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.
```

```
compute pv_var=mean(var1 to var5).

*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.

compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).

formats stat (f8.2)/ SE (f8.2).
list cases/var= !grp stat SE.

!enddefine.
```

Cuadro 15.3. Sintaxis de la macro mcr_SE_GrpPct.sps

```
define GRPPCT (nrep = !charend('/')/
               within = !charend('/') /
               grp = !charend('/') /
               wgt = !charend('/') /
               rwgt = !charend('/') /
               cons = !charend('/')/
               infile = !charend('/')).

get file !infile /keep !within !grp !wgt !concat(!rwgt,1) to !con-
cat(!rwgt,!nrep)).

*** COMPUTE ESTIMATE ***.

weight by !wgt.
aggregate outfile='c:\temp\temp1.sav' /break=!within /N_all=n.
aggregate outfile=* /break=!within !grp /N_grp=n.
exe.
match files file=* /table='c:\temp\temp1.sav' /by !within.
compute stat=100*(n_grp/n_all).
save outfile=!quote(!concat('c:\temp\',!grp,'.sav')).

* REPLICATES.

!do !i= 1 !to !nrep.
get file !infile /keep !within !grp !wgt !concat(!rwgt,1) to !con-
cat(!rwgt,!nrep)).
weight by !concat(!rwgt,!i).
aggregate outfile='c:\temp\temp2.sav' /break=!within /N_all=n.
aggregate outfile=* /break=!within !grp /N_grp=n.
exe.
match files file=* /table='c:\temp\temp2.sav' /by !within.
compute statR=100*(n_grp/n_all).
save outfile=!quote(!concat('c:\temp\',!grp,'_',!i,'.sav')).
erase file='c:\temp\temp2.sav'.
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('c:\temp\',!grp,'_1.sav')).

!Do !e = 2 !to !nrep.
add files file=* /file=!quote(!concat('c:\temp\',!grp,'_',!e,'.sav')).
!Doend.

sort cases by !within !grp.

match files file=* /table=!quote(!concat('c:\temp\',!grp,'.sav')) /by !within
!grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

compute var=(statr-stat)**2.
exec.

save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
          break=!within !grp/
          stat= mean(stat)/
          var=sum(var).

compute var=!cons*var.
```

```
*** COMPUTE STANDARD ERROR ***.

compute SE=sqrt(var).
exec.

formats stat (f8.3)/ SE (f8.3).
list cases/var= !within !grp stat SE.

!enddefine.
```

Cuadro 15.4. Sintaxis de la macro mcr_SE_PctLev.sps

```
define PCTLEV (nrep = !charend('/')/
               within = !charend('/') /
               grp = !charend('/') /
               wgt = !charend('/') /
               rwgt = !charend('/') /
               cons = !charend('/')/
               infile = !charend('/')).

*** COMPUTE ESTIMATE ***.

get file !infile /keep !within !concat(!grp,'1') !concat(!grp,'2')
!concat(!grp,'3') !concat(!grp,'4')
   !concat(!grp,'5') !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)).
weight by !wgt.
aggregate outfile='c:\temp\temp10.sav' /break=!within /N_all=n.

!do !j=1 !to 5.
get file !infile /keep !within !concat(!grp,'1') !concat(!grp,'2')
!concat(!grp,'3') !concat(!grp,'4')
   !concat(!grp,'5') !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)).
weight by !wgt.
aggregate outfile=* /break=!within !concat(!grp,!j) /N_grp=n.
exe.
match files file=* /table='c:\temp\temp10.sav' /by !within.
compute !concat('stat',!j)=100*(n_grp/n_all).
rename var (!concat(!grp,!j)=!grp).
save outfile=!quote(!concat('c:\temp\temp',!j,'.sav')) /keep=!within !grp
!concat('stat',!j).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !within !grp.
save outfile='c:\temp\all.sav'.

* REPLICATES.

!do !i= 1 !to !nrep.
get file !infile /keep !within !concat(!grp,'1') !concat(!grp,'2')
!concat(!grp,'3') !concat(!grp,'4')
   !concat(!grp,'5') !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)).
weight by !concat(!rwgt,!i).
erase file='c:\temp\temp20.sav'.
aggregate outfile='c:\temp\temp20.sav' /break=!within /N_all=n.

!do !j=1 !to 5.
get file !infile /keep !within !concat(!grp,'1') !concat(!grp,'2')
!concat(!grp,'3') !concat(!grp,'4')
   !concat(!grp,'5') !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)).
weight by !concat(!rwgt,!i).

aggregate outfile=* /break=!within !concat(!grp,!j) /N_grp=n.
exe.
match files file=* /table='c:\temp\temp20.sav' /by !within.
compute !concat('statR',!j)=100*(n_grp/n_all).
rename var (!concat(!grp,!j)=!grp).
save outfile=!quote(!concat('c:\temp\temp',!j,'.sav')) /keep=!within !grp
!concat('statR',!j).
!doend.
```

```
match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !within !grp.
save outfile=!quote(!concat('c:\temp\',!grp,'_',!i,'.sav')).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('c:\temp\',!grp,'_1.sav')).

!Do !e = 2 !to !nrep.
add files file=* /file=!quote(!concat('c:\temp\',!grp,'_',!e,'.sav')).
!Doend.

sort cases by !within !grp.

match files file=* /table='c:\temp\all.sav' /by !within !grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
         b=stat1 to stat5/
         c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
         break=!within !grp/
           stat1 to stat5= mean(stat1 to stat5)/
           var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.

compute pv_var=mean(var1 to var5).

*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.

compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).

formats stat (f8.2)/ SE (f8.2).
list cases/var= !within !grp stat SE.

!enddefine.
```

Cuadro 15.5. Sintaxis de la macro mcr_SE_reg.sps

```
define REGnoPV (nrep = !charend('/')/
                ind = !charend('/') /
                dep = !charend('/') /
                grp = !charend('/') /
                wgt = !charend('/') /
                rwgt = !charend('/') /
                cons = !charend('/')/
                infile = !charend('/')).

get file !infile /keep !grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep) !dep
!ind).

*** COMPUTE ESTIMATE ***.

*sort cases by !grp.
split file by !grp.

weight by !wgt.

REGRESSION
  /DEPENDENT !dep
  /METHOD=ENTER !ind
  /OUTFILE=COVB('C:\temp\coef.sav') .
get file='C:\temp\coef.sav'.
select if (rowtype_='EST').
rename var(CONST_=b0).
VARSTOCASES  /MAKE stat FROM b0 !ind
 /INDEX = ind(stat) /KEEP = !grp /NULL = KEEP.

erase file='c:\temp\all.sav'.
sort cases by !grp ind.
save outfile='c:\temp\all.sav'.

* REPLICATES.

!do !i=1 !to !nrep.
get file !infile /keep !grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep) !dep
!ind).

*sort cases by !grp.
split file by !grp.

weight by !concat(!rwgt,!i).

REGRESSION
  /DEPENDENT !dep
  /METHOD=ENTER !ind
  /OUTFILE=COVB('C:\temp\coef.sav') .
get file='C:\temp\coef.sav'.
select if (rowtype_='EST').
rename var(CONST_=b0).
VARSTOCASES  /MAKE statR FROM b0 !ind
 /INDEX = ind(statR) /KEEP = !grp /NULL = KEEP.

save outfile=!quote(!concat('C:\temp\',!dep,!i,'.sav'))

!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!dep,'1.sav')).

!Do !e = 2 !to !nrep.
```

```
add files file=* /file=!quote(!concat('C:\temp\',!dep,!e,'.sav')).
!Doend.

sort cases by !grp ind.

match files file=* /table='c:\temp\all.sav' /by !grp ind.
exec.


*** COMPUTE SAMPLING VARIANCE (U) ***.

compute var=(statr-stat)**2.
exec.

save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
          break=!grp IND/
          stat= mean(stat)/
          var=sum(var).

compute var=!cons*var.

*** COMPUTE STANDARD ERROR ***.

compute se=sqrt(var).
exec.

formats stat (f8.3)/ se (f8.3).
list cases/var= !grp ind stat se.

!enddefine.
```

Cuadro 15.6. Sintaxis de la macro mcr_SE_reg_PV.sps

```
define REG_PV (nrep = !charend('/')/
               ind = !charend('/') /
               dep = !charend('/') /
               grp = !charend('/')  /
               wgt   = !charend('/') /
               rwgt   = !charend('/') /
               cons = !charend('/')/
               infile = !charend('/')).

*** COMPUTE REGRESSION COEFFICIENTS ***.

!do !j=1 !to 5.

get file !infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep) !IND
!concat('pv1',!dep)
  !concat('pv2',!dep) !concat('pv3',!dep) !concat('pv4',!dep) !con-
cat('pv',5,!dep) .

*sort cases by !grp.
split file by !grp.

weight by !wgt.

REGRESSION
  /DEPENDENT !concat('pv',!j,!dep)
  /METHOD=ENTER !ind
  /OUTFILE=COVB(!quote(!concat('C:\temp\coef',!j,'.sav'))) .
get file=!quote(!concat('C:\temp\coef',!j,'.sav')).

select if (rowtype_='EST').
rename var(CONST_=b0).
VARSTOCASES  /MAKE !concat('stat',!j) FROM b0 !ind
 /INDEX = ind(!concat('stat',!j)) /KEEP =  !grp /NULL = KEEP.
sort cases by !grp IND.
save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !GRP ind.
erase file='c:\temp\all.sav'.
save outfile='c:\temp\all.sav'.

* REPLICATES.

!do !i=1 !to !nrep.
!do !j=1 !to 5.

get file !infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep) !IND
!concat('pv1',!dep)
  !concat('pv2',!dep) !concat('pv3',!dep) !concat('pv4',!dep) !con-
cat('pv',5,!dep) .

*sort cases by !grp.
split file by !grp.

weight by !CONCAT(!rwgt,!i).

REGRESSION
  /DEPENDENT !concat('pv',!j,!dep)
```

```
  /METHOD=ENTER !ind
  /OUTFILE=COVB(!quote(!concat('C:\temp\coef',!j,'.sav'))) .
get file=!quote(!concat('C:\temp\coef',!j,'.sav')).
select if (rowtype_='EST').
rename var(CONST_=b0).
VARSTOCASES  /MAKE !concat('statR',!j) FROM b0 !ind
 /INDEX = ind(!concat('statR',!j)) /KEEP =  !grp /NULL = KEEP.
sort cases by !grp IND.
save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !GRP ind.

erase file=!quote(!concat('C:\temp\',!dep,!i,'.sav')).
save outfile=!quote(!concat('C:\temp\',!dep,!i,'.sav')).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!dep,'1','.sav')).
cache.

!Do !e = 2 !to !nrep.
add files/file=*/file=!quote(!concat('C:\temp\',!dep,!e,'.sav')).
!Doend.

SORT CASES BY !grp IND.
match files file=*/table= 'c:\temp\all.sav'/by !grp IND.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
          b=stat1 to stat5/
          c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
          break=!grp IND/
          stat1 to stat5= mean(stat1 to stat5)/
          var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.

compute pv_var=mean(var1 to var5).


*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.

compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).
```

```
*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).

formats stat (f8.2)/ SE (f8.2).
list cases/var= !grp IND stat SE.

!enddefine.
```

Cuadro 15.7. Sintaxis de la macro mcr_SE_cor.sps

```
define CORnoPV (nrep = !charend('/')/
                var1 = !charend('/') /
                var2 = !charend('/') /
                grp = !charend('/')  /
                wgt = !charend('/') /
                rwgt = !charend('/') /
                cons = !charend('/')/
                infile = !charend('/')).

get file !infile /keep !grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!var2 !var1.

*** COMPUTE ESTIMATE ***.

*sort cases by !grp.
split file by !grp.

weight by !wgt.
descr !var1 !var2 / stat=mean stddev /save.
compute Y=!concat('Z',!var1)*!concat('Z',!var2).
split file off.
aggregate outfile=* /break=!grp /sumY=sum(Y) /n=n.
compute stat=sumY/n.

erase file='c:\temp\all.sav'.
save outfile='c:\temp\all.sav' /keep=cnt stat.

* REPLICATES.

!do !i=1 !to !nrep.
get file !infile /keep !grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!var2 !var1).

*sort cases by !grp.
split file by !grp.

weight by !concat(!rwgt,!i).
descr !var1 !var2/ stat=mean stddev /save.
compute Y=!concat('Z',!var1)*!concat('Z',!var2).
split file off.
aggregate outfile=* /break=!grp /sumY=sum(Y) /n=n.
compute statR=sumY/n.


erase file=!quote(!concat('C:\temp\',!var2,!i,'.sav')).
save outfile=!quote(!concat('C:\temp\',!var2,!i,'.sav')) /keep=cnt statR.
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!var2,'1.sav')).

!Do !e = 2 !to !nrep.
add files file=* /file=!quote(!concat('C:\temp\',!var2,!e,'.sav')).
!Doend.

sort cases by !grp.
match files file=* /table='c:\temp\all.sav' /by !grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

compute var=(statr-stat)**2.
```

```
exec.

save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
         break=!grp/
           stat= mean(stat)/
           var=sum(var).

compute var=!cons*var.

*** COMPUTE STANDARD ERROR ***.

compute SE=sqrt(var).
exec.

formats stat (f8.3)/ SE (f8.3).
list cases/var= !grp stat SE.

!enddefine.
```

Cuadro 15.8. Sintaxis de la macro mcr_SE_cor_1PV.sps

```
define COR_1PV (nrep = !charend('/')/
                nopv = !charend('/') /
                pv = !charend('/') /
                grp = !charend('/') /
                wgt = !charend('/') /
                rwgt = !charend('/') /
                cons = !charend('/')/
                infile = !charend('/')).

*** COMPUTE REGRESSION COEFFICIENTS ***.

!do !j=1 !to 5.

get file !infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!concat('pv1',!pv)
  !concat('pv2',!pv) !concat('pv3',!pv) !concat('pv4',!pv) !concat('pv',5,!pv)
!nopv.

*sort cases by !grp.
split file by !grp.

weight by !wgt.
split file layered by !grp.

!let !x=!nopv.
!let !y=!concat('pv',!j,!pv).

descr !x !y /stat=mean stddev /save.
compute P=!concat('Z',!x)*!concat('Z',!y).
split file off.

aggregate outfile=* /break=!grp /sumP=sum(P) /n=n.
compute !concat('stat',!j)=sumP/n.

save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')) /keep=cnt !con-
cat('stat',!j).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !GRP.
erase file='c:\temp\all.sav'.
save outfile='c:\temp\all.sav'.

* REPLICATES.

!do !i=1 !to !nrep.
!do !j=1 !to 5.

get file !infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!concat('pv1',!pv)
  !concat('pv2',!pv) !concat('pv3',!pv) !concat('pv4',!pv) !concat('pv',5,!pv)
!nopv.

*sort cases by !grp.
split file by !grp.

weight by !concat(!rwgt,!i).
split file layered by !grp.
```

```
!let !x=!nopv.
!let !y=!concat('pv',!j,!pv).

descr !x !y /stat=mean stddev /save.
compute P=!concat('Z',!x)*!concat('Z',!y).
split file off.

aggregate outfile=* /break=!grp /sumP=sum(P) /n=n.
compute !concat('statR',!j)=sumP/n.

save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')) /keep=cnt !con-
cat('statR',!j).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !GRP.

erase file=!quote(!concat('C:\temp\',!pv,!i,'.sav')).
save outfile=!quote(!concat('C:\temp\',!pv,!i,'.sav')).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!pv,'1','.sav')).
cache.

!Do !e = 2 !to !nrep.
add files/file=*/file=!quote(!concat('C:\temp\',!pv,!e,'.sav')).
!Doend.

sort cases by !grp.
match files file=*/table= 'c:\temp\all.sav'/by !grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
          b=stat1 to stat5/
          c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
          break=!grp/
          stat1 to stat5= mean(stat1 to stat5)/
          var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.

compute pv_var=mean(var1 to var5).

*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.

compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.
```

```
compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).

formats stat (f8.3)/ SE (f8.1).
list cases/var= !grp stat SE.

!enddefine.
```

Cuadro 15.9. Sintaxis de la macro mcr_SE_cor_2PV.sps

```
define COR_2PV (nrep = !charend('/')/
               pv1 = !charend('/') /
               pv2 = !charend('/') /
               grp = !charend('/') /
               wgt = !charend('/') /
               rwgt = !charend('/') /
               cons = !charend('/')/
               infile = !charend('/')).

*** COMPUTE REGRESSION COEFFICIENTS ***.

!do !j=1 !to 5.

get file !infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!concat('pv1',!pv2)
  !concat('pv2',!pv2) !concat('pv3',!pv2) !concat('pv4',!pv2) !con-
cat('pv',5,!pv2)
  !concat('pv1',!pv1) !concat('pv2',!pv1) !concat('pv3',!pv1) !con-
cat('pv4',!pv1) !concat('pv',5,!pv1).

*sort cases by !grp.
split file by !grp.

weight by !wgt.
split file layered by !grp.

!let !x=!concat('pv',!j,!pv1).
!let !y=!concat('pv',!j,!pv2).

descr !x !y /stat=mean stddev /save.
compute P=!concat('Z',!x)*!concat('Z',!y).
split file off.

aggregate outfile=* /break=!grp /sumP=sum(P) /n=n.
compute !concat('stat',!j)=sumP/n.

save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')) /keep=cnt !con-
cat('stat',!j).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'

           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !GRP.
erase file='c:\temp\all.sav'.
save outfile='c:\temp\all.sav'.

* REPLICATES.

!do !i=1 !to !nrep.
!do !j=1 !to 5.

get file !infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!concat('pv1',!pv2)
  !concat('pv2',!pv2) !concat('pv3',!pv2) !concat('pv4',!pv2) !con-
cat('pv',5,!pv2)
  !concat('pv1',!pv1) !concat('pv2',!pv1) !concat('pv3',!pv1) !con-
cat('pv4',!pv1) !concat('pv',5,!pv1).

*sort cases by !grp.
```

```
split file layered by !grp.

weight by !CONCAT(!rwgt,!i).
split file by !grp.

!let !x=!concat('pv',!j,!pv1).
!let !y=!concat('pv',!j,!pv2).

descr !x !y /stat=mean stddev /save.
compute P=!concat('Z',!x)*!concat('Z',!y).
split file off.

aggregate outfile=* /break=!grp /sumP=sum(P) /n=n.
compute !concat('statR',!j)=sumP/n.

save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')) /keep=cnt !con-
cat('statR',!j).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !GRP.

erase file=!quote(!concat('C:\temp\',!pv2,!i,'.sav')).
save outfile=!quote(!concat('C:\temp\',!pv2,!i,'.sav')).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!pv2,'1','.sav')).
cache.

!Do !e = 2 !to !nrep.
add files/file=*/file=!quote(!concat('C:\temp\',!pv2,!e,'.sav')).
!Doend.

sort cases by !grp.
match files file=*/table= 'c:\temp\all.sav'/by !grp.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
         b=stat1 to stat5/
         c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
         break=!grp/
           stat1 to stat5= mean(stat1 to stat5)/
           var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.

compute pv_var=mean(var1 to var5).

*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.
```

```
compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).

formats stat (f8.3)/ SE (f8.1).
list cases/var= !grp stat SE.

!enddefine.
```

Cuadro 15.10. Sintaxis de la macro mcr_SE_dif.sps

```
define difNOpv (nrep = !charend('/')/
                dep = !charend('/') /
                stat = !charend ('/') /
                within = !charend('/') /
                compare = !charend('/') /
                categ = !charend ('/') /
                wgt = !charend('/') /
                rwgt = !charend('/') /
                cons = !charend('/')/
                infile = !charend('/')).

get file !infile /keep !dep !within !compare !wgt
!concat(!rwgt,1) to !concat(!rwgt,!nrep) .

*** COMPUTE ESTIMATE ***.

weight by !wgt.
aggregate outfile = *
   /break=!within !compare /stat=!stat(!dep).
exe.
casestovars /id=!within /index=!compare .

!let !n=!length(!categ).
!let !m=!LENGTH(!substr(!blanks(!n),2)).

!let !var=" ".
!do !a=1 !to !m.
!let !s=!LENGTH(!concat(!blanks(!a),' ')).
!do !b=!s !to !n.

!let !d=!substr(!categ,!a,1).
!let !e=!substr(!categ,!b,1).
compute !concat('stat',!d,'_',!e)=
        !concat('stat.',!d)-!concat('stat.',!e).
!let !var=!concat(!var,'stat',!d,'_',!e,' ').
!doend.
!doend.

VARSTOCASES  /MAKE stat FROM !var
 /INDEX = contrast(stat) /KEEP = !within /NULL = DROP.

save outile='c:\temp\all.sav'.

* REPLICATES.

!do !i= 1 !to !nrep.
get file !infile /keep !within !compare !wgt !concat(!rwgt,1) to !con-
cat(!rwgt,!nrep) !dep).
weight by !concat(!rwgt,!i).

aggregate outfile = *
   /break=!within !compare /stat=!stat(!dep).
exe.
casestovars /id=!within /index=!compare .

!let !n=!length(!categ).
!let !m=!LENGTH(!substr(!blanks(!n),2)).

!let !var=" ".
!do !a=1 !to !m.
!let !s=!LENGTH(!concat(!blanks(!a),' ')).
!do !b=!s !to !n.
```

```
!let !d=!substr(!categ,!a,1).
!let !e=!substr(!categ,!b,1).
compute !concat('stat',!d,'_',!e)=
        !concat('stat.',!d)-!concat('stat.',!e).
!let !var=!concat(!var,'stat',!d,'_',!e,' ').
!doend.
!doend.

VARSTOCASES  /MAKE statR FROM !var
 /INDEX = contrast(statR) /KEEP = !within /NULL = DROP.

save outile=!quote(!concat('c:\temp\',!dep,!i,'.sav')).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('c:\temp\',!dep,'1','.sav')).

!Do !e = 2 !to !nrep.
add files file=* /file=!quote(!concat('c:\temp\',!dep,!e,'.sav')).
!Doend.

sort cases by !within CONTRAST.
match files file=* /table='c:\temp\all.sav' /by !within CONTRAST.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

compute var=(statr-stat)**2.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/ break=!within contrast/ stat= mean(stat)/ var=sum(var).

compute var=!cons*var.

*** COMPUTE STANDARD ERROR ***.

compute se=sqrt(var).
exec.

formats stat (f8.3)/ se (f8.3).
list cases/var= !within contrast stat se.

string contr (a3).
compute contr=substr(contrast,5,3).
casestovars /id=!within /index=contr /groupby=index /drop=contrast var.

!enddefine.
```

Cuadro 15.11. Sintaxis de la macro mcr_SE_dif_PV.sps

```
define dif_pv (nrep = !charend('/')/
               dep = !charend('/') /
               stat = !charend ('/') /
               within = !charend('/') /
               compare = !charend('/') /
               categ = !charend ('/') /
               wgt = !charend('/') /
               rwgt = !charend('/') /
               cons = !charend('/')/
               infile = !charend('/')).

*** COMPUTE ESTIMATE ***.

!do !j=1 !to 5.
get file !infile /keep !concat('pv1',!dep)
  !concat('pv2',!dep) !concat('pv3',!dep) !concat('pv4',!dep) !concat('pv',5,!dep)
  !within !compare !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep) .
weight by !wgt.
aggregate outfile = *
   /break=!within !compare /stat=!stat(!concat('pv',!j,!dep)).
exe.
casestovars /id=!within /index=!compare .

!let !n=!length(!categ).
!let !m=!LENGTH(!substr(!blanks(!n),2)).

!let !var=" ".
!do !a=1 !to !m.
!let !s=!LENGTH(!concat(!blanks(!a),' ')).
!do !b=!s !to !n.

!let !d=!substr(!categ,!a,1).
!let !e=!substr(!categ,!b,1).
compute !concat('stat',!d,'_',!e)=
        !concat('stat.',!d)-!concat('stat.',!e).
!let !var=!concat(!var,'stat',!d,'_',!e,' ').
!doend.
!doend.

VARSTOCASES  /MAKE !concat('stat',!j) FROM !var
 /INDEX = contrast(!concat('stat',!j)) /KEEP = !within /NULL = DROP.
save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !within contrast.
sort cases by !within CONTRAST.
save outile='c:\temp\all.sav'.

* REPLICATES.

!do !i= 1 !to !nrep.
!do !j=1 !to 5.
get file !infile /keep !concat('pv1',!dep)
  !concat('pv2',!dep) !concat('pv3',!dep) !concat('pv4',!dep) !concat('pv',5,!dep)
  !within !compare !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep) .
weight by !concat(!rwgt,!i).
aggregate outfile = *
   /break=!within !compare /stat=!stat(!concat('pv',!j,!dep)).
exe.
casestovars /id=!within /index=!compare .

!let !n=!length(!categ).
!let !m=!LENGTH(!substr(!blanks(!n),2)).

!let !var=" ".
```

```
!do !a=1 !to !m.
!let !s=!LENGTH(!concat(!blanks(!a),' ')).
!do !b=!s !to !n.

!let !d=!substr(!categ,!a,1).
!let !e=!substr(!categ,!b,1).
compute !concat('stat',!d,'_',!e)=
        !concat('stat.',!d)-!concat('stat.',!e).
!let !var=!concat(!var,'stat',!d,'_',!e,' ').
!doend.
!doend.

VARSTOCASES  /MAKE !concat('statR',!j) FROM !var
 /INDEX = contrast(!concat('statR',!j)) /KEEP = !within /NULL = DROP.
save outfile=!quote(!concat('c:\temp\temp',!J,'.sav')).
!doend.

match files file='c:\temp\temp1.sav'
           /file='c:\temp\temp2.sav'
           /file='c:\temp\temp3.sav'
           /file='c:\temp\temp4.sav'
           /file='c:\temp\temp5.sav'
           /by !within contrast.

save outfile=!quote(!concat('C:\temp\',!dep,!i,'.sav')).
!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('c:\temp\',!dep,'1','.sav')).

!Do !e = 2 !to !nrep.
add files file=* /file=!quote(!concat('c:\temp\',!dep,!e,'.sav')).
!Doend.

sort cases by !within CONTRAST.
match files file=* /table='c:\temp\all.sav' /by !within CONTRAST.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
         b=stat1 to stat5/
         c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'c:\temp\regmod.sav'.

aggregate outfile=*/
         break=!within contrast/
          stat1 to stat5= mean(stat1 to stat5)/
          var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.

compute pv_var=mean(var1 to var5).

*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.

compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).
```

```
formats stat (f8.2)/ SE (f8.2).

string contr (a3).
compute contr=substr(contrast,5,3).
list cases/var= !within contr stat SE.
save outfile='c:\temp\temp.sav' .
get file='c:\temp\temp.sav'/keep=!within contr stat se.
casestovars /id=!within /index=contr /groupby=index.

!enddefine.
```

Cuadro 15.12. Sintaxis de la macro mcr_SE_PV_WLEQRT.sps

```
define PVWLEQRT  (nrep = !charend('/')
                  /stat = !charend('/')
                  /pv = !charend('/')
                  /WLE = !charend('/')
                  /grp = !charend('/')
                  /wgt = !charend('/')
                  /rwgt = !charend('/')
                  /cons = !charend('/')
                  /infile = !charend('/')).

get file=!infile /keep=!grp !wgt !concat(!rwgt,1) to !concat(!rwgt,!nrep)
!concat(pv,1,!pv)
  !concat(pv,2,!pv) !concat(pv,3,!pv) !concat(pv,4,!pv) !concat(pv,5,!pv)
!wle.

select if not missing (!wle).
means !wle /cell=mean min max.

* COMPUTE CUMULATIVE COUNT WITHIN COUNTRIES.
*-----------------------------------------------.
sort cases by cnt.
autorecode cnt /into cnt# /print.

!do !s=1 !to 5.
set seed = !s.
!let !v=!concat(!wle,!s).
!let !c=!concat('cumfreq',!s).

compute !v=!wle+rv.normal(0,.01).
sort cases by cnt# !v.

do if ($casenum=1 or lag(cnt#) <> cnt#).
compute !c=w_fstuwt.
else if (cnt#=lag(cnt#)).

compute !c=w_fstuwt + lag(!c).
end if.
sort cases by cnt.
!doend.
save outfile='C:\temp\temp.sav'.

* DEFINE CUTSCORES.
*-----------------------------------------------.
weight by w_fstuwt.
aggregate outfile=*  /break=cnt  /total=max(cumfreq1).
compute cut25=total/(100/25).
compute cut50=total/(100/50).
compute cut75=total/(100/75).

match files file='C:\temp\temp.sav'
  /table=*
  /by cnt.
exe.

* CREATE PERCENTILE GROUPS.
*--------------------------------------.
do repeat c=cumfreq1 cumfreq2 cumfreq3 cumfreq4 cumfreq5 /q=quart1 quart2
quart3 quart4 quart5.
if (c<cut25) q=1.
if (c>=cut25 & c<cut50) q=2.
if (c>=cut50 & c<cut75) q=3.
if (c>=cut75) q=4.
formats q (f1.0).
```

```
end repeat.
save outfile='C:\temp\quarters.sav'.

*** COMPUTE STATISTIC ***.
!do !j=1 !to 5.
get file='C:\temp\quarters.sav' /keep !grp !wgt
  !concat(!rwgt,1) to !concat(!rwgt,!nrep) !concat(pv,1,!pv)
  !concat(pv,2,!pv) !concat(pv,3,!pv) !concat(pv,4,!pv) !concat(pv,5,!pv)
quart1 quart2 quart3 quart4 quart5.
weight by !wgt.
rename var (!concat('quart',!j)=quart).
erase file=!quote(!concat('C:\temp\temp',!j,'.sav')).
aggregate outfile = !quote(!concat('C:\temp\temp',!j,'.sav')) /break=!grp
quart
   /!concat('stat',!j)=!stat(!concat('pv',!j,!pv)).
!doend.

match files file='C:\temp\temp1.sav'
           /file='C:\temp\temp2.sav'
           /file='C:\temp\temp3.sav'
           /file='C:\temp\temp4.sav'
           /file='C:\temp\temp5.sav'
           /by !grp quart.
erase file='C:\temp\all.sav'.
save outfile='C:\temp\all.sav'.

* REPLICATES.

!do !i= 1 !to !nrep.
!do !j=1 !to 5.
get file='C:\temp\quarters.sav' /keep !grp !wgt !concat(!rwgt,1) to !con-
cat(!rwgt,!nrep) !concat(pv,1,!pv)
  !concat(pv,2,!pv) !concat(pv,3,!pv) !concat(pv,4,!pv) !concat(pv,5,!pv)
quart1 quart2 quart3 quart4 quart5.
weight by !concat(!rwgt,!i).
rename var (!concat('quart',!j)=quart).
sort cases by!grp quart.
erase file=!quote(!concat('C:\temp\temp',!j,'.sav')).
aggregate outfile = !quote(!concat('C:\temp\temp',!j,'.sav')) /break=!grp
quart
   /!concat('statR',!j)=!stat(!concat('pv',!j,!pv)).
!doend.

match files file='C:\temp\temp1.sav'
           /file='C:\temp\temp2.sav'
           /file='C:\temp\temp3.sav'
           /file='C:\temp\temp4.sav'
           /file='C:\temp\temp5.sav'
           /by !grp quart.
erase file=!quote(!concat('C:\temp\',!pv,!i,'.sav')).
save outfile = !quote(!concat('C:\temp\',!pv,!i,'.sav')).

!doend.

*** COMBINE RESULTS ***.

get file =!quote(!concat('C:\temp\',!pv,'1','.sav')).
cache.

!Do !e = 2 !to !nrep.
add files/file=*/file=!quote(!concat('C:\temp\',!pv,!e,'.sav')).
!Doend.

sort cases by !grp quart.
```

```
match files file=*/table= 'C:\temp\all.sav'/by !grp quart.
exec.

*** COMPUTE SAMPLING VARIANCE (U) ***.

do repeat a=statr1 to statr5/
         b=stat1 to stat5/
         c=var1 to var5.
compute c=(a-b)**2.
end repeat.
save outfile = 'C:\temp\regmod.sav'.

aggregate outfile=*/
          break=!grp quart/
          stat1 to stat5= mean(stat1 to stat5)/
          var1 to var5 = sum(var1 to var5).

do repeat a=var1 to var5.
compute a=!cons*a.
end repeat.

compute pv_var=mean(var1 to var5).

*** CALCULATING MEASUREMENT VARIANCE (Bm) ***.

compute stat=mean(stat1 to stat5).

do repeat a=stat1 to stat5/b=pvar1 to pvar5.
compute b=(a-stat)**2.
end repeat.

compute pvmerr=.25*(sum(pvar1 to pvar5)).

*** COMPUTE STANDARD ERROR [V = U + (1+1/M)Bm] ***.

compute SE=sqrt(pv_var+1.2*pvmerr).

formats stat (f8.2)/ SE (f8.2).
list cases/var= !grp quart stat SE.

!enddefine.
```

Referencias bibliográficas

Baumert, J., S. Gruehn, S. Heyn, O. Köller y **K.U. Schnabel** (1997), *Bildungsverläufe und Psychosoziale Entwicklung im Jugendalter (BIJU): Dokumentation - Band 1,* Max-Planck-Institut für Bildungsforschung, Berlin.

Baumert, J., S. Heyn y **O. Köller** (1994), *Das Kieler Lernstrategien-Inventar (KSI),* Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel, Kiel.

Beaton A.E., I.V.S. Mullis, M.O. Martin, E.J. Gonzalez, D.L. Kelly y **T.A. Smith** (1997), *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS),* Boston College, Chestnut Hill.

Beaton, A.E. (1987), *The NAEP 1983-1984 Technical Report,* Educational Testing Service, Princeton.

Beaton, A.E. (1988), *The NAEP 1985-86 Reading Anomaly: A Technical Report,* Educational Testing Service, National Assessment of Educational Progress, Princeton.

Bloom, B.S. (1979), *Caractéristiques individuelles et apprentissage scolaire,* Éditions Labor, Bruxelles.

Bryk, A.S. y **S.W. Raudenbush** (1992), *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods,* Sage Publications, Beverly Hills.

Cornfield, J. (1951), "A Method for Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix", *Journal of the National Cancer Institute* Vol. 11, Oxford University Press, Oxford, pp. 1269-1275.

Dunn, O.J. (1961), "Multiple Comparisons among Means", *Journal of the American Statistical Association,* Vol. 56, American Statistical Association, Alexandria, pp. 52-64.

Eignor, D., C. Taylor, I. Kirsch y **J. Jamieson** (1998), "Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Students", *TOEFL Research Report No. 60,* Educational Testing Service, Princeton.

Ganzeboom, H.B.G., P. de Graaf y **D. J. Treiman,** con **J. de Leeuw,** (1992), "A Standard International Socioeconomic Index of Occupational Status", *Social Science Research,* Vol. 21(1), Academic Press, New York, pp. 1-56.

Goldstein, H. (1995), *Multilevel Statistical Models,* segunda edición, Edward Arnold, London.

Goldstein, H. (1997), "Methods in School Effectiveness Research", *School Effectiveness and School Improvement,* 8, Swets & Zeitlinger, Lisse, The Netherlands, pp. 369-395.

Gonzalez, E.J. y **A.M. Kennedy** (2003), *PIRLS 2001 User Guide for the International Database,* Boston College, Chestnut Hill. Guilford, J.P. (1954), Psychometric Methods, MacGraw Hill, New York.

Husen, T. (1967), *International Study of Achievement in Mathematics: a Comparison of Twelve Countries,* Almqvist & Wiksells, Uppsala.

Judkins, D.R. (1990), "Fay's Method of Variance Estimation", *Journal of Official Statistics,* Vol. 6, No. 3, Statistics Sweden, Stockholm, pp. 223-239.

Kish, L. y **M.R. Frankel** (1974), "Inference from Complex Sample", Journal of the Royal Statistical Society (B), 36, p. 1-37. Royal Statistical Society, London.

Kish, L. (1987), *Statistical Design for Research,* John Wiley & Sons, New York.

Marsh, H. W., R.J. Shavelson y **B.M. Byrne** (1992), "A Multidimensional, Hierarchical Self-concept", en R. P. Lipka and T. M. Brinthaupt (eds.), *Studying the Self: Self-Perspectives across the Life-Span,* State University of New York Press, Albany.

Monseur, C. y **R.J. Adams** (2002) "Plausible Values: How to Deal with Their Limitations", artículo presentado en el *International Objective Measurement Workshop,* New Orleans, 6-7 April.

OECD (Organisation for Economic Co-operation and Development) (1998), *Education at a Glance – OECD Indicators,* OECD, Paris.

OECD (1999a), *Measuring Student Knowledge and Skills – A New Framework for Assessment,* OECD, Paris.

OECD (1999b), *Classifying Educational Programmes – Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2001), *Knowledge and Skills for Life – First Results from PISA 2000,* OECD, Paris.

OECD (2002a), *Programme for International Student Assessment – Manual for the PISA 2000 Database,* OECD, Paris.

OECD (2002b), *Sample Tasks from the PISA 2000 Assessment – Reading, Mathematical and Scientific Literacy,* OECD, Paris.

OECD (2002c), *Programme for International Student Assessment – PISA 2000 Technical Report,* OECD, Paris.

OECD (2002d), *Reading for Change: Performance and Engagement across Countries,* OECD, Paris.

OECD (2003a), *Literacy Skills for the World of Tomorrow – Further Results from PISA 2000, OECD,* Paris.

OECD (2003b), *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills,* OECD, Paris.

OECD (2004a), *Learning for Tomorrow's World – First Results from PISA 2003,* OECD, Paris.

OECD (2004b), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005), *PISA 2003 Technical Report,* OECD, Paris.

Owens, L. y **J. Barnes** (1992), *Learning Preferences Scales,* Australian Council for Educational Research, Camberwell.

Rust, K. y **S. Krawchuk** (2002), *Replicate Variance Estimation Methods for International Surveys of Student Achievement,* International Conference on Improving Surveys, Copenhagen.

Rust, K.F. y **J.N.K. Rao** (1996), "Variance Estimation for Complex Surveys Using Replication Techniques", *Statistical Methods in Medical Research*, Vol. 5, Hodder Arnold, London, pp. 283-310.

Rust, K.F. (1996), TIMSS 1995 working paper.

Warm, T.A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika*, Vol. 54(3), Psychometric Society, Williamsburg, Va., etc., pp. 427-450.

Westat (2000), *WesVar complex samples 4.0* [programa de ordenador]. Westat. Rockville.

Wright, B.D. y **M.H. Stone** (1979), *Best Test Design: Rasch Measurement,* MESA Press, Chicago.

Wu, M. y **R.J. Adams** (2002), "Plausible Values – Why They Are Important", artículo presentado en el *International Objective Measurement Workshop,* New Orleans, 6-7 April.

Wu, M.L., R.J. Adams y **M.R. Wilson** (1997), *Conquest: Multi-Aspect Test Software* [programa de ordenador], Australian Council for Educational Research, Camberwell.